Java Data Mining: Strategy, Standard, and Practice

The Morgan Kaufmann Series in Data Management Systems

Series Editor: Jim Gray, Microsoft Research

Java Data Mining: Strategy, Standard, and Practice
Mark Hornick, Erik Marcade and Sunil Venkayala

Joe Celko's Analytics and OLAP in SQL
Joe Celko

Data Preparation for Data Mining Using SAS
Mamdouh Refaat

Querying XML: XQuery, XPath, and SQL/XML in Context
Jim Melton and Stephen Buxton

Data Mining: Concepts and Techniques, Second Edition
Jiawei Han and Micheline Kamber

Database Modeling and Design: Logical Design, Fourth Edition
Toby J. Teorey, Sam S. Lightstone, and Thomas P. Nadeau

Foundations of Multidimensional and Metric Data Structures
Hanan Samet

Joe Celko's SQL for Smarties: Advanced SQL Programming, Third Edition
Joe Celko

Moving Objects Databases
Ralf Hartmut Güting and Markus Schneider

Joe Celko's SQL Programming Style
Joe Celko

Data Mining, Second Edition: Concepts and Techniques
Ian Witten and Eibe Frank

Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration
Earl Cox

Data Modeling Essentials, Third Edition
Graeme C. Simsion and Graham C. Witt

Location-Based Services
Jochen Schiller and Agnès Voisard

Database Modeling with Microsft® Visio for Enterprise Architects
Terry Halpin, Ken Evans, Patrick Hallock, Bill Maclean

Designing Data-Intensive Web Applications
Stephano Ceri, Piero Fraternali, Aldo Bongio, Marco Brambilla, Sara Comai, and Maristella Matera

Mining the Web: Discovering Knowledge from Hypertext Data
Soumen Chakrabarti

Advanced SQL: 1999—Understanding Object-Relational and Other Advanced Features
Jim Melton

Database Tuning: Principles, Experiments, and Troubleshooting Techniques
Dennis Shasha and Philippe Bonnet

SQL: 1999—Understanding Relational Language Components
Jim Melton and Alan R. Simon

Information Visualization in Data Mining and Knowledge Discovery
Edited by Usama Fayyad, Georges G. Grinstein, and Andreas Wierse

Transactional Information Systems: Theory, Algorithms, and Practice of Concurrency Control and Recovery
Gerhard Weikum and Gottfried Vossen

Spatial Databases: With Application to GIS
Philippe Rigaux, Michel Scholl, and Agnès Voisard

Information Modeling and Relational Databases: From Conceptual Analysis to Logical Design
Terry Halpin

Component Database Systems
Edited by Klaus R. Dittrich and Andreas Geppert

Managing Reference Data in Enterprise Databases: Binding Corporate Data to the Wider World
Malcolm Chisholm

Understanding SQL and Java Together: A Guide to SQLJ, JDBC, and Related Technologies
Jim Melton and Andrew Eisenberg

Database: Principles, Programming, and Performance, Second Edition
Patrick and Elizabeth O'Neil

The Object Data Standard: ODMG 3.0
Edited by R. G. G. Cattell and Douglas K. Barry

Data on the Web: From Relations to Semistructured Data and XML
Serge Abiteboul, Peter Buneman, and Dan Suciu

Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations
Ian Witten and Eibe Frank

Joe Celko's SQL for Smarties: Advanced SQL Programming, Second Edition
Joe Celko

Joe Celko's Data and Databases: Concepts in Practice
Joe Celko

Developing Time-Oriented Database Applications in SQL
Richard T. Snodgrass

Web Farming for the Data Warehouse
Richard D. Hackathorn

Management of Heterogeneous and Autonomous Database Systems
Edited by Ahmed Elmagarmid, Marek Rusinkiewicz, and Amit Sheth

Object-Relational DBMSs: Tracking the Next Great Wave, Second Edition
Michael Stonebraker and Paul Brown, with Dorothy Moore

A Complete Guide to DB2 Universal Database
Don Chamberlin

Universal Database Management: A Guide to Object/Relational Technology
Cynthia Maro Saracco

Readings in Database Systems, Third Edition
Edited by Michael Stonebraker and Joseph M. Hellerstein

Understanding SQL's Stored Procedures: A Complete Guide to SQL/PSM
Jim Melton

Principles of Multimedia Database Systems
V. S. Subrahmanian

Principles of Database Query Processing for Advanced Applications
Clement T. Yu and Weiyi Meng

Advanced Database Systems
Carlo Zaniolo, Stefano Ceri, Christos Faloutsos, Richard T. Snodgrass, V. S. Subrahmanian, and Roberto Zicari

Principles of Transaction Processing
Philip A. Bernstein and Eric Newcomer

Using the New DB2: IBM's Object-Relational Database System
Don Chamberlin

Distributed Algorithms
Nancy A. Lynch

Active Database Systems: Triggers and Rules For Advanced Database Processing
Edited by Jennifer Widom and Stefano Ceri

Migrating Legacy Systems: Gateways, Interfaces, & the Incremental Approach
Michael L. Brodie and Michael Stonebraker

Atomic Transactions
Nancy Lynch, Michael Merritt, William Weihl, and Alan Fekete

Query Processing for Advanced Database Systems
Edited by Johann Christoph Freytag, David Maier, and Gottfried Vossen

Transaction Processing: Concepts and Techniques
Jim Gray and Andreas Reuter

Building an Object-Oriented Database System: The Story of O_2
Edited by François Bancilhon, Claude Delobel, and Paris Kanellakis

Database Transaction Models for Advanced Applications
Edited by Ahmed K. Elmagarmid

A Guide to Developing Client/Server SQL Applications
Setrag Khoshafian, Arvola Chan, Anna Wong, and Harry K. T. Wong

The Benchmark Handbook for Database and Transaction Processing Systems, Second Edition
Edited by Jim Gray

Camelot and Avalon: A Distributed Transaction Facility
Edited by Jeffrey L. Eppinger, Lily B. Mummert, and Alfred Z. Spector

Readings in Object-Oriented Database Systems
Edited by Stanley B. Zdonik and David Maier

Java Data Mining: Strategy, Standard, and Practice

A Practical Guide for Architecture, Design, and Implementation

Mark F. Hornick
Erik Marcadé
Sunil Venkayala

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

MORGAN KAUFMANN PUBLISHERS

Publisher Diane D. Cerra
Publishing Services Manager George Morrison
Project Manager Marilyn E. Rash
Assistant Editor Asma Palmeiro
Cover Design Brian May, Maycreate LLC
Production Services Graphic World
Interior Printer The Maple Press Company
Cover Printer Phoenix Color Corp.

Morgan Kaufmann Publishers is an imprint of Elsevier.
500 Sansome Street, Suite 400, San Francisco, CA 94111

This book is printed on acid-free paper.

Library of Congress Cataloging-in-Publication Data
Application Submitted

ISBN 10: 0-12-370452-9
ISBN 13: 978-0-12-370452-8

For information on all Morgan Kaufmann publications,
visit our Web site at *www.mkp.com* or *www.books.elsevier.com*

Printed in the United States of America
06 07 08 09 10 10 9 8 7 6 5 4 3 2 1

To Suzanne, Amanda, and Tim for their enthusiasm and support. - M.H.

To Caroline, Laetitia and Guillaume. - E.M.

To my parents, wife Meera and daughter Shreya. - S.V.

Contents

Preface

The birth of a standard is an amazing event, highlighting the ability of individuals from vastly different and often competing companies to come together to design an interface for a domain such as data mining. For JSR-73, we drew on experts from data mining tool and application vendors, as well as users of data mining technology. Data mining, as a field, is remarkably diverse in scope, encompassing capabilities from a broad range of disciplines: artificial intelligence, machine learning, statistics, data analysis, and visualization. Producing a standard in such a space is a challenging and fascinating adventure.

Within a year or so of embarking on the JDM 1.0 standard, various expert group members suggested that we'd have to write a book about Java Data Mining (JDM) someday. And indeed, here we are. Our main motivation for writing this book is to introduce data mining to a much broader audience, one that may have never used or encountered data mining before. As such, we focus less on the technical and scholarly details of data mining than on its practical understanding and application. We have tried to include a reasonably broad set of references for individuals who want to dive down to the next level of detail. However, we have strived to make data mining concepts, process, and use through JDM more accessible to Java developers, who usually do not encounter data mining, and the colleagues they will work with to develop advanced analytic applications.

Advanced analytic applications—those augmented with advanced and predictive analytics such as data mining—provide greater business intelligence, yielding insight into business problems and guidance for improved decision making. Such applications are becoming most valuable to businesses, and hence can increase revenue and profits—both for the vendors who sell them and for the businesses that use them.

Readers of this book will find a somewhat unconventional approach to data mining. Other books on data mining provide much detail on algorithms and techniques. Although this information is important to those studying machine learning or wanting to become a data analyst, other potential users of data mining are left wondering how these algorithms or techniques will be applied to solve problems. As vendors of data mining technology strive to make data mining more accessible to a broader range of users, such as business analysts, information technology (IT) specialists, and database administrators (DBAs), it is no longer the details that users require, but the big picture. Users ask, "How can I use this powerful technology to provide value within my business?" In this book, we strive to approach this and other questions from several perspectives: the software developer, the software and systems architect, and the business and data analyst. We explore these perspectives in the following section, "Guide to Readers."

In this book, we provide insight into three key aspects of the Java Data Mining standard. The first aspect, covered in Part I, focuses on strategies for solving data mining–related business and scientific problems, and on the strategy the JDM Expert Group pursued in the design of the JDM standard. After an introduction to the data mining field, we discuss solving problems in various industries using data mining technology.

Although every industry has unique problems to solve, requiring custom and innovative solutions, each industry also shares many problems that can benefit from cross-industry solutions. For example, industries such as retail, financial services, and healthcare, as well as the public sector, all have customers. The cross-industry solution spaces include customer acquisition, customer retention, customer lifetime value, and targeted marketing.

Because data mining solutions typically do not take form or produce value in a vacuum, we then discuss the overall process, based on the industry standard data mining process CRISP-DM. Because users of data mining technology need to be minimally conversant in the terminology and concepts to problem solve with their colleagues, we introduce the mining functions and algorithms defined in JDM. With this foundation, we explore the JDM strategy, answering questions such as: What drove the design of JDM? What is the role of standards? Lastly, before embarking on details of the Java Data Mining standard, we provide a "getting started" code example that follows the CRISP-DM process.

The second aspect, covered in Part II, focuses on the standard itself. This part introduces various concepts defined by or assimilated into the standard using examples based on business problems. After this, we explore the design of the JDM API and more detailed code examples to give readers a better understanding of how to use JDM to build applications and solve problems. Although JDM is foremost dedicated to being a standard Java language API, Java Data Mining also defines an XML schema representation for JDM objects, as well as a web services interface to enable the use of JDM functionality in a Services Oriented Architecture (SOA) environment. Part II also discusses these with specific examples of their use.

The third aspect, covered in Part III, focuses on using JDM in practice, building applications and tools that use the Java Data Mining API. We begin this part with several business scenarios (e.g., targeted marketing, key factor analysis, and customer segmentation). Because JDM is designed to be used by both application designers and data mining tool designers, we introduce code for building a simple tool graphical user interface (GUI), which manipulates JDM-persistent objects as well as enables the building and testing of a model. Having introduced web services in Part II, we give an example of a web services based application. Since data mining can impact the Information Technology (IT) infrastructure of most companies, we explore the impact of data mining along several dimensions, including hardware, software, data access, performance, and administration tools. Since the practice of using data mining often involves the use of commercial implementations, we introduce two such JDM implementations, from Oracle and KXEN. We also provide some guidelines or insights for implementers new to JDM.

