Data Warehouse and Data Mining

Concepts, techniques and real life applications

Dr. Jugnesh Kumar

www.bpbonline.com

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.

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

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*:

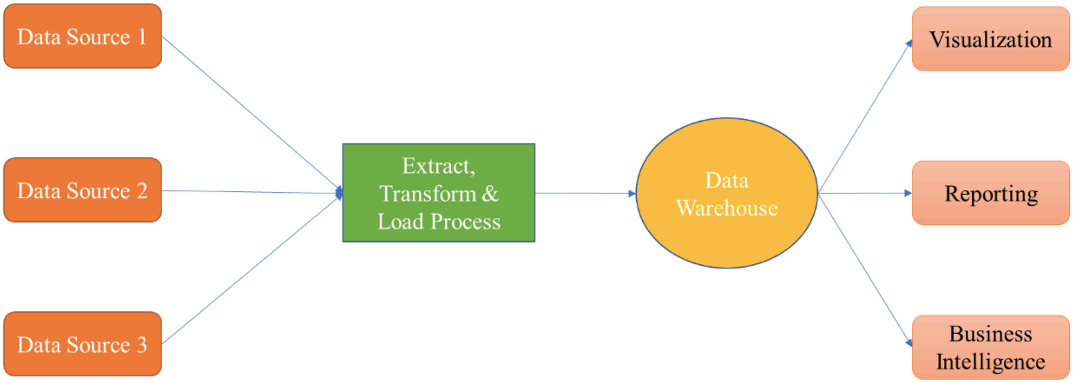

Figure 1.1: *The architecture of data warehousing*

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*:

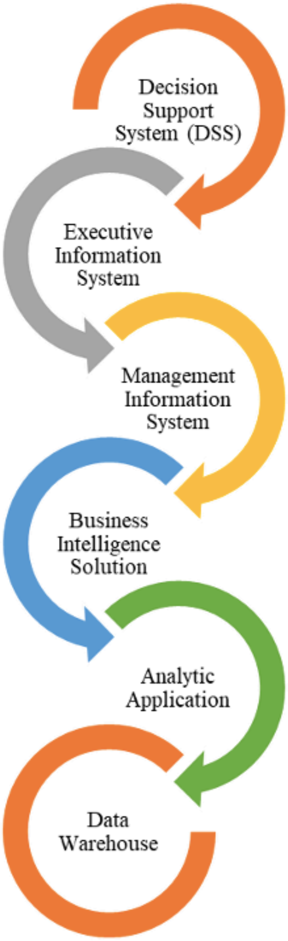
Decision Support System (DSS)
Executive Information System
Management Information System
Business Intelligence Solution
Analytic Application
Data Warehouse

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*: