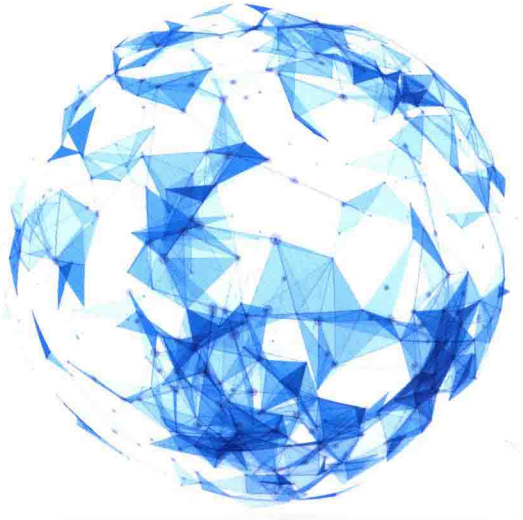
热销图书《Spark 原理、机制及应用》的续作，本书理论结合实践，以通熟易懂的方式深入解析大数据治理与安全在学术界和工业界的应用场景与重要作用，是不可多得的技术实践类前沿佳作。

大数据技术丛书



Big Data Governance and Security
From Theory to Implementation

大数据治理与安全
从理论到开源实践

刘驰 胡柏青 谢一◎等编著

机械工业出版社
China Machine Press

大数据技术丛书

Big Data Governance and Security
From Theory to Implementation

大数据治理与安全
从理论到开源实践

刘　驰　胡柏青　谢　一
施盟捷　陈喆毓　林秋霞　◎编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

大数据治理与安全：从理论到开源实践 / 刘驰等编著 . —北京：机械工业出版社，2017.8
（大数据技术丛书）

ISBN 978-7-111-57997-7

I. 大… II. 刘… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2017）第 222527 号

本书主要从理论和实践两个部分对大数据治理与安全技术展开详尽描述。其中理论篇主要从大数据治理的概念、作用、重要性，以及大数据治理的原则、范围及评估内容做出了详细介绍；之后从大数据安全、隐私和审计三个方面，探讨了大数据安全所面临的挑战，以及解决这些问题的技术与方案、作用与意义。开源实践篇分别从 Apache 的四个开源组件 Falcon、Atlas、Ranger 和 Sentry 以及 Kerberos 软件框架与工具介绍其在大数据治理与安全方面的功能与实践应用方案。

本书适用于大数据应用技术爱好者以及具有一定开发经验的读者，也可以作为大数据相关课程的教学参考书，供云计算、大数据相关专业方向的本科生、研究生阅读，亦可作为相关从业人员与一线软件开发人员的参考资料。

大数据治理与安全：从理论到开源实践

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：陈佳媛　　责任校对：李秋荣
印　　刷：北京诚信伟业印刷有限公司　　版　　次：2017 年 9 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：24.75
书　　号：ISBN 978-7-111-57997-7　　定　　价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

Preface 前　言

在大数据时代，随着信息量与日俱增，数据价值也得到越来越多人的认可。但大数据在迅猛发展的同时也带来不少问题，如怎样管理数据、实现数据价值最大化等，这些问题始终未得到完美的解答。在不同时间段，针对不同业务需求，数据的价值也不尽相同。为了最大化大数据的价值，互联网数据共享不可避免。然而，由于各个企业和部门之间相互独立，数据所在的系统甚至数据存储结构存在较大差异，数据之间难以进行信息共享，从而造成信息孤岛这一普遍现象。同时，互联网庞大的使用群体，也使得互联网数据在实现共享时，难以保障数据的安全性以及数据隐私。

为了解决这些问题，大数据治理与安全成为当下学术界与工业界最热门的研究领域之一。大数据治理主要在于建立一个统一标准化平台，从不同数据源中获取数据，在对数据进行生命周期管理的同时允许各方对数据进行相应操作（例如数据审计、数据筛选以及数据迁移等），从而实现数据价值最大化。而在数据业务流程中，这个统一标准化平台能够针对不同用户，根据不同的时间点以及 IP 地址，对不同的元数据进行权限设置，以保证数据使用的安全性。

