

FUNDAMENTALS OF  
**Business Analytics**



R N Prasad  
Seema Acharya

WILEY

FUNDAMENTALS OF  
**Business Analytics**

2nd Edition



# FUNDAMENTALS OF Business Analytics

2nd Edition

R. N. Prasad

*Founder and Chief Practice Officer of Confluence Consulting Circle  
Formerly, Senior Principal Consultant at Infosys, India*

Seema Acharya

*Lead Principal, Education, Training and Assessment Department,  
Infosys Limited*

WILEY

FUNDAMENTALS OF  
**Business Analytics**

*2nd Edition*

Copyright © 2016 by Wiley India Pvt. Ltd., 4435-36/7, Ansari Road, Daryaganj, New Delhi-110002.

Cover Image from © Sergey Nivens/Shutterstock

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, electronic, mechanical, photocopying, recording or scanning without the written permission of the publisher.

**Limits of Liability:** While the publisher and the authors have used their best efforts in preparing this book, Wiley and the authors make no representation or warranties with respect to the accuracy or completeness of the contents of this book, and specifically disclaim any implied warranties of merchantability or fitness for any particular purpose. There are no warranties which extend beyond the descriptions contained in this paragraph. No warranty may be created or extended by sales representatives or written sales materials. The accuracy and completeness of the information provided herein and the opinions stated herein are not guaranteed or warranted to produce any particular results, and the advice and strategies contained herein may not be suitable for every individual. Neither Wiley India nor the authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

**Disclaimer:** The contents of this book have been checked for accuracy. Since deviations cannot be precluded entirely, Wiley or its authors cannot guarantee full agreement. As the book is intended for educational purpose, Wiley or its authors shall not be responsible for any errors, omissions or damages arising out of the use of the information contained in the book. This publication is designed to provide accurate and authoritative information with regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services.

**Trademarks:** All brand names and product names used in this book are trademarks, registered trademarks, or trade names of their respective holders. Wiley is not associated with any product or vendor mentioned in this book.

Other Wiley Editorial Offices:

John Wiley & Sons, Inc. 111 River Street, Hoboken, NJ 07030, USA

Wiley-VCH Verlag GmbH, Pappelallee 3, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada, M9W 1L1

First Edition: 2011

Second Edition: 2016

ISBN: 978-81-265-6379-1

ISBN: 978-81-265-8248-8 (ebk)

[www.wileyindia.com](http://www.wileyindia.com)

Printed at:

Note: CD is not part of ebook



# Foreword

Business Analytics deals with the methodology of making available dependable ‘facts’ using large volumes of high-quality data from multiple sources. The challenges are in ensuring that data are of good quality, data in different formats are properly integrated, and aggregated data are stored in secure systems, and then delivered to users at the right time in the right format. Businesses have gone past the traditional approaches of ‘MIS Reporting’ and moved into the realm of statistical/quantitative analysis and predictive modeling. Organizations world-wide have evolved systems and processes to transform transactional data into business insights. There are suites of technical as well as management concepts that have come together to deliver analytics relevant to the business.

Business intelligence systems and analytical applications provide organizations with information from the enormous amount of data hidden in their various internal systems, and equip the organization with abilities to influence the business direction. They provide answers to complex business questions, such as:

- How do we get more insights about our customers, partners and employee talents?
- What will be the products and services our customers will need?
- How do we target our products and services at the most profitable geographies, customers and segments?
- How do we detect and prevent frauds?
- How do we enhance the experience of ‘Doing Business with Us’ for all stakeholders?

This book is an attempt at providing foundational knowledge associated with the domain of Business Analytics. It has basic and holistic coverage of the topics needed to handle data. The coverage include organizational perspectives, technical perspectives, as well some of the management concepts used in this field.

This book aims to provide a good coverage of all the critical concepts including:

- What problems does the technology of Data Warehouse (DW)/Business Intelligence (BI)/Advanced Analytics (AA) solve for businesses?
- When is an organization ready for DW/BI/AA projects?
- How do industries approach BI/DW/AA projects?

- What is the significance of data, meta-data and data quality challenges, and data modeling?
- What data are to be analyzed? How do we analyze data? What are the characteristics of good metrics?
- How do we frame Key Performance Indicators (KPIs)?
- What makes for a good Dashboard and Analytical Report?
- What is the impact of mobile computing, cloud computing, ERP applications, and social networks?

The book is authored by team members with twenty person-years of good business, industry and corporate - university experience. They have related their practical experiences in the field of business analytics in simple language. Hands-on experience may be gained by using the suggested open - source software tools and the step-by-step lab guide provided by the authors; the exercises and project ideas will bring enhanced exposure.

This book promises to get you started in the world of business analytics and harness the power therein for a fruitful and rewarding career.

**Prof R Natarajan**  
**Former Chairman, AICTE**  
**Former Director, IIT Madras**



# Preface

The idea of a book was conceived owing to the demand by students and instructor fraternities alike, for a book which includes a comprehensive coverage of business intelligence, data warehousing, analytics and its business applications.

## **Salient Features of the Book**

The following are few salient features of the book:

- The book promises to be a single source of introductory knowledge on business intelligence which can be taught in one semester. It will provide a good start for first time learners typically from the engineering and management discipline.
- The book covers the complete life cycle of BI/Analytics project: Covering operational/transactional data sources, data transformation, data mart/warehouse design-build, analytical reporting and dashboards.
- To ensure that concepts can be practiced for deeper understanding at low cost or no cost, the book is accompanied with step by step Hands-On manual (in CD) on:
  - Advanced MS Excel to explain the concept of analysis
  - R Programming to explain analysis and visualization
- Business Intelligence subject cannot be studied in isolation. The book provides a holistic coverage beginning with an enterprise context, developing deeper understanding through the use of tools, touching a few domains where BI is embraced and discussing the problems that BI can help solve.
- The concepts are explained with the help of illustrations and real life industrial strength application examples.
- The pre-requisites for each chapter and references to books currently available are stated.
- In addition the book has the following pedagogical features:
  - Industrial application case studies.
  - Cross word puzzles/do it yourself/exercises to help with self-assessment. The solutions to these have also been provided.

- Glossary of terms.
- Summary at the end of every chapter.
- References/web links/bibliography – generally at the end of every concept.

## Target Readers

While designing this book, the authors have kept in mind the typical requirements of the undergraduate engineering/MCA/MBA students who are still part of the academic program in Universities. The attempt is to help the student fraternity gain sound foundational knowledge of “Analytics” irrespective of the engineering discipline or special electives they may have chosen. The best way to benefit from this book will be to learn by completing all the hands-on lab assignments, exploring the sources of knowledge compiled in this book as well as studying one or two organizations in detail that are leveraging analytics for their business benefits.

## Structure of the Book

The book has 13 chapters.

- Chapter 1 introduces the readers to the enterprise view of IT applications and discusses some of the salient requirements of information users.
- Chapter 2 explains the types of digital data: structured, semi-structured and unstructured. The reader will learn the nuances of storing, retrieving and extracting information from the various types of digital data.
- Chapter 3 aims to differentiate between OLTP (On-line Transaction Processing) and OLAP (On Line Analytical Processing) systems.
- Chapters 4 and 5 set the stage for a novice BI learner by enunciating all the terms, concepts and technology employed in the BI space.
- Chapters 6, 7, 8 and 9 are core chapters that discuss the building of a data warehouse (Extraction Transformation Loading process), analyzing the data from multiple perspectives, comprehending the significance of data quality and metrics, presenting the data in the apt format to the right information users.
- Chapters 10 and 11 focus on the significance of statistics in analytics. They seek to explain analytics through discussing applications.
- Chapter 12 introduces data mining algorithms such as Market Basket Analysis, k-means clustering and decisions trees with the subsequent implementation of the stated algorithms in R statistical programming tool.
- The book finally concludes with Chapter 13 on the confluence of BI with mobile technology; cloud computing, social CRM, etc.

