Informatica PowerCenter中文权威作品

Broadview
www.broadview.com.cn

Informatica PowerCenter 权威指南

杜绍森 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

Informatica PowerCenter 权威指南

杜绍森 著

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

内 容 简 介

在大数据时代，掌控数据首先需要掌握数据的处理能力。俗话说：“工欲善其事，必先利其器。”Informatica PowerCenter 作为业界广泛使用的数据处理工具之一，被全球多数大型机构、组织认可并采用。

本书全面地介绍了 Informatica PowerCenter 的主要功能及高级特性。

本书分为 3 个部分：第一部分为基础篇，包括第 1～4 章，系统介绍了 PowerCenter 的基础组件和常用功能，并在其中穿插了大量实践案例；第二部分为高级篇，包括第 5～8 章，系统介绍了 PowerCenter 并行、集群、性能调优和字符集管理等高级内容；第三部分为扩展篇，包括第 9 章，简要介绍了 CDC 的相关知识，PowerCenter 与 SAP、MPP、Hadoop 集成，以及非结构化和半结构化数据处理能力。

本书适合 PowerCenter 的入门者及有一定 PowerCenter 使用经验的用户参考，也可作为各数据仓库、大数据专业培训机构的培训教材。

图书在版编目（CIP）数据

Informatica PowerCenter 权威指南 / 杜绍森著. —北京：电子工业出版社，2015.9
ISBN 978-7-121-27045-1

Ⅰ. ①I… Ⅱ. ①杜… Ⅲ. ①企业管理—数据管理 Ⅳ. ①F270.7

中国版本图书馆 CIP 数据核字（2015）第 203208 号

责任编辑：徐津平
特约编辑：赵树刚
印　　刷：北京京师印务有限公司
装　　订：北京京师印务有限公司
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：22.75　　字数：475 千字
版　　次：2015 年 9 月第 1 版
印　　次：2015 年 9 月第 1 次印刷
印　　数：3000 册　　定价：69.00 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

犹豫了很久，以我现今的职位给作者写序，是否有些自吹自擂？但读完书稿，我决定了：一本好书，介绍一个好产品，既然与我受用，何不推荐给更多的人呢？

“IT”是信息（Information）和技术（Technology）的缩写，它的发展不过三十多年的时间。在IT发展前期的大部分时间里，其所有进步大部分体现在“T”上，例如计算性能、存储容量、网络拓展及打印效果等。直到近些年“大数据时代”的出现，人们才开始了对于信息数据，也就是“I”的关注。我相信，这个变化是IT发展的必然，是一个破茧成蝶的过程，并且这个关注也一定会延续很多年。

同样，正是因为“大数据”日益深入人心，企业的IT规划和发展越来越与“大数据”相关联，PowerCenter才得以从一个IT人员得心应手的工具，蜕变为大数据应用的一个重要环节。记得在2014年国务院工业和信息化部颁布的大数据白皮书中，就明确地将“数据准备”定义为大数据发展的第一个环节。由此，作为在数据集成领域里历年排名第一的PowerCenter，也就承担起了“帮助企业实现大数据应用的第一步”的重要使命。

本书前6章中规中矩，如同一本深入浅出的教科书，将具备一些基本IT知识的人士引进数据迁移的奇妙世界，加上作者风趣的调侃，学来丝毫不觉得枯燥单调。第7章开始是实战描述，实际上是一系列的应用经验分享，这些宝贵的经验之谈，可以让初学者在未来的实践中少走弯路，还可以将本书作为可以随时受教的参考书。更值得一提的是，不同于普通的产品手册，本书作者以其十几年的理论研究和教育培训，以及主导或参与诸多中外企业“数据集成项目”实施的经验，将PowerCenter的很多功能细节描述得淋漓尽致。本书对于有意进行ETL教学的教育培训机构，不失为一本经典的教材；而对于有意培养自己成为ETL应用高手的IT人士，则是一本有益而又有趣的读物。

曾经有不少朋友问我：当成了ETL的行家里手以后，下一个职业目标会有哪些发展方

向？所以，我想借此序的一角，分享一些我的认知，供大家参考。

第一，云数据集成和管理。根据 IDC 的预测，2017 年全球 SaaS 和云软件模式将占软件开支的 1/6。越来越多的云应用系统承诺并交付更简单、更快捷和更智能的业务营运方法，所以，掌握云数据集成，会让你在不可阻挡的云服务趋势下游刃有余。

第二，下一代数据洞察。不同于第一代商业智能（BI）对展示形态和分析过去的重视，大数据时代的数据洞察，更加关注数据的质量而不是数据的展现形式，更重视预测未来的行为模式而非过去的行为分析。所以，要想成为大数据分析专家，你必须懂得数据质量管理和前瞻性的分析。当然，保障分析结果正确的前提是确保数据的统一性、完整性，并找到数据的关联性。

第三，数据治理。大数据时代，越来越多的企业将数据纳入其固定资产；在金融和医疗行业，数据相关的合规性成为政府监管的重要指标；为了应对客户需求和市场业务模式的变化，许多企业开始考虑应用整合和迁移……这些巨大的变化，不断催生出数据治理的高手，他们必须在行业规范、企业应用系统、数据的关联性和安全性方面具备独特的技能。因此，了解行业特性、行业应用，使之与数据集成相结合，便成为你进行数据治理的更高境界。

近年来，关于大数据的定义一直在调整，而大数据应用的目标却始终没变，那就是：发现数据价值，帮助企业降低成本并实现业务创新。在过去短短的两三年里，中国作为自然的大数据国家，已经在大数据的理论研究和实际应用方面取得了巨大的进展。大数据的应用会推动各行各业诞生越来越多的数据科学家，那是行业知识和数据治理兼备的卓越人才。IT 的发展已经实现了由“计算机科学”向“数据科学”的转换，近年来，“数据科学”又开始向行业应用进行大规模迁移。所以，数据科学家既是数据价值的挖掘者，更是行业产品和流程的创新者，他们的价值不是向企业的高管提供分析报表，他们本身就是企业的高管，他们在用数据作为依据，实现企业面向客户、市场、产品和流程方面的创新。

千里之行，始于足下。与各位读者共勉。

Informatica 大中国区总经理　王晨杰

自 序

初识 Informatica，大概是在十四五年前的一个偶然的机会，公司接到一个叫作决策支持系统（DSS）的项目。尽管当时作为工程师和客户一起整理了项目的需求，完成了需求的确认和签字，但我现在几乎无法记起任何关于需求的内容，对项目实施过程的某些环节却仍然记忆犹新。项目开始时，公司安排了两位工程师参与 ETL（Extract Transformation Load），一位是我，另一位与我现在仍是同事，我们当时使用了一个叫作 PowerMart 的工具，版本是 5.1。这就是我和 Informatica 的第一次亲密接触。当时 Informatica 的总代理也是曾经大名鼎鼎的 Sybase，据说我们的这个项目是 Informatica 进入中国后的第二个项目。

从此，我开始了自己漫长的 Informatica 之路。当时我所就职的公司敏锐地察觉到数据仓库/商业智能是未来的趋势之一，开始着手准备发展数据仓库方面的业务。当年有个著名的第三方调研机构，叫 IDC。我所就职的公司通过查阅 IDC 报告，发现 Informatica 是当时 ETL 市场份额最高的公司，于是果断决定采用这个工具作为自己的数据仓库的开发平台。当年公司的果断、决心，让我至今想起，仍非常钦佩。在 IT 人才严重短缺的年代，虽然年纪很轻、经验不足，但我还是作为经营分析项目的项目经理、技术经理等开始了自己的数据管理生涯。

此后一段时间，我在不经意间进入悲催的计费岁月，加班、加班、通宵、通宵……每个项目都以几年来计算，历经两个完整的移动计费项目，在此期间认识了很多好朋友和师长。这是我与 Informatica 断绝联系的几年，也是在技术方面拓展能力的几年。