本书总体分为两部分。第一篇：理论篇，包括第 1 章和第 2 章。第 1 章从大数据治理的概念以及作用两方面，阐述大数据治理的重要性，并对大数据治理的原则、范围及评估内容做了详尽介绍。第 2 章从大数据安全、隐私和审计三个方面出发，探讨了大数据安全所面临的挑战与问题，以及解决这些问题的技术与方案。

第二篇：开源实现篇，包括第 3～7 章。作者对开源社区中的大数据治理与安全相关的开源项目做了充分的介绍和实践，将内容根据不同组件分类，汇总成为该篇的主要内容。该篇全面介绍了 Apache Falcon、Apache Atlas、Apache Ranger、Apache Sentry 与 Kerberos 等大数据治理与安全开源组件的技术概况、配置与使用、场景设计与实现以及具体应用举例等多方面的内容。

第 3 章深入介绍建立在 Hadoop 环境下的数据过程及数据集管理系统 Apache Falcon 的

技术概况与架构特点。在此基础上，对集群上进行数据保留、生命周期管理、数据血统及追踪等功能进行介绍。并且设计与实现了日常生产环境中可能用到的数据处理场景，可作为相关从业者的参考。最后作者举例说明了 Falcon 在数据流程管理领域的使用前景。

第 4 章全面介绍元数据管理框架 Apache Atlas 的技术概况、配置使用与具体使用场景等核心内容。本章首先介绍 Apache Atlas 在元数据管理方面的突出优势，进而对 Hive、Sqoop、Storm 及 Falcon 等多种元数据导入方式进行了介绍，并对元数据的管理做了十分深入的阐述。在此基础上，对 Atlas 的实时数据、非实时数据等元数据管理场景进行了设计与实现，可以作为类似场景下构建与使用的参考。

第 5 章讲述安全认证框架 Apache Ranger 的技术概况、发展近况、插件集成和功能验证等内容。本章首先介绍 Apache Ranger 在 Hadoop 生态系统中实施安全认证的优势和特点，并对 Hadoop 生态组件如 HDFS、Hive、HBase 等如何进行安全数据访问控制做出详细阐述。最后给出了 Ranger 四种不同策略的实际场景，对其安全功能进行了验证。

第 6 章对 Cloudera 公司发布的高度模块化的权限管理组件 Apache Sentry 做了深入的介绍，弥补了 Hadoop 文件系统 HDFS 缺乏对数据和元数据细粒度权限访问支持的问题。从 Sentry 的特点、优势、发展近况三个方面，对其架构中的 Binding、Policy Engine 和 Policy Provider 三大核心组件进行了详细的阐述。并介绍了 Sentry 的搭建与部署步骤，以及其与 Impala 的集成步骤和在各类场景下 Sentry 的设计与使用方法。

第 7 章除了对网络认证协议 Kerberos 的特点与组成、架构与应用等做了介绍以外，还对大数据应用下的诸多组件与 Kerberos 的集成做了详细的实践介绍，包括 HDFS、Yarn、Zookeeper、Hive、HBase、Sqoop、Hue、Spark、Solr、Kafka、Storm 与 Impala，几乎涵盖了大部分学术界与工业界所涉及的各类组件，能够为高校科研人员与企业开发人员提供有效的参考与帮助。

作者认为大数据治理与安全理论部分已经有一些书籍进行了较好的阐述，而实践应用部分却十分匮乏。因此本书着重在实践部分使用大量篇幅进行详细的讲解描述。若读者想要查阅大数据治理与安全的相关理论内容，作者推荐桑尼尔·索雷斯的《大数据治理》和张邵华的《大数据治理与服务》两本书作为进一步的参考。

本书的作者除了封面和内封提到的六位之外，还有王文杰、段雄、吴琪、方久鑫、童楚云、陈超源、徐杰、陈喆、吴岳秋、吴成、张晶。

大数据发展迅速，而大数据治理与安全作为其分支，发展更是日新月异。由于作者水平有限，书中难免有不足与谬误之处，若读者发现问题并不吝告知，不胜感激。

本书讲述的相关组件，请读者到 www.bitlinc.cn 进行下载。

刘驰

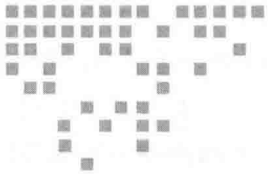
lincbit@gmail.com

Contents 目　录

第一篇 Part 1

理 论 篇

理论篇主要介绍大数据治理技术的相关背景知识，包括大数据安全、隐私和审计技术。其中，第1章总体介绍大数据治理的基本概念以及框架，使读者了解大数据治理的现状以及为此制定的相关原则、范围以及评估细则。第2章分别就大数据安全、隐私以及审计技术做了更加细致的介绍，让读者对大数据治理面临的挑战和解决方案有一个更加全面的了解。本篇的目的是使读者了解大数据治理的背景，包括急需解决的问题和常用的技术方案，使得读者对大数据治理有一个理论上的认知，为后续开源实现篇的学习打好基础。



Chapter 1 第 1 章

大数据治理技术

1.1 概述

1.1.1 大数据治理的基本概念

现如今，我们已被数据包围，数据正在逐渐将我们淹没。来自于社交媒体、网络日志、GPS 信号、RFID 标签、网络音频、数字图片等方面的数据扑面而来。大数据被炒得火热，大数据时代已然来临。而大数据本身是一个比较抽象的概念，如果我们仅仅从字面来理解，它表示数据规模的庞大。但是仅仅数量上的庞大这一简单的理解显得有些狭隘，难以区分这一概念和以往的“海量数据”“超大规模数据”等概念的区别。而现如今，当谈到大数据定义时都运用比较有代表性的 3V 定义，即认为大数据需满足以下 3 个特点：规模性（Volume）、多样性（Variety）和高速性（Velocity）。而 IDC 认为还应该添加数据具有的价值性（Value），IBM 认为大数据必然具有真实性（Veracity）。当然每个人对大数据有不同的理解，当我们面对实际问题时，没必要拘泥于这些现有的定义，只要符合业务规则即可。

伴随着网络和信息技术的不断发展与普及，人类产生的数据量正在呈指数级增长，在历史上从未有哪个时代产生如此海量的数据。数据的产生已经完全不受时间、地点的限制，大约每两年就会翻一倍，换句话说，每两年产生的数据量相当于之前产生的全部数据量。并且根据现有的数据量监测，这个速度还会在很长一段时间内保持下去。信息数据的单位由 TB → PB → EB → ZB 的级别暴增，而这样的数据很明显已经远远超出了我们人力所能处理的范围，因此大数据应运而生。它的重要性也因此而得之。

伴随着数据行业的昌盛发展，很自然就产生了一个对应的问题：这些数据作为原材料应该怎么管理？虽然数据管理并不新鲜，很早以前我们也一直在做，但随着数据爆炸性地呈指数级增长，我们如今所讲的数据和以往已经大大不同。而这也不仅仅体现在数据的大

小上，同时也体现在数据的内容、来源、结构上。举个简单的例子，现如今 Facebook 的日均新增数据量可达 600TB 左右，未来必然会更高。那么处理如此大量的数据，我们不禁要问：以往的算法还可能吗？应用还能正常运行吗？答案是否定的。随着数据的变化，我们的算法也要升级，同样，我们以往的数据管理方式与思路也无法完全适应，也需要创新。因此大数据治理的概念应运而生。

既然已提出大数据治理的概念，那么它应该和大数据管理有明显的区别。COBIT5[⊖]对两者进行了精准的区分定义。

1. 管理定义

管理（Management）是指按照治理机构设定的方向展开计划、建设、运营和监控活动，以实现企业目标。

基于此定义，管理包含计划、建设、运营和监控 4 个关键活动，并且活动必须符合治理机构所设定的方向和目标。

2. 治理定义

治理（Governance）是指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，然后根据方向和目标来监督绩效与规范。

基于此定义，治理包括评估、指导和监督 3 个关键活动，并且输出结果与设定方向必须和预期的目标一致。

从上述定义可做如下总结。

1）关键活动不同：管理包含计划、建设、运营和监控 4 个关键活动，治理包含评估、治理和监督 3 个关键活动。

2）过程不同：根据 COBIT 5 的定义，管理包括 4 个域，APO（调整、计划和组织）、BAI（建立、获取和实施）、DSS（交付、服务和支持）、MEA（监视、评价和评估），每个域又包含若干个流程。而治理包含如下过程，框架的设置与维护、确保资源化、风险化、收益交付、利益相关透明。

3）分工不同：治理相当于决策者，制定决策；管理相当于执行者，负责制定和实施决策的过程。

目前最权威的大数据治理的定义由桑尼尔·索雷斯[⊜]提出，主要包含如下 6 个部分：

⊖ COBIT（Control Objectives for Information and related Technology）是目前国际上通用的信息系统审计标准，由信息系统审计与控制协会在 1996 年公布。这是一个在国际上公认的、权威的安全与信息技术管理和控制的标准，目前已经更新至 5.0 版。它在商业风险、控制需要和技术问题之间架起了一座桥梁，以满足管理的多方面需要。该标准体系已在世界 100 多个国家的重要组织与企业中运用，指导这些组织有效地利用信息资源，有效地管理与信息相关的风险。

⊜ 桑尼尔·索雷斯是信息资产公司 LLC 的创始人和执行合伙人（LLC 专注于帮助组织构建信息治理计划），他曾任 IBM 的信息治理总监，其合作客户遍布六大洲和众多行业，他是较早提出大数据安全与治理理念的先驱之一。

1）大数据治理应该被纳入现有的信息治理框架内。

2）大数据治理的工作就是制定策略。

3）大数据必须被优化。

4）大数据的隐私保护很重要。

5）大数据必须被货币化，即创造商业价值。

6）大数据治理必须协调好多个职能部门的目标和利益。

根据上述相关定义可知，为了形成有效的治理体系，治理和管理必须相互作用，相互配合，才能取得最优效果。很多技术上的相关领域涉及治理框架、数据优化、隐私保护等。

大数据的大规模性、高速性和多样性等特征，使得它不同于小量数据。将小量数据的隐私保护方法用在大数据上会有很大的局限性：大数据的多样性带来的多源数据融合使得传统的匿名化和模糊化技术几乎无法生效；大数据的大规模性与高速性带来的实时性分析使得传统的加密和密码学技术遇到了极大的瓶颈。此外，大规模的数据采集技术、新型存储技术以及高级分析技术使得大数据的隐私保护面临更大的挑战。因此数据的隐私保护与安全也是大数据治理的重要关注点之一。

而在数据治理的框架下，元数据的管理也显得尤为重要。元数据按照数据类别信息进行区分可分为技术元数据与业务元数据。

技术元数据是存储关于数据仓库系统技术细节的数据，是开发和管理数据仓库的使用的数据，它主要包括以下信息：数据仓库结构的描述，包括仓库模式、视图、维、层次结构和导出数据的定义，以及数据集的位置和内容；业务系统、数据仓库和数据集的体系结构和模式。

业务元数据从业务角度描述了数据仓库中的数据，它提供了介于使用者和实际系统之间的语义层，使得不懂计算机技术的业务人员也能够“读懂”数据仓库中的数据。业务元数据主要包括以下信息：使用者的业务术语所表达的数据模型、对象名和属性名；访问数据的原则和数据的来源；系统所提供的分析方法以及公式和报表的信息。还包括企业概念模型，这是业务元数据所应提供的重要信息，它表示企业数据模型的高层信息、整个企业的业务概念和相互关系。

而对于元数据的管理又可分为以下两部分。

1）数据质量的管理：就像超市对物品进行清理一样，我们的数据也需要定期清理。

2）信息生命周期的管理：对大数据进行存档，并在没必要继续保存某些数据时将它删除。

大数据安全与治理体系下需要解决的问题如图 1-1 所示。

图 1-1 大数据安全与治理体系

本书中，通过将 Apache 的 Ranger、Atlas、Falcon 以及 Hadoop 生态下的其他组件进行整合，形成完整的大数据安全与治理体系，以此来完成安全与隐私保护、元数据管理、

