Building a Data Warehouse

With Examples in SQL Server

Vincent Rainardi

Apress®

Building a Data Warehouse: With Examples in SQL Server

ISBN-13 (pbk): 978-1-59059-931-0

ISBN-10 (pbk): 1-59059-931-4

ISBN-13 (electronic): 978-1-4302-0527-2

ISBN-10 (electronic): 1-4302-0527-X

Printed and bound in the United States of America 9 8 7 6 5 4 3 2 1

Lead Editor: Jeffrey Pepper
Technical Reviewers: Bill Hamilton and Asif Sayed
Editorial Board: Steve Anglin, Ewan Buckingham, Tony Campbell, Gary Cornell, Jonathan Gennick, Jason Gilmore, Kevin Goff, Jonathan Hassell, Matthew Moodie, Joseph Ottinger, Jeffrey Pepper, Ben Renow-Clarke, Dominic Shakeshaft, Matt Wade, Tom Welsh
Senior Project Manager: Tracy Brown Collins
Copy Editor: Kim Wimpsett
Associate Production Director: Kari Brooks-Copony
Production Editor: Kelly Winquist
Compositor: Linda Weidemann, Wolf Creek Press
Proofreader: Linda Marousek
Indexer: Ron Strauss
Artist: April Milne
Cover Designer: Kurt Krames
Manufacturing Director: Tom Debolski

Distributed to the book trade worldwide by Springer-Verlag New York, Inc., 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax 201-348-4505, e-mail orders-ny@springer-sbm.com, or visit http://www.springeronline.com.

For information on translations, please contact Apress directly at 2855 Telegraph Avenue, Suite 600, Berkeley, CA 94705. Phone 510-549-5930, fax 510-549-5939, e-mail info@apress.com, or visit http://www.apress.com.

The source code for this book is available to readers at http://www.apress.com.

For my lovely wife, Ivana.

Contents at a Glance

Contents

About the Author

VINCENT RAINARDI is a data warehouse architect and developer with more than 12 years of experience in IT. He started working with data warehousing in 1996 when he was working for Accenture. He has been working with Microsoft SQL Server since 2000. He worked for Lastminute.com (part of the Travelocity group) until October 2007. He now works as a data warehousing consultant in London specializing in SQL Server. He is a member of The Data Warehousing Institute (TDWI) and regularly writes data warehousing articles for SQLServerCentral.com.

Preface

Friends and colleagues who want to start learning data warehousing sometimes ask me to recommend a practical book about the subject matter. They are not new to the database world; most of them are either DBAs or developers/consultants, but they have never built a data warehouse. They want a book that is practical and aimed at beginners, one that contains all the basic essentials. There are many data warehousing books on the market, but they usually cover a specialized topic such as clickstream, ETL, dimensional modeling, data mining, OLAP, or project management and therefore a beginner would need to buy five to six books to understand the complete spectrum of data warehousing. Other books cover multiple aspects, but they are not as practical as they need to be, targeting executives and project managers instead of DBAs and developers.

Because of that void, I took a pen (well, a laptop really) and spent a whole year writing in order to provide a practical, down-to-earth book containing all the essential subjects of building a data warehouse, with many examples and illustrations from projects that are easy to understand. The book can be used to build your first data warehouse straightaway; it covers all aspects of data warehousing, including approach, architecture, data modeling, ETL, data quality, and OLAP. I also describe some practical issues that I have encountered in my experience—issues that you'll also likely encounter in your first data warehousing project—along with the solutions.

It is not possible to show examples, code, and illustrations for all the different database platforms, so I had to choose a specific platform. Oracle and SQL Server provide complete end-to-end solutions including the database, ETL, reporting, and OLAP, and after discussions with my editor, we decided to base the examples on SQL Server 2005, while also making them applicable to future versions of SQL Server such as 2008. I apologize in advance that the examples do not run on SQL Server 2000; there is just too big a gap in terms of data warehousing facilities, such as SSIS, between 2000 and 2005.

Throughout this book, together we will be designing and building a data warehouse for a case study called Amadeus Entertainment. A data warehouse consist of many parts, such as the data model, physical databases, ETL, data quality, metadata, cube, application, and so on. In each chapter, I will cover each part one by one. I will cover the theory related to that part, and then I will show how to build that part for the case study. Specifically, Chapter 1 introduces what a data warehouse is and what the benefits are. In Chapters 2–6, we will design the architecture, define the requirements, and create the data model and physical databases, including the SQL Server configuration. In Chapters 7–10 we will populate the data stores using SSIS, as well as discuss data quality and metadata. Chapters 11–12 are about getting the data out by using Reporting Services and Analysis Services cubes. In Chapters 13–15, I'll discuss the application of data warehouse for BI and CRM as well as CDI, unstructured data, and search. I close the book with testing and administering a data warehouse in Chapters 16–17.

The supplementary material (available on the book's download page on the Apress web site, `http://www.apress.com`) provides all the necessary material to build the data warehouse for the case study. Specifically, it contains the following folders:

Scripts: Contains the scripts to build the source system and the data warehouse, as explained in Chapters 5 and 6.

Source system: Contains the source system databases required to build the data warehouse for the case study in Chapters 7 and 8.

ETL: Contains the SSIS packages to import data into the data warehouse. Chapters 7 and 8 explain how to build these packages.

Report: Contains the SSRS reports explained in Chapter 11.

Cubes: Contains the SSAS projects explained in Chapter 12.

Data: Contains the backup of data warehouse database (the DDS) and Analysis Services cube, which are used for reporting, OLAP, BI, and data mining in Chapters 11, 12, and 13.