Wrapping up in Part IV, we explore the evolution of data mining standards, which puts JDM in the broader context of other data mining standards. We also contrast the approaches taken by various data mining standards bodies. Since we note that no standard is ever complete, and JDM 1.1 itself covers only a subset of the possible data mining functions and algorithms, we highlight directions for JDM 2.0. We introduce features under consideration such as transformations, time series, and apply for association models, among others.

Acknowledgments

We first want to acknowledge the Java Data Mining expert group members who participated in the long process required to produce the JSR-73 standard. Their unwavering support through weekly conference calls and face-to-face meetings over the 4 years of the standards development is greatly appreciated. We also acknowledge the additional contributions of Hankil Yoon, Ka Kit Chan, Jim Dadashev, and Somesh Marepalli to the Technology Compatibility Kit (TCK) implementation, and Marwane Jai Lamimi to the Reference Implementation (RI).

We are very grateful for the general and specialist input provided by Frank Byrum, Jim Melton, and Osmar Zaiane on the developing manuscript. Over the past year, their detailed comments on both structure and content were a tremendous asset. We thank Jacek Myczkowski and Don Deutsch for their valuable comments on the final manuscript, as well as their support of the standards efforts for JSR-73 and JSR-247 at Oracle. We thank the JDM expert group members Michal Prussak, Alex Russakovsky, and Michael Smith who also provided valuable comments on the final manuscript, and David Urena and Samy Mechiri who contributed to the source code used in Part III of this book.

Of course, all remaining errors (which we expect exist despite careful review) are entirely our responsibility.

We offer our appreciation and gratitude to the wonderful people at Morgan Kaufmann Publishers as they guided us through the process of book writing and publishing. We thank Jim Melton, one of our reviewers, for putting us in contact with Diane Cerra, our talented and patient publisher, to begin this journey. We thank Diane, Asma Palmeiro, Misty Bergeron, Marilyn Rash, and Bruce Siebert who worked to make this book possible.

Guide to Readers

Data mining is becoming a mainstream technology used in business intelligence applications supporting industries such as financial services, retail, healthcare, telecommunications, and higher education, and lines of business such as marketing, manufacturing, customer experiences, customer service, and sales. Many of the business problems that data mining can solve cut across industries such as customer retention and acquisition, cross-sell, and response modeling. Due to the cost, skillsets, and complexity required to bring data mining results into an established business process, early adopters were typically big companies and research labs with correspondingly large budgets and access to statisticians and machine learning experts. In recent years, however, data mining products have simplified data mining considerably by automating the process—making the fruits of the technology more widely accessible. New algorithms and heuristics have evolved to provide good results with little or no experimentation or data preparation. In addition, the availability of data mining has increased with in-database data mining capabilities.

Java Data Mining (JDM) furthers the adoption of data mining by providing a standard Java and web services Application Programming Interface (API) for data mining. This book introduces data mining to software developers and application architects who may have heard of the benefits of data mining but are unsure how to realize these benefits. This book is also targeted at business and data analysts who want to learn how JDM helps in developing vendor-neutral data mining solutions. It does not require a reader to be familiar with data mining, statistics, or machine learning technologies.

We have organized this book into three main parts: *strategy, standard*, and *practice*. In Part I, JDM Strategy, we introduce data mining in general, uses of data mining in solving industry problems, data mining processes and techniques, the role of data mining standards, and a high-level introduction to the JDM Application Programming Interface (API). Most of this part doesn't require the reader to know the Java language.

In Part II, JDM Standard, we explain the concepts used in JDM by example, explore the JDM API design and its usage, and introduce the Java Data Mining XML schema and web services. This part requires readers to know the Java language, XML, and XML schema. It gives a brief introduction to web services in Chapter 11 before discussing the JDM web services.

In Part III, JDM Practice, we illustrate practical problem solving using the JDM API. We begin by developing a sample data mining tool using JDM and a sample data mining web service using JDM. We then introduce two JDM vendor implementations, exploring their functionality, architecture, and design tradeoffs before giving some guidance to others interested in implementing a JDM-compliant system.

In Part IV, Wrapping Up, we discuss the evolution of data mining standards, where they have been and where they might go. We give a preview of some of the features proposed for JDM 2.0.

For the Software Developer

For software developers, and in particular Java and web services developers, this book introduces data mining and how to use JDM to develop data mining solutions. Part I introduces data mining and various types of business problems that can be solved using data mining, illustrates a standard process used to conduct a data mining project, describes data mining techniques used to solve business problems, explains the JDM standard strategy and why the JDM standard is necessary, and provides an overview of the JDM API. Even though software developers are not typically involved in the initial solving of a data mining problem, it is important to know about concepts to understand the JDM API and how to develop data mining solutions.

Part II will familiarize developers with JDM concepts and the API. Readers of this part are required to know the Java language, Object Oriented Programming, the Unified Modeling Language and XML to understand the Java examples, API design concepts, JDM XML schema, and web services. This part introduces JDM concepts using examples, describes the design and usage of the Java API, and illustrates the Java Data Mining XML schema and web services interfaces.

Part III describes the use of the JDM API in practice with sample applications and detailed code examples both for the Java and web services API. It also provides JDM vendor implementation details and explains the process for other data mining vendors in adopting the JDM standard.

After reading this book, we expect the data mining knowledge gap between developers and data analysts will be greatly reduced to help them communicate more effectively when developing a data mining solution.

For the Software Architect

Data mining is often integrated with existing software applications and business processes. Understanding of data mining processes provides greater insight for architects to enable this technology in existing or new applications. For example, an architect needs to understand how data mining works to add intelligent customer offers using data mining to an existing call center application.

For architects who want to be hands-on with the JDM API (e.g., to develop prototypes), all parts of this book are useful. Part I and Part III are particularly useful for architects. Part I introduces data mining in general and provides examples of how it is currently being applied to solve business problems. Most important, it introduces the data mining process and the role of the information technology department in implementing a data mining project.

Part II will be useful to understand the API-level concepts for the architects who want to be hands-on with the API, to develop prototypes, or to mentor the developers about the use of the API.

In Part III, we provide deeper insight into how JDM can be used in practice. Chapter 16, which discusses vendor implementations, is particularly useful for data mining software architects who are interested in developing JDM compatible API's and extensions.

After reading this book, architects should be comfortable in integrating JDM-based data mining solutions with their applications and be able to develop a strategy to operationalize data mining results with their existing applications.

For the Business/Data Analyst

For business and data analysts who want to extract actionable information from corporate data, this book provides an introduction to data mining and how it is used to solve various business problems across industries. In Part I, the data mining usage scenarios and process of implementing a data mining project will be particularly useful for the analyst unfamiliar with data mining. Chapter 5, JDM Strategy, enables analysts to understand why the JDM standard is important in implementing a data mining solution. Typically, analysts are not involved in the software implementation of the solution, yet Part II may be useful for understanding the data mining concepts used by JDM to facilitate communication with developers and data mining experts, and for using tools based on JDM.

For an analyst who is already familiar with data mining and who has expertise in data mining and statistics, this book gives details of Java Data Mining and its usage in developing data mining solutions. Data mining tools can often generate JDM-compatible code to easily deploy a solution to a JDM-compatible Data Mining Engine (DME).

After reading this book, an analyst previously unfamiliar with data mining should be able to better understand how data mining can help in solving business problems. A data mining expert analyst will be able to understand the supported data mining features in JDM and be able to communicate easily with the software architects/developers to implement a data mining solution.