数据生命周期管理等问题。本书中的大数据治理框架如图 1-2 所示。读者初看时可能难以有清晰直观的认识，当读完本书再回头观看此图时定会有不一样的理解。

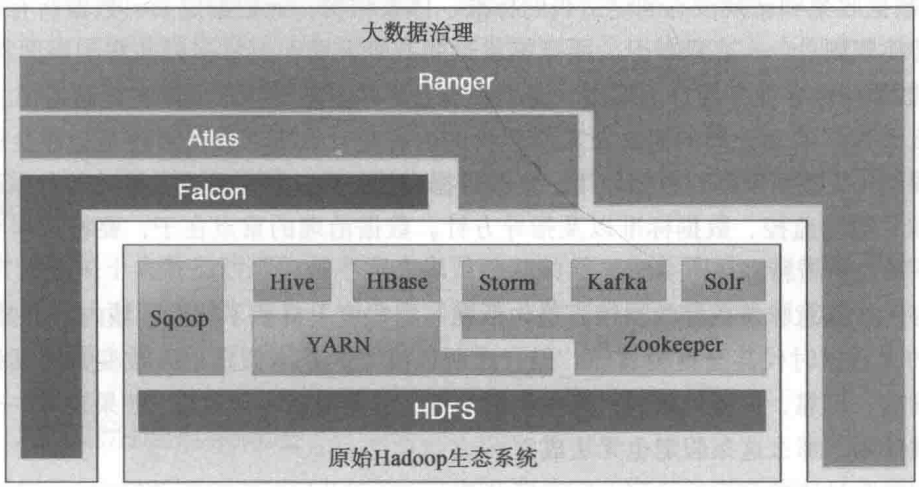


图 1-2　大数据治理框架

大数据的快速发展，使它成为 IT 领域的又一大新兴产业。据估算，国外大数据行业约有 1000 亿美元的市场，而且每年以 10% 的速度增长，增速是软件行业的 2 倍。而我国的大数据行业因起步稍晚，增速更为迅猛。而目前中国政府和企业对数据治理的重视程度也不断提升，在通信行业、银行行业、能源行业、互联网行业都已经开展了大数据治理的相关工作。在这个过程中，学术界和工业界做了很多探索，建立了较为科学、完整的数据治理理论体系和框架。本文从理论到实践引导读者加深理解，上文所提及的治理框架、数据安全、隐私保护、数据质量管理、数据生命周期管理都将在实践篇给出具体的实现。

1.1.2　大数据治理的意义和重要作用

如今，我们的生活已经被数据所淹没，但是目前主流的软件往往无法在合理的时间内完成对数据的撷取、管理、处理并整理成为帮助企业经营决策的重要资讯这些工作，而随着数据量的逐步扩增，这一现象会更加明显。所以企业经常要面对超出其基础设施和流程处理能力的大量数据，而从数据中挖掘出对制定有效决策有实际价值的情报更是难上加难。如今，由于种类、数量日益成倍增加的数据从社交媒体及各种在线渠道汹涌而来，导致处理上述数据的迫切性也日益加强，企业面临着更多的技术难题和挑战。

大数据不断从各种渠道、以多种格式涌入，其中蕴含着大量商业价值，但仅利用传统的数据处理方法和技术无法处理它们。故而早在 2009 年年初，《大数据资产：智慧企业如何在数据治理中胜出》的作者 Tony Fisher 就指出，如果基础数据不可靠，多数企业或大数据计划会失败，或者效果会低于预期。导致上述结果的关键原因是数据进入生命周期的不一致，数据不准确，数据不可靠。这些原因可能是多样性的：

1）大数据计划中的数据识别不完整。目前还不清楚如何获取数据，如何使用数据，哪些业务目标要满足，哪些人有权拥有数据。

2）数据收集和转换没有制定适当的标准、体系结构、元数据定义、数据所有权、策略和数据转换规则。

3）数据传输在业务用户上下文、安全性、数据和业务流程方面没有正确定义。

那么大数据治理计划的意义及其所包含的内容是什么呢？数据治理是指在企业数据生命整个周期（从数据采集到数据使用，直至数据存档）中，制定由业务推动的数据政策、数据所有权、数据监控、数据标准以及指导方针。数据治理的重点在于，要将数据明确作为企业的一种资产看待。

更好的数据意味着更好的决策，这句话在一定程度上反映了数据领域内的主要关注点，在当今的大数据时代甚至更为真切。但它之所以成立的基本假定也未改变，那就是“基本数据是准确、可靠、值得信赖的，来龙去脉清楚，并且具有一致性”。如果没有一个可靠的数据治理计划，那么这条假定也无法成立。

我们都听过诸如此类的说辞：“IT 技术融入业务对我们的企业至关重要”“IT 技术促成各种业务功能的实现”。但对企业上下进行实际的评估，能实现上述说辞的情况却是屈指可数。对大多数企业而言，IT 技术与各种业务目标之间仍存在差距，首席信息官及各高级主管仍在努力设法使 IT 技术能配合各种业务目标，从而促进企业战略目标的实现。在对成功企业进行分析后，可以得出一个很明确的结论，那就是“有效的数据治理计划”是成功企业的法宝。

任何大数据计划都应该考虑数据的以下特性：数量大、种类多、产生频率高、质量可靠性低、模糊性高。那么数据处理团队想要完全识别、定义并分析这些数据，就要征询企业各方利益相关者的意见。这样做才能让企业拥有者、数据拥有者以及数据治理部门在数据治理初期就避免一些错误，确保框架的正确搭建及实施，从而达到数据集规划与业务流程紧密联系且合理有效的目的。

现在，伴随着大数据运用时代的到来，所谓“数据驱动”已然成为未来全世界的发展趋势。现在大数据已经应用于全球的生产、分配及消费活动等，并且对于国家经济的运营体制、社会民生和国家的治理生产、制造能力等都会产生非常重要的影响。在未来，国家之间的竞争可能会从资本和土地等资源的争夺转移到大数据的争夺。所以，现在大数据已经成为每个国家的战略资源的基础设施，同时，大数据治理也成为多个国家提升现代治理能力的一个重要标杆。

随着互联网、云计算等网络相关的新技术的不断完善和知识普及，我们的社会已经进入大数据时代，大量数据的产生和流转都将成为再平常不过的事。到 2016 年年底，全球近 50% 的人口在使用互联网，人人都拥有一台或多台网络终端设备，随时随地都可以上网，所以全球的数据量也在飞速增长。2020 年，预计全球的数据使用量将会达到 40ZB，每个行业都将产生并使用大数据，大数据也将成为发展的新趋势。而大数据治理将为社会经济能

力发展提供新的动力。

在这个大数据时代，世界上各个国家都将大数据看作国家的核心资产。因此，对大数据的开发、利用和保护的概念就越来越强，可能还会产生对于大数据的争夺。大数据概念的出现就使得国家的强弱对比不仅体现在经济发展层面，还体现在一个国家大数据治理实力如何。所以对于大数据安全与治理的挑战也才刚刚开始。

1.2　框架

在讲述了关于大数据治理的基本概念以及治理的意义和作用后，我们对大数据治理已经有了简单的认识。接下来将会从 3 个维度阐述大数据治理的框架，目的是让读者更加深刻地认识、理解大数据治理。

1.2.1　大数据治理框架概述

大数据治理框架从全局视角描述了大数据治理的主要内容，下面我们从大数据治理原则、治理范围、治理的实施与评估 3 个维度给出大数据治理的全貌，展现大数据治理的重要性以及如何进行大数据治理，如图 1-3 所示。

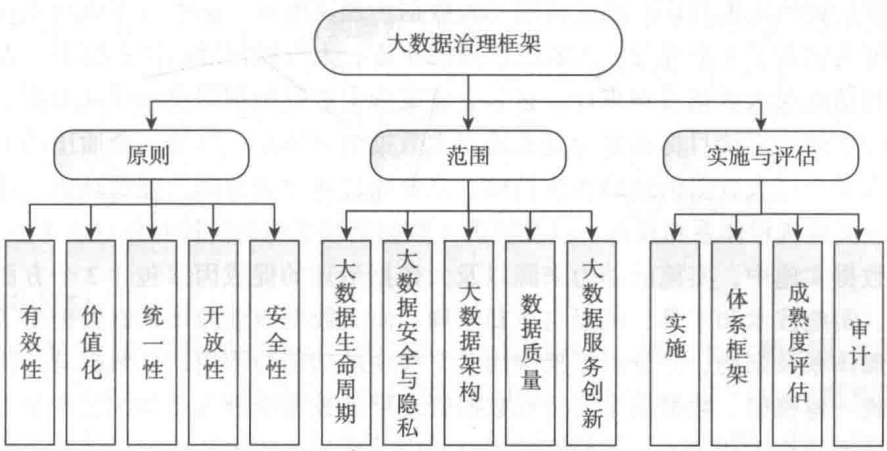


图 1-3　大数据治理框架

其中大数据治理的原则给出了大数据治理过程中所遵循的、首要的、基本的指导性法则，即有效性原则、价值化原则、统一性原则、开放性原则、安全性原则，这 5 个部分分别从各个层面、各个角度解释了大数据治理所应遵循的原则的重要性与必要性。其中，有效性原则体现了大数据治理过程中数据的标准、质量、价值、管控的有效性、高效性；价值化原则体现了大数据治理过程中以数据资产为价值核心，最大化大数据平台的数据价值；统一性原则能够形成一套规范的、有条理的、可遵循的准则，能够节约很大的成本、时间，对大数据的治理具有重要意义和作用；开放性原则是为了提高数据治理的透明度，不让海

量数据信息在封闭的环境中沉睡，同时共享信息，安全合理地共享数据，使数据之间形成关联，形成一个良好的数据标准；安全性原则体现了安全的重要性、必要性，保障大数据平台的数据安全和数据治理过程中数据的安全可控。

大数据治理的范围描述了大数据治理的关键域，即大数据治理决策层应该在哪些关键领域内做出决策。该维度共包含5个关键领域：大数据生命周期、大数据架构、大数据安全与隐私、数据质量以及大数据服务创新。这5个关键领域就是大数据治理的主要决策领域，规定了大数据治理主要应用的地方以及方向。其中，大数据生命周期是指数据产生、获取到销毁的全过程，在大数据治理中生命周期的管理更注重在成本可控的情况下有效地管理并使用大数据，从而创造出更大的价值。大数据生命周期管理包含了数据捕获、数据维护、数据合成、数据利用、数据发布、数据归档和数据清除。大数据架构是指大数据在IT环境下进行存储、使用以及管理的逻辑或物理架构，主要包含了大数据来源、大数据存储、大数据分析以及大数据应用和服务4个部分。大数据安全与隐私提供了大数据隐私管理的几个步骤，来对大数据云计算时代的数据进行隐私安全保障。数据质量领域总结了大数据产生质量问题的原因，以及应该从哪几个方面入手去有效提升大数据质量。大数据服务创新领域提出应该从基于数据本身进行创新、基于业务需求进行创新、基于数据分析的创新3个方面进行探讨，来体现对大数据服务的创新。

大数据治理的实施与评估维度描述了大数据治理实施和评估中需要重点关注的关键内容，该维度共包含了4个部分：大数据治理的实施、大数据治理的体系框架、大数据治理的成熟度评估以及大数据治理审计。它为企业实施大数据治理提供指导性方案。其中，大数据治理的实施的直接目标就是为企业建立大数据治理体系，形成一个通用的大数据治理架构。而为了实现大数据治理的实施目标，需要通过建立大数据治理的环境、建立完善的大数据治理实施流程体系和规范，以及明确大数据治理实施的阶段目标这3个方面来完成。同时在大数据实施中，实施的动力来源以及大数据治理的促成因素包含3个方面：治理实施的环境、实施技术和工具、流程与活动管理。而大数据治理的体系框架提出了一个通用的数据治理体系及架构，并分析了架构内各个模块的功能与作用，从数据持久化层、数据集成层、统一建模层、数据质量层、元数据管理层和数据治理人员组织层5个方面对大数据治理的体系结构进行阐述。大数据治理过程中，通过成熟度评估可以了解当前大数据治理实施的状态和实施方向，成熟度可以帮助了解治理的重要性。根据能力成熟度分类的方法，将成熟度分为5个等级，等级由低到高分别为：初始级、受管级、定义级、定量管理级、优化级。大数据治理的审计不仅可以提高大数据治理的实施水平，还能从更全面的角度为大数据治理提供实施意见，而且大数据审计还可以满足企业监管的需要，改善大数据在治理过程中的安全和隐私。

相关组织及企业可根据上述3个维度的指导原则，从大数据治理原则、治理范围、治理的实施与评估3个维度了解大数据的治理工作，按照治理原则中所遵循的指导性法则、治理范围中的治理关键域以及实施与评估维度中的关键内容，持续稳步地推进大数