

Data Warehouse and Data Mining

Concepts, techniques and real life applications

Dr. Jugnesh Kumar



www.bpbonline.com

Copyright © 2024 BPB Online

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor BPB Online or its dealers and distributors, will be held liable for any damages caused or alleged to have been caused directly or indirectly by this book.

BPB Online has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, BPB Online cannot guarantee the accuracy of this information.

First published: 2024

Published by BPB Online

WeWork

119 Marylebone Road

London NW1 5PU

UK | UAE | INDIA | SINGAPORE

ISBN 978-93-55517-340

www.bpbonline.com

Dedicated to

My beloved wife Varsha and my son Nikhil Kumar

About the Author

Dr. Jugnesh Kumar is working as a professor in the Computer Science and Engineering department at Echelon Institute of Technology, Faridabad. He has more than 18 years of teaching experience. He holds an M.Tech and PhD in computer science and engineering. He has successfully organized international and national-level conferences. He has published more than 35 articles in Scopus-indexed, SCI-high-impact journals and international conferences.

About the Reviewer

Amitava Nandi is a Senior IT professional with a profound expertise in the field of data engineering and data analytics. He is renowned for his ability to deliver cutting-edge and optimal data solutions in big data engineering and data warehousing using cloud platforms like Azure, AWS. He has an impressive track record of working with various clients and partners, including industry giants such as Siemens Healthineers, TataNue, FedEx, Virgin Atlantic Airlines, and BJ's Retail.

Acknowledgement

I want to express my deepest gratitude to my family and friends for their unwavering support and encouragement throughout this book's writing, especially my friend Dr Anubhav Kumar and Dr Rinku Kumar.

I am also grateful to BPB Publications for their guidance and expertise in bringing this book to fruition. It was a long journey of revising this book, with valuable participation and collaboration of reviewers, technical experts, and editors.

I would also like to acknowledge the valuable contributions of my colleagues and co-worker during many years working in the tech industry, who have taught me so much and provided valuable feedback on my work.

Finally, I would like to thank all the readers who have taken an interest in my book and for their support in making it a reality. Your encouragement has been invaluable.

Preface

Welcome to the dynamic world of data warehousing and data mining! In an era where information is revered as the new gold, the ability to harness, manage, and extract insights from vast repositories of data has become indispensable.

This book is an exploration into the realms of data warehousing and data mining, designed to be a comprehensive guide for both beginners and seasoned professionals. It delves into the fundamental principles, methodologies, and advanced techniques essential for understanding, building, and leveraging robust data infrastructure and extracting valuable knowledge from it.

Data Warehousing introduces the foundational concepts behind the creation and management of centralized repositories, offering a blueprint for designing efficient data storage systems. It covers the intricacies of schema design, extraction-transformation-loading (ETL) processes, and optimization strategies, providing a solid groundwork for constructing data warehouses tailored to diverse business needs.

Data Mining on the other hand, navigates the terrain of extracting meaningful patterns, trends, and associations from extensive datasets. It illuminates the various algorithms, statistical techniques, and machine learning methodologies used to unearth hidden insights, empowering practitioners to derive actionable intelligence and make informed decisions.

Chapter 1: Introduction to Data Warehousing - The key objective of this chapter on Data Warehousing is to provide readers with a comprehensive understanding of the fundamental aspects and components of data warehousing. It aims to define data warehousing and its purpose, explore the differences between database management systems and Data Warehouses, introduce the concept of data marts, emphasize the significance of metadata, and explain the multidimensional data model and various schema designs.

Chapter 2: Data Warehouse Process and Architecture - Data warehouse architecture are designed to efficiently store, manage, and analyze large volumes of data collected from various sources. These architectures aim to provide a structured framework for organizing and processing data in a way that facilitates business intelligence, reporting and data analysis.

Chapter 3: Data Warehouse Implementation - Data warehouse implementation involves a sequential set of tasks critical for building a functional data warehouse based on the

client's requirements. It encompasses the phases of planning, data gathering, data analysis, and business actions, all of which contribute to the successful implementation of the data warehouse.

Chapter 4: Data Mining Definition and Task - Data mining is the process of discovering patterns, relationships, and insights from large volumes of data. It involves applying various techniques and algorithms to extract meaningful and actionable information from datasets. The goal of data mining is to uncover hidden patterns, identify trends, and make predictions or decisions based on the discovered knowledge.

Chapter 5: Data Mining Query Languages - The DBMiner data mining system introduced the Data Mining Query Language (DMQL), which is derived from the widely used Structured Query Language (SQL). DMQL is designed to facilitate ad hoc and interactive data mining by providing specific commands for defining primitives. It can be applied to both databases and data warehouses, making it a versatile language for data mining tasks.

Chapter 6: Data Mining Techniques - Data mining techniques are methods and processes used to discover patterns, relationships, anomalies, and valuable insights from large datasets. These techniques are applied to extract useful information and knowledge from data, and they play a crucial role in various fields, including business, healthcare, finance, and scientific research. Data mining techniques are selected based on the specific problem, dataset, and desired outcomes

Chapter 7: Mining Complex Data Objects - Mining complex data objects refers to the process of discovering valuable patterns, structures, and insights within datasets that contain intricate and multi-dimensional data objects. These complex data objects can take various forms, such as images, text documents, time series, graphs, or any other data type that exhibits intricate relationships and properties. The goal of mining complex data objects is to extract meaningful knowledge from these diverse and often unstructured data sources.

Coloured Images

Please follow the link to download the
Coloured Images of the book:

<https://rebrand.ly/1ca64a>

We have code bundles from our rich catalogue of books and videos available at <https://github.com/bpbpublications>. Check them out!

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

Did you know that BPB offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.bpbonline.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at :

business@bpbonline.com for more details.

At www.bpbonline.com, you can also read a collection of free technical articles, sign up for a range of free newsletters, and receive exclusive discounts and offers on BPB books and eBooks.

Piracy

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at **business@bpbonline.com** with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit **www.bpbonline.com**. We have worked with thousands of developers and tech professionals, just like you, to help them share their insights with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions. We at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit **www.bpbonline.com**.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

<https://discord.bpbonline.com>



Table of Contents

1. Introduction to Data Warehousing	1
Introduction	1
Structure	1
Objectives	2
Data warehousing	2
History of data warehouse	3
<i>Decision support system development (1970s–1980s)</i>	<i>4</i>
<i>Online analytical processing was first introduced in the 1980s</i>	<i>4</i>
<i>Concepts for data warehousing in the 1980s and 1990s</i>	<i>5</i>
<i>Ralph Kimball’s dimensional modelling first appeared in the 1990s</i>	<i>5</i>
<i>ETL and data integration advancements between the 1990s and 2000s</i>	<i>5</i>
<i>Columnar databases and in-memory computing adoption between 2000 and 2010</i>	<i>5</i>
<i>Big Data and Cloud data warehousing (2010s–present)</i>	<i>5</i>
Data warehouse works	6
Sources of data	6
Extract, Transform, and Load	6
Data organization and storage	7
Types of data warehouses	8
Enterprise data warehouse	8
Data mart	9
Operational data store	13
General stages of data warehouse	14
Analysis of requirements	15
Modelling data	15
Extraction of data	15
Data loading	16
Transformation of data	16
Data management and storage	16
Management of metadata	16

<i>Analysis and querying</i>	16
<i>Upkeep and evolution</i>	16
<i>Need of data warehouse</i>	18
Uses and trends of data warehouse.....	19
<i>Data warehouse applications in various industries</i>	21
<i>Banking</i>	21
<i>Retail</i>	22
<i>Healthcare</i>	22
<i>Marketing</i>	22
<i>Manufacturing</i>	22
<i>Telecommunications</i>	23
<i>Trends of data warehouse</i>	23
Database management system versus data warehouse	24
Metadata.....	26
<i>Types of metadata</i>	29
<i>Descriptive metadata</i>	29
<i>Administrative metadata</i>	30
<i>Structural metadata</i>	31
<i>Technical metadata</i>	32
<i>Provenance metadata</i>	33
<i>Rights metadata</i>	34
<i>Preservation metadata</i>	34
<i>Role of metadata</i>	36
<i>Metadata repository</i>	37
<i>Benefits of metadata</i>	39
<i>Challenges for metadata management</i>	40
Multidimensional data model.....	42
Data cubes	43
<i>Data cube classification</i>	44
<i>Operations in data cubes</i>	45
<i>Advantages of data cubes</i>	46
<i>Data cube disadvantages</i>	47

Schemas for multidimensional database	47
<i>Star schema</i>	47
<i>Snowflake schema</i>	48
<i>Star Schema</i>	48
<i>Fact table</i>	50
<i>Snowflake schema</i>	52
<i>Fact table</i>	53
<i>Fact constellation in data warehouse</i>	56
<i>General structure of fact constellation</i>	58
<i>Fact constellation schema architecture</i>	58
Conclusion	60
Exercises	60
2. Data Warehouse Process and Architecture	63
Introduction	63
Structure	63
Objectives	64
Objectives of data warehouse architecture.....	64
<i>Single-tier architecture</i>	67
<i>Two-tier architecture</i>	67
<i>Three-tier architecture</i>	68
<i>Properties of data warehouse architecture</i>	69
Types of data warehouse architecture.....	71
<i>Single-tier architecture</i>	71
<i>Two-tier architecture</i>	71
<i>Three-tier architecture</i>	73
Advantages of data warehouse architecture.....	74
Disadvantages of data warehouse architecture.....	74
Data warehouse database	75
Online transaction processing and online analytical processing	76
<i>Online transaction processing</i>	76
<i>Online analytical processing</i>	76
<i>Advantages and disadvantages of Online transaction processing</i> <i>and online analytical processing</i>	78

<i>Types of online analytical processing</i>	79
<i>Examples of all types, advantages and disadvantages of different OLAP systems</i>	80
Servers	83
Data warehouse manager	84
Conclusion	85
Exercises	85
3. Data Warehouse Implementation	87
Introduction	87
Structure	88
Objectives	88
Introduction of data warehouse implementation	88
Data warehouse implementation guidelines	91
Components of data warehouse implementation	93
Benefits of implementing a data warehouse	94
Computation of data cubes.....	95
<i>Subtracting two cubes for differences</i>	98
Modeling Online Analytical Processing data.....	98
Online Analytical Processing queries manager	100
Data warehouse back-end tools	101
Complex aggregation at multiple granularities	103
Tuning and testing of data warehouse.....	103
<i>Tuning and testing of data warehouse: Enhancing performance and reliability</i>	103
Conclusion	106
Exercises	106
4. Data Mining Definition and Task	109
Introduction	109
Structure	109
Objectives	110
Introduction to data mining	110
<i>Defining business objectives/problem definition</i>	110
<i>Importance</i>	110
<i>Stakeholder meetings</i>	110

<i>Problem definition</i>	110
<i>Prioritization</i>	111
<i>Scope definition</i>	111
<i>Deliverables</i>	111
<i>Benefits</i>	111
Data mining tasks	112
<i>Data mining versus data analysis</i>	114
Data mining functionality	115
Knowledge Discovery in Databases versus data mining	117
Data mining techniques	119
<i>Classification</i>	119
<i>Clustering</i>	121
<i>Association rule mining</i>	122
<i>Regression analysis</i>	124
<i>Correlation versus causation</i>	125
<i>Anomaly detection</i>	126
<i>Global outliers</i>	126
<i>Contextual outliers</i>	126
<i>Collective outliers</i>	127
<i>Sequence pattern mining</i>	128
Text mining	129
Data mining tools and applications	131
Conclusion	132
Exercises	132
5. Data Mining Query Languages	133
Introduction	133
Structure	133
Objectives	134
Introduction to data mining query languages	134
Syntax of Data Mining Query Languages	135
<i>Structured Query Language</i>	135
<i>GROUP BY</i>	136

ORDER BY.....	136
WINDOW operations (using OVER).....	136
Creating a view as part of data preprocessing	136
Various clauses to explore the attributes or dimensions	137
Data Mining Query Language	138
Multidimensional Expressions.....	138
MDX query clauses	139
MDX functions.....	139
Datalog	140
XQuery.....	140
XPath and XQuery.....	140
CSS selectors	140
Web scraping libraries.....	140
JSONPath	140
SPARQL.....	141
RESTful APIs	141
Web Scraping Frameworks	141
Graph Query Languages.....	141
Standardization of Data Mining Languages	141
Data specification.....	142
DMQL for characterization.....	143
DMQL for discrimination.....	143
Specifying knowledge	144
Hierarchy specification	144
Pattern presentation and visualization specification	145
Data mining languages and standardization of data mining	146
Conclusion	147
Exercises	148
6. Data Mining Techniques.....	149
Introduction	149
Structure	149
Objectives	150

Data mining techniques	150
Association rules	151
<i>Mathematical explanation of association rule and all of these components</i>	152
<i>Types of association rules in data mining</i>	153
<i>Algorithms for generating association rules in data mining</i>	154
Clustering techniques	154
<i>Types of clustering</i>	155
<i>Clustering methods</i>	156
<i>Partitioning method</i>	156
<i>Hierarchical methods</i>	157
<i>Density-based method</i>	157
<i>Grid-based method</i>	159
<i>Model-based methods</i>	160
<i>Constraint-based method</i>	161
Decision tree	161
Rough sets	162
Support vector machines and fuzzy techniques	163
<i>Support vector machines</i>	163
<i>Hyperplane</i>	165
<i>Support vectors</i>	165
<i>Margin</i>	165
<i>Fuzzy techniques</i>	166
Conclusion	167
Exercises	167
7. Mining Complex Data Objects	169
Introduction	169
Structure	169
Objectives	170
Mining complex data objects	170
<i>Time-series data mining</i>	170
<i>Sequential pattern mining in symbolic sequences</i>	170
<i>Data mining of biological sequences</i>	170

Graph pattern mining	170
Statistical modeling of networks	170
Mining spatial data	171
Mining cyber-physical system data.....	171
Mining multimedia data	171
Mining web data	171
Mining text data.....	171
Mining spatiotemporal data.....	171
Mining data streams	172
Spatial databases	172
Multimedia databases	174
Multimedia databases and temporal data.....	175
Difference between spatial and temporal data.....	175
Different types of multimedia applications	177
Challenges with multimedia database	177
Architecture for multimedia data mining	179
Applications of multimedia mining	180
Time series and sequence data mining	180
Components of time series	180
Categories of time series movements	181
Text databases and mining	182
Word Wide Web.....	183
Streaming data processing: An in-depth exploration.....	184
Types of windows for aggregating streaming data.....	185
Conclusion	185
Exercises	186
Index	187-196

CHAPTER 1

Introduction to Data Warehousing

Introduction

Data warehousing is a central component of business intelligence, focusing on the collection, storage, and analysis of data from various sources within an organization. It involves the process of gathering and managing large volumes of structured and unstructured data to support decision-making processes.

Structure

This chapter will cover the following topics:

- Data warehousing
- History of data warehouse
- Data warehousing works
- Types of data warehousing
- Uses and trends of data warehouse
- Database management system versus data warehouse
- Metadata

- Multidimensional data model
- Data cubes
- Schemas for multidimensional database

Objectives

The key objective of this chapter on data warehousing is to provide readers with a comprehensive understanding of the fundamental aspects and components of data warehousing. It aims to define data warehousing and its purpose, explore the differences between database management systems and Data Warehouses, introduce the concept of data marts, emphasize the significance of metadata, and explain the multidimensional data model and various schema designs. Ultimately, the chapter seeks to equip readers with the knowledge necessary to grasp the essential concepts and trends in data warehousing.

Data warehousing

A data warehouse acts as a safe electronic storage facility for the information that companies and organizations keep. Its main goal is to compile a priceless archive of old data that can be accessed and examined to learn insightful things about how the company operates. A data warehouse is regarded as a key component of business intelligence and is a part of the larger information infrastructure used by contemporary businesses. They can effectively assess their past successes and failings, which helps them make well-informed decisions about their future endeavors and architected data warehousing, as shown in *Figure 1.1*:

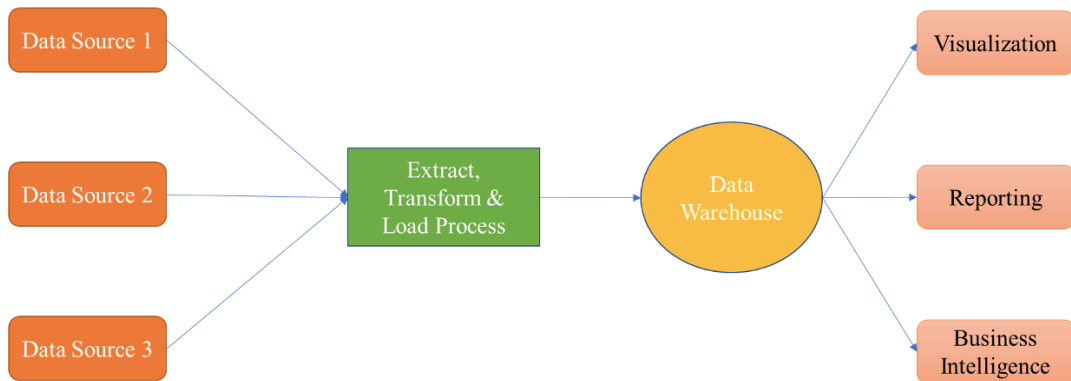


Figure 1.1: The architecture of data warehousing

Large amounts of structured, semi-structured, and unstructured data from various sources within an organization are kept in a centralized, integrated repository known as a Data Warehouse. It is specially made to support the organization's reporting and analytical needs. A data warehouse's main goal is to give decision-makers a consolidated and

historical view of the data, making it simpler for them to access and analyze the data for business intelligence and reporting needs. It differs from operational databases, which are designed for everyday use and transactional processing. A data warehouse, on the other hand, concentrates on long-term data storage and analysis.

The data warehouse system is also known by the following name, as mentioned in *Figure 1.2*:



Figure 1.2: Various names of data warehouse

History of data warehouse

Data warehousing's beginnings can be traced back to the 1980s when companies began to struggle with the management and analysis of massive amounts of data produced by their operational systems. Here is a history of data warehouse , as shown in *Figure 1.3*: