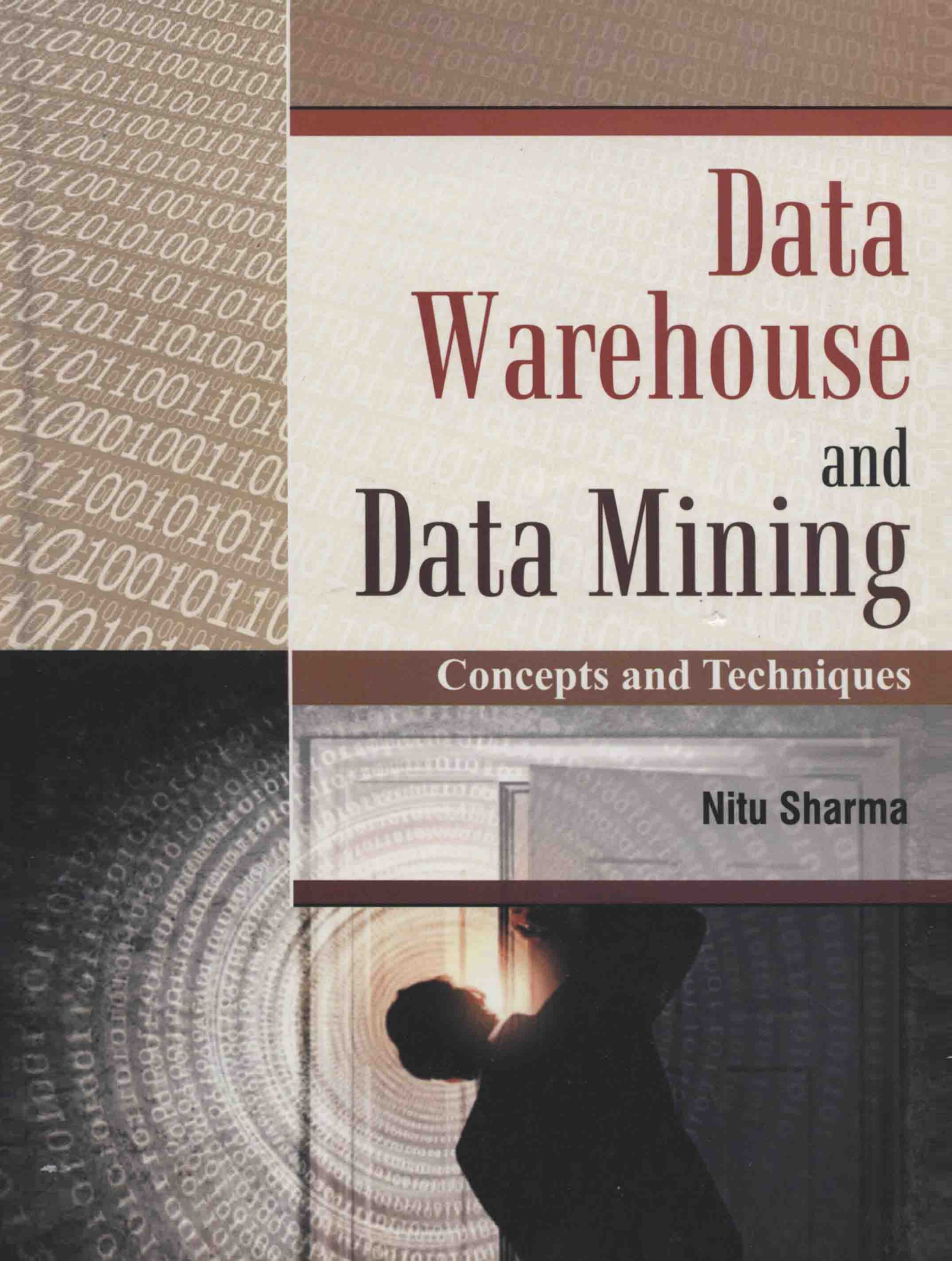
Data Warehouse and Data Mining

Concepts and Techniques

Nitu Sharma



Dedicated to:

My Nation

&

My Parents

(*Mr. Rakesh Kr. Sharma & Ms. Shashi Sharma*)

GLOBAL VISION PUBLISHING HOUSE
F-4, 1st Floor, 'Hari Sadan' 20, Ansari Road,
Daryaganj, New Delhi-110002 (INDIA)
Tel.: 23261581, 23276291, 43037885 Mob: 9810644769
Email: nsingh_2004@vsnl.net, info@globalvisionpub.com
Website: www.globalvisionpub.com

Data Warehouse and Data Mining: Concepts and Techniques

First Edition 2013

ISBN: 978-81-8220-592-5

PRINTED IN INDIA

Published by Dr. N.K. Singh for Global Vision Publishing House, New Delhi-2 and *Printed at* G. S. Offset, Naveen Shahdara, Delhi-32.

Data Warehouse and Data Mining

Concepts and Techniques

CONTENTS

1

HISTORY OF DATA WAREHOUSING AND MINING

CHAPTER OVERVIEW

1.1. INTRODUCTION

This chapter deals with the history and function of data warehouse, and discuss about the absence of a data warehousing architecture, and various approach for storage of data like normalized and dimensional approach. Various design methodology like top-down, bottom-up, and hybrid design has also been explained. This chapter also includes the various type of field where we can use different mining like sensor, pattern, in science and engineering etc.

1.2. HISTORY OF DATA WAREHOUSING

The concept of data warehousing dates back to the late 1980s when IBM researchers Barry Devlin and Paul Murphy developed the "business data warehouse". In essence, the data warehousing concept was intended to provide an architectural model for the flow of data from operational systems to decision support environments. The concept attempted to address the various problems associated with this flow-mainly, the high costs associated with it. In the absence of a data warehousing architecture, an enormous amount of redundancy was required to support multiple decision support environments. In larger corporations it was typical for multiple decision support environments to operate independently. Each environment served different users but often required much of the same data. The process of gathering, cleaning and integrating data from various sources, usually long existing operational systems (usually referred to as legacy systems), was typically in part replicated for each environment. Moreover, the operational systems were frequently re-examined as new decision support requirements emerged. Often new requirements necessitated gathering, cleaning and integrating new data from "data marts" that were tailored for ready access by users.

Many large businesses found themselves with data scattered across multiple platforms and variations of technology, making it almost impossible for any one individual to use data from multiple sources. A key idea within data warehousing is to take data from multiple platforms/technologies (As varied as spreadsheets, DB2 databases, IDMS records, and VSAM files) and place them in a common location that uses a common querying tool. In this way operational databases could be held on whatever system was most efficient for the operational business, while the reporting/ strategic information could be held in a common location using a common language.

Key developments in early years of data warehousing were:

- 1960— General Mills and Dartmouth College, in a joint research project, develop the terms dimensions and facts.
- 1970— ACNielsen and IRI provide dimensional data marts for retail sales.
- 1970— Bill Inmon begins to define and discuss the term: Data Warehouse
- 1975— Sperry Univac Introduce MAPPER (Maintain, Prepare, and Produce Executive Reports) is a database management and reporting system that includes the world's first 4GL. It was the first platform specifically designed for building Information Centers (a forerunner of contemporary Enterprise Data Warehousing platforms)
- 1983— Teradata introduces a database management system specifically designed for decision support.
- 1983— Sperry Corporation Martyn Richard Jones defines the Sperry Information Center approach, which while not being a true DW in the Inmon sense, did contain

many of the characteristics of DW structures and process as defined previously by Inmon, and later by Devlin. First used at the TSB England and Wales

- 1984— Metaphor Computer Systems, founded by David Liddle and Don Massaro, releases Data Interpretation System (DIS). DIS was a hardware/software package and GUI for business users to create a database management and analytic system.
- 1988— Barry Devlin and Paul Murphy publish the article An architecture for a business and information system in IBM Systems Journal where they introduce the term "business data warehouse".
- 1990— Red Brick Systems, founded by Ralph Kimball, introduces Red Brick Warehouse, a database management system specifically for data warehousing.
- 1991— Prism Solutions, founded by Bill Inmon, introduces Prism Warehouse Manager, software for developing a data warehouse.
- 1992— Bill Inmon publishes the book Building the Data Warehouse.
- 1995— The Data Warehousing Institute, a for-profit organization that promotes data warehousing, is founded.
- 1996— Ralph Kimball publishes the book The Data Warehouse Toolkit.
- 2000— Daniel Linstedt releases the Data Vault, enabling real time auditable Data Warehouses warehouse.

1.3. FUNCTIONALITY OF DATA WAREHOUSES

Data warehouses exist to facilitate complex, data-intensive and frequent adhoc queries. Data warehouses must provide far greater and more efficient query support than is demanded of transactional databases. The data warehouse access component supports enhanced spreadsheet functionality, efficient query processing, structured queries, adhoc queries, data mining and materialized views. Particularly enhanced spreadsheet functionality includes support for state-of-the art spreadsheet applications as well as for Online Analytical Processing applications programs. These provide preprogrammed functionalities such as the following:

1. *Roll-up*: Data is summarized with increasing generalization
2. *Drill-down*: Increasing levels of detail are revealed
3. *Pivot*: Cross tabulation that is, rotation, performed
4. *Slice and dice*: Performing projection operations on the dimensions
5. *Sorting*: Data is sorted by ordinal value
6. *Selection*: Data is available by value or range
7. *Derived or computer attributes*: Attributes are computed by operations on stored and derived values.

1.4. DATABASES VS. DATA WAREHOUSES

A database is a collection of related data and a database system is a database and database software together. It is also a collection of information as well as a supporting system. Databases are transactional such as relational, object-oriented, network or hierarchical.

Traditional databases support on-line transaction processing (OLTP), which includes insertions, updates, and deletions, while also supporting information query requirements. Traditional databases are optimized to process queries that may touch a small part of the database and transactions that deal with insertions or updates of a few tuples per relation to process.

Thus databases must strike a balance between efficiency in transaction processing and supporting query requirements (ad hoc user requests), That is, they can't further optimized for the applications such as OLAP, Decision Support System and data mining.

But a data warehouse is typically optimized for access from a decision maker's needs. Data warehouses are designed specifically to support efficient extraction, processing and presentation for analytic and decision-making purposes.

In contrast to databases, data warehouses generally contain very large amounts of data from multiple sources that may include databases from different data models and sometimes files acquired from independent systems and platforms.

Multidatabases provide access to disjoint and usually heterogeneous databases and are volatile. Whereas a data warehouse is frequently a store of integrated data from multiple sources, processed for storage in a multidimensional model and nonvolatile. Data warehouses also support time-series and trend analysis, both of which require more historical data.

In transactional systems, transactions are the unit and are the agent of change to the database, but data warehouse information is much more coarse-grained and is refreshed according to a careful choice of incremental refresh policy. Warehouse updates are handled by the warehouse's acquisition component that provides all required processing. As data warehouses encompass large volumes of data, they are more or less double the size of source databases.

The sheer volume of data likely to be in terabytes is an issue that has been dealt with through enterprise-wide data warehouses, virtual data warehouses and data marts. Enterprise-wide data warehouses are huge projects in need of massive investment of time and resources. Virtual data warehouses are bound to provide views of operational databases that are materialized for efficient access.

A data mart is an easy-to-access repository of a subset of highly focused data for a single function or department (*i.e.*, finance, sales, and marketing) and is considerably smaller