Broadview
www.broadview.com.cn

IBM

大数据时代的分析之道，
你也能轻松获得的
分析洞察力！

Smarter Analytics

智慧的分析洞察

程永◎编著

唤醒数据·深入洞察·赢得优势

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

Smarter Analytics

智慧的分析洞察

程永◎编著

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

内 容 简 介

本书主要从体系结构和方法论层面讲述智慧的分析洞察、IBM“3A5 步”、信息议程、构建新一代数据中心、大数据管理、元数据管理、数据治理和主数据管理等相关概念、方法、模型和示例。针对信息供应链的每个领域以 IBM 相关产品简单举例。

本书适合 IT 从业者、CIO、数据库架构师、企业的架构师、IT 部门经理、数据库管理和开发人员阅读。同时还适合业务部门人士和互联网业务相关的业务经理，以及需要基于互联网进行业务创新的部门经理阅读。如果你正在思考如何开展业务创新，可以看看这本书，书中有不少适用于各个行业的例子，会启发你的业务创新。当然，本书对从事咨询业务的专家，像 ITSP 咨询师，以及从事教育科学研究领域的人士，也有一定参考价值。

图书在版编目（CIP）数据

智慧的分析洞察 / 程永编著. —北京：电子工业出版社，2013.11
ISBN 978-7-121-21620-6

Ⅰ. ①智… Ⅱ. ①程… Ⅲ. ①数据库管理系统－研究 Ⅳ. ①TP311.13

中国版本图书馆 CIP 数据核字(2013)第 237124 号

责任编辑：徐津平
印　　刷：北京中新伟业印刷有限公司
装　　订：北京中新伟业印刷有限公司
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：787×980　1/16　印张：16.75　字数：305 千字
印　　次：2013 年 11 月第 1 次印刷
印　　数：3000 册　　定价：75.00 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。

关于作者

程永，IBM 大中华软件集团资深信息管理专家，信息架构师。

程永从事信息管理相关工作 13 年以上，是 IBM 认证的 SOA 架构师。在大数据管理、数据治理、元数据管理、主数据管理、信息集成、数据仓库、业务分析等领域具有丰富经验。支持过许多政府、金融、烟草和能源电力行业的大型项目，并在 IBM 开发者园地（IBM developerWorks）、IT168 等网站撰写过 35 篇专业文章。

在金融行业，程永成功地支持了国内数家大型银行的数据中心、零售 CRM、风险信贷管理、数据仓库、主数据管理、Core Banking、电子影像和容灾等几十个大型项目的建设，帮助多个合作伙伴构建了银行新一代数据中心、分析型 CRM、风险信贷管理、主数据管理和 Core Banking 等解决方案并成功推广（Repeatable Solution）。

在政府行业，程永成功地支持了国内多个省市及部委的数据中心、电子政务并联审批、税务核心征管、发票管理、纳税评估、水利调度管理、财政资金监管、医疗卫生平台、公积金管理和电子监察等几十个大型项目，帮助多个合作伙伴构建了政府新一代数据中心、电子政务并联审批、医疗卫生平台、公积金管理等解决方案并成功推广。

在烟草行业，程永成功地支持过国内多个省市的烟草营销平台、工商协同、数据中心、电子政府和应用集成平台等几十个大型项目，并帮助合作伙伴构建了多个可推广的解决方案。

另外，程永还在煤炭、交通、通信、电力等多个行业支持过很多大型项目，如煤炭行业人员执法、任务调度、高速综合收费、智慧交通、通信网管综合分析、综合客服、电力数据交换平台、状态检修等项目。

推荐序一

IBM 广告的创意总是层出不穷。最近我看到一条很有意思的 IBM 广告，一个劫匪出发去便利店打劫，到了计划的犯罪地点，警察已经在那里布控，抢劫计划宣布失败。电影《少数派报告》中曾经描述的场景如今几乎成为了现实：通过罪行先知系统，于案发前终结犯罪。警察的职责在大数据时代将发生转变，从单纯的案发后追捕罪犯，到分析犯罪数据，识别犯罪模式，部署警力，预防犯罪，从而有效地降低犯罪率。这个广告并非空谈，IBM 的智慧城市方案已协助美国部分城市将重大犯罪率降低了 30%。

那么如何构建这样的系统呢？数据是关键，信息议程需要企业的认真思考。过去的几十年，IT 技术迅猛发展，企业构建了很多应用系统，应用议程是企业信息系统构建的传统思考方式。但是，随着应用系统的不断发展、更新，越来越多的企业发现，应用所产生的数据是孤立的，形成竖井，数据真正的价值有待开发。企业需要规划系统构建的信息议程，打通信息的“任督二脉”。本书从信息议程入手，深入浅出，向读者呈现了 IBM 所倡导的大数据理念，帮助读者了解构建新一代数据中心需要的基础知识。

“大处着眼、小处着手”是本书的一大特色。所谓“大处着眼”是指本书在论述新一代数据中心建构时，是从信息议程这一关系企业数据中心建设的长远目标切入。在几十年的应用系统建设中，我们有太多“头疼医头，脚疼医脚”的惨痛案例，归根到底就是缺乏大视野。对于许多成长中的企业来说，也许今天的分析洞察仅仅局限于单个数据平台的报表分析，但这并不意味着在明天没有针对跨数据平台的商业智能需求，也不意味着在更远的将来没有针对非结构化乃至大数据的分析需求。而信息议程的理念在于，我们对新一代数据中心的建设，不仅要有整体布局、有一个统领全局的时间表，还要对时间表上的每一步部署都胸中有丘壑。这也是信息化建设的终极经济之道，是避免重复建设、规避推倒重来最行之有效的方法。简而言之，本书首先基于作者多年从事数据中心建设的经验，总结出下一代数据中心建设的宏观战略，而信息议程正是这一战略的精髓。“小处着手”体现了本书的务实特性。从数据平台的整合，到分析模型的架构；从结构化信息的管理，到非结构化乃至大数据的加工利用，每一个环节都包含丰富的方法论和技术细节。依托于信息

议程这样的宏观架构，本书重点阐述了信息议程的每个环节所要解决的问题，以及对应的解决方案。但在探讨战术层面的问题时，本书绝不拘泥于烦琐凌乱的技术细节，而是提炼了作者的技术实践，择其精华而述之。

让业务人员也能看懂的 IT 读物，是本书的又一大特色。我们始终认为，智慧的分析洞察存在于我们工作和生活的点滴中，而非少数从事数据管理和分析人员的专利。因此，无论你从事何种工作，都有必要了解或应用到智慧的分析洞察。也正因为此，本书不同于一般的 IT 读物，以大段的程序代码示人，让业务从业者望而却步。本书在阐述任何一个问题的时候，都是从实际用例或者应用情景出发，让读者能身临其境，快速进入角色，更准确地理解所要探讨的问题，并把握所要传递的思想。基于这些具有普适意义的用例，即使将其称为企业数据中心建设的百科全书或 DIY 手册也不为过。

本书语言生动而不浮华，叙事有条理而不刻板，案例丰富而不累赘，这些都有助于我们以一种轻松的心态来了解智慧的分析洞察，让那些将分析洞察仅作为一种谈资或对其持高深莫测态度的人从根本上改变想法，构建下一代数据中心绝非可望而不可即，利用智慧的分析洞察来改变我们的工作和生活，处处掌握主动，不再是梦想！

智慧的城市，让我们从智慧的分析洞察开始！

刘隶放

信息管理北方区技术经理

IBM 大中华区软件集团

2013 年 7 月 7 日于北京

推荐序二

我们生活在巨大信息革命带来的生活方式变革中，从电报、电话、寻呼机，到手机、智能手机、平板电脑，每一次新技术催生的新产品都改变着我们的生活方式。今天，由大数据推动的又一次新技术变革将使我们生活在一个被数据左右的时代（甚至可以说被数据绑架的时代），离开了数据，生活将变得无所适从。

观察一下很多人的生活方式在发生怎样的变化，比如，买东西时访问各种购物网站，比较价格、功能、性能，还关注别人对这些商品的购物体验和使用反馈，一旦购买，我们又去提供这类评论。相信每个人的手上都有各种各样的卡和证，如身份证、医疗卡、购物卡、信用卡、借记卡，甚至门卡和饭卡等，所有这些卡在携带和使用过程中，都在消费和产生着各种信息，并被记录下来，我们制造出的信息将以数据的形式存在。我们每个人不只是数据的消费者，同时也是数据的制造者。

商家知道如何使用这些数据，挖掘其价值，帮助企业开展业务创新。大数据时代已经来临，我们拥有了更多的数据，如何利用好这些数据，是很多企业面临的挑战和机遇。

数据作为信息的载体，电子化过程经历了几十年的发展，其脉络清晰可见，可简单分成三个大的时代，第一个时代是计算机发展的早期，数据多以裸文件方式存在、管理和使用。第二个时代是 20 世纪 80 年代开始的关系型数据库时代，直到今天，关系型数据库都是高价值数据的主要存储管理方式，在未来相当长的时间里，关系型数据库都将继续发挥重要作用。第三个时代就是已经到来的大数据时代，这个时代充分吸收文件型数据时代的高效快速和关系型数据时代的简洁、易操作，把数据的大容量和易用性结合起来，使海量数据的使用变为可能。

本书是一本承上启下的书，书中有大量的知识和经验分享，详细介绍了在关系型数据库时代，数据处理的技术，包括数据库技术、数据仓库技术、主数据、元数据、数据中心建设、常见的主流产品等。同时，本书将读者领入大数据时代，用不少笔墨讲述大数据的基本概念和常见部件，以及大数据平台的体系结构，大数据平台下的前沿技术等。

作者提出混合型数据中心的概念很有新意，非常符合未来的数据中心规划蓝图，其描绘了未来 10 年，甚至更长一段时间内，企业的数据架构形态。

作者程永是 IBM 资深数据库专家，长期从事大型企业的数据库项目和数据中心建设。既有深厚的理论基础，又有丰富的实践经验。书中不少例子都是作者工作经验的积累和总结。

相信本书能把读者从传统关系型数据库时代带到大数据时代！

刘胜利

信息管理产品技术总监

IBM 大中华区软件集团

2013 年 7 月，北京

推荐序三

“大数据”已经成为时下最火热的IT词汇。相关数据指出，互联网上的数据将每年增长50%，每两年翻一番，而目前世界90%以上的数据是最近几年产生的。此外，数据并非单纯指人们在互联网上发布的信息，随着硬件技术的进步，越来越多的设备（计算机、智能终端、工业设备、交通监控等等）具备了更强的数据生成和处理能力，海量的数据信息给我们带来了之前无法企及的各种可能性——最重要的是它将对每个人的生活产生重大影响。

我们如何有效地处理庞大的数据信息呢？如何从各种类型的数据中，快速获得有价值的信息？事实上，我们看到非常成功的互联网企业，以及一些领先的传统企业，在这方面都进行了卓有成效的工作。或许我们可以说，数据将成为一种核心竞争力。

本书最大的价值是结合作者多年在信息管理领域的经验，以信息议程、“3A5 步”方法论和参考架构贯穿全书，为读者提供有效的指南；同时对一些行业提供参考案例，具有相当的实践性。程永作为 IBM 资深信息管理专家，多年来热心于分享自己的工作经验。我非常高兴看到这本书付梓出版。

王小虎

渠道及区域拓展技术总经理

IBM 大中华区软件集团

致　谢

IBM 始终是渴求进步者实现梦想的地方，感谢 IBM 提供的无与伦比的学习经验和工作挑战。

感谢 Ernie Hu、Tom Chan、刘胜利和刘隶放，本书能够出版，离不开他们的鼎力支持和帮助。

感谢王小虎、张红卫、王媛媛和刘广，在过去 6 年对我的一贯支持和鼓励。

特别感谢 Mentor 李磊对本书提供的技术指导，过去几年他一直在信息议程、BI 和大数据等多个领域引导我前进。同时，特别感谢 Mentor 张利民在数据库方面给予的大力指导，从 2005 年到现在，这种指导和帮助一直在持续。

总体来说，我要感谢以下人员，没有他们，本书不可能完成。

- 感谢阎卫防、秦磊、管连、吴湘洲、刘慎锋和王小宜，为出版本书提供的大力支持。
- 感谢朱宏、仲崇国和 Sammi Wang 为出版本书提供市场、法律和流程等全方位支持。
- 感谢刘皎、陈曦、金春霞、朱丽萍、陈斌，以及出版社的各位编辑在本书出版和发行过程中提供的帮助。
- 感谢王清华过去 3 年在 SPSS、分析型 CRM 和分析型税务解决方案方面给予的支持和帮助。感谢张瑞锋在 Cognos 方面给予的支持和帮助。
- 感谢杨国彦、雷林、周雄志、张忠和区波长久以来在数据库、数据仓库和模型方面提供的意见和见解。
- 感谢陈威、杨俞平、张萌、蔡玉全、郝多慧、颜悦悦、王敏、陈赟和艾飞在大数据管理方面的独到见解。

◎ 感谢刘春霞在信息整合和数据治理方面提供的意见。

◎ 感谢李英伟和邓俊宁在主数据管理方面所做的卓越工作。

◎ 感谢廖安舟在数据归档、测试、审计方面的意见。

◎ 感谢张光业在合作伙伴解决方案构建方面的丰富经验。

感谢我的妻子和家人，正是你们长久的包容和支持使本书得以完成。特别是我的妻子，本书耗费了我无数的个人时间，耽误了太多次出游和聚会，没有她的理解和支持，本书不可能完成。

感谢各业务合作伙伴和客户在过去多年合作过程中的鼎力支持，正是你们的支持使一个个项目终获成功。

最后，感谢与我合作多年的所有 IBM 同事，是你们陪伴我学习和成长，使我写书的梦想成为现实。

作者序

《智慧的分析洞察》是我在过去 5 年中支持合作伙伴实现智慧的分析洞察和构建新一代数据中心解决方案的经验总结，我从 2012 年 1 月开始利用业余时间编写本书，历经 1 年半，终于成稿。在编写过程中大量借鉴了 IBM 方法论、模型、白皮书、解决方案、实际案例、各产品信息中心，以及其他资料。

本书最初的创作灵感来自于我在 2009 年领导开发的一个高级课程“信息随需应变和业务分析”，该课程在 2009 年年底完成课件开发，随后在 2010 年、2011 年和 2012 年分别于北京、上海、南京和成都等地举行过多次培训，取得了良好的效果。与此同时，在实际工作过程中，我发现客户和合作伙伴非常需要一本能全面阐述、贯穿信息供应链各个环节的书来指导各种解决方案的构建。于是，我于 2011 年下半年开始构思本书，并于 2012 年 1 月动笔，原本计划 1 年内完成（可惜工作越来越忙，业余时间太少），结果花了 1 年半才彻底完成。

本书主要从总体上阐述智慧的分析洞察、IBM“3A5 步”模型、信息议程、构建新一代数据中心、大数据管理、元数据管理、数据治理和主数据管理等相关概念、方法、模型和示例，最后简单介绍了 IBM 的相关产品。通过阅读本书，读者可以了解到：

- 如何通过对数据的唤醒，获得深入洞察力，帮助企业获得独特的竞争优势。
- 如何通过信息议程，帮助企业构建信息单一视图，使信息成为战略资产。
- 如何基于参考模型，构建新一代混合型数据中心（大数据平台）。
- 如何基于企业级 Hadoop 和流数据分析计算平台，进行大数据静止和移动分析。
- 如何基于模型驱动，构建企业级元数据管理体系结构。
- 如何基于数据治理统一流程模型和模型扩展，提升业务价值。
- 如何基于 IBM 信息管理和业务，分析相关产品，实现智慧的分析洞察。

关于本书的故事总线（storyline）请参见本书第 1.2 节内容（考虑到很多人不会看前言，所以直接放在第 1 章了）。

由于水平有限，难免存在对各个领域知识理解不到位及各种人为或机械错误，欢迎广大读者批评指正（欢迎读者将意见发到邮箱 smarteranalytics@sina.cn），再版时将根据反馈进行修正。

目　录