我与 Informatica 的缘分并没有结束。有一天，原来的同事告诉我 Informatica 在招售前工程师，我就毅然决定去应聘，满足自己转向咨询领域的一点梦想，后来发现售前和咨询还是有所区别的，这是后话。因此，8 年前我加入了 Informatica 中国，开始成为一名专职的售前工程师。当年的 Informatica 只有这一个产品，人不是很多。我仍清晰地记着当时的

版本为 PowerCenter 8.1.1。现在的 Informatica 已经与早期差别非常大了，但是很多人还是习惯把 Informatica 的数据集成产品 PowerCenter 叫作 Informatica。

为什么要写一本关于 PowerCenter 的书呢？其实我内心里一直有这样的冲动。PowerCenter 是一个非常好的产品，在国内也有近千家用户，有大量的开发者和管理者，随着大数据的推广，还有大量的后继者会陆续开始 PowerCenter 之旅。一本中文材料会帮助所有的用户更加快速、更加全面地了解 PowerCenter，充分利用自己在 PowerCenter 方面的投资。这个冲动持续了很久，包括期间陆续说服数位同事参与，但是大家都有繁忙的本职工作，一次一次被耽搁。直到 2015 年春节前，内心的冲动促使我开始动笔了。前两天坐在我对面的北区销售总监说，我写这本书像写回忆录，想想确实有道理，我的自序也是回忆录的样子，希望大家能够谅解。

如何写好这本书？这也是我非常纠结的一个问题。什么样的深度？适合什么样的人群？是否有读者愿意来读？如何帮助读者了解 PowerCenter？和同事们讨论过很多次，似乎还是没有下定决心。当我下笔的时候，尤其是写到 50 多页的时候，我觉得我已经坚定了这本书的方向：不求全面，但求让读者快速地掌握 PowerCenter；不求精深，但求将最常用的功能展现给读者；不求华丽的词藻，但求读者能读懂。

PowerCenter 是什么？它是 ETL 工具。什么是 ETL？大数据及数据仓库 70%左右的工作都在做 ETL，在 Gartner 报告中它被划为 Data Integration 产品。Informatica 也曾定义自己是一家 Data Integration 公司。我是这样解释 ETL 工具的：它是 Data Integration 产品在数据仓库、大数据项目中的一个应用场景，它同时还有其他的应用场景，比如数据交换、数据安全，这些也是 PowerCenter 在后期的扩展。

希望本书能够成为分享我这些年掌握的 PowerCenter 相关知识的一个载体，成为初学者的入门教材，成为有经验者的开发人员的一本参考书。

杜绍森
2015 年 8 月

目 录

第 1 章
PowerCenter Hello World 世界

Hello World 程序是学习 IT 的入门术语，是初步了解一种技术的有效途径，因此本书也从 Hello World 开始。除了 PowerCenter Hello World，本书还提供了 Informatica Hello World 的个人版。

1.1 Informatica Hello World

Informatica 曾经是一个 ETL 工具的供应商，也是最好的 ETL 产品的供应商。在开始使用 ETL 工具，并加入到 Informatica 之后的很长一段时间内，我都认为 Informatica 是一家这样的公司。那时的 Informatica 只有一个产品，叫作 PowerCenter。那时，它的 Logo 下有一句话“The Data Integration Company”。

此后，“幸福的日子”结束了，Informatica 陆续有了很多新产品，包括除 PowerCenter 外的 Data Quality、Data Archive、Master Data Management、Data Masking、Ultra Message、Data Integration Hub 等，这时，其 Logo 下的“The Data Integration Company”显著的 IT 蓝依然没有变化。这时的我也在思考什么是 Data Integration，并有了自己对 Data Integration 的诠释。我将 Infomratica 所谓的 Data Integration 诠释为三类数据集成，即下游集成、中游集成和上游集成。

下游集成指的是数据仓库，这是典型的数据集成项目，它有一个显著特点：从数据流