Glossary of terms is also provided at the end of the book.

## CD Companion

The companion CD consists of a collection of two step-by-step lab guides:

- Analysis using “Advanced Microsoft Excel”
- Introduction to R programming

## How to Get the Maximum Out of the Book

The authors recommend that the chapters of the book be read in sequence from Chapter 1 to Chapter 13. The authors recommend that readers complete the pre-requisites stated for the chapter before plunging into the chapter. The authors have provided useful references at the end of each chapter that should be explored for deeper gains. After completing the chapter, the readers should solve the “Test Exercises”.

We welcome your suggestions/feedback to assist us in our endeavors to help the learners gain a good foundation in this subject that is critical to the industry.

Our heartfelt gratitude to the teacher's community who have adopted our book while teaching and evangelizing Business Intelligence subject in their respective colleges. We heard your requirements to include new topics related to Big Data, Statistics, Lab with R, Lab with Advanced Excel and Industry examples of applications of Analytics. We have incorporated these topics while retaining the original content to ensure utmost alignment with your college curriculum. Our earnest hope that the Second Edition of the book will immensely benefit the student's community.

**RN Prasad**

**Seema Acharya**

**July 2016**





# Acknowledgments

We don't remember having read the acknowledgement page in any book that we have read so far. And we have read quite a few! Now as we pen down the acknowledgement for our book, we realize that this is one of the most important pages. As is the case with any book, the creation of this one was an extended collaborative effort. It's a collection of ideas, case briefs, definitions, examples, and good practices from various points in the technology field. We have an army of people who have been with us on this journey and we wish to express our sincere gratitude to each one of them.

We owe it to the student and the teacher community whose inquisitiveness and incessant queries on business intelligence and analytics led us to model this book for them.

We would like to thank our friends – the practitioners from the field – for filling us in on the latest in the business intelligence field and sharing with us valuable insights on the best practices and methodology followed during the life cycle of a BI project.

A special thanks to Ramakrishnan for his unrelenting support and vigilant review.

We have been fortunate in gaining the good counsel of several of our learned colleagues and advisors who have read drafts of the book at various stages and have contributed to its contents.

We would like to thank Sanjay B. for his kind assistance. We have been extraordinarily fortunate in the editorial assistance we received in the later stages of writing and final preparation of the book. Meenakshi Sehrawat and her team of associates skillfully guided us through the entire process of preparation and publication.

And finally we can never sufficiently thank our families and friends who have encouraged us throughout the process, offering inspiration, guidance and support and enduring our crazy schedules patiently as we assembled the book.





# About the Authors

## RN Prasad

RN Prasad is the founder and Chief Practice Officer of *Confluence Consulting Circle* with over 37 years of IT industry experience. Prior to this, he was a Senior Principal Consultant at Infosys, India. He is a trusted advisor in BI & Analytics technology adoption and Product Management practices. He works with companies to develop data platforms, patent innovative ideas, and design smart-decision systems leveraging SMAC.

His interest areas include Big Data applications, Business Analytics solutions, BI self-service portals, Data Virtualization, Digital Transformation, Analytics & Big Data related executive education, Analytics certification assessments design, University curriculum innovation consulting and IGIP (International Society for Engineering Pedagogy)-based faculty certification.

## Seema Acharya

Seema Acharya is Lead Principal with the Education, Training and Assessment Department of Infosys Limited. She is an educator by choice and vocation, and has rich experience in both academia and the software industry. She is also the author of the book, “Big Data and Analytics”, ISBN: 9788126554782, publisher – Wiley India.

She has co-authored a paper on “Collaborative Engineering Competency Development” for ASEE (American Society for Engineering Education). She also holds the patent on “Method and system for automatically generating questions for a programming language”.

Her areas of interest and expertise are centered on Business Intelligence, Big Data and Analytics, technologies such as Data Warehousing, Data Mining, Data Analytics, Text Mining and Data Visualization.





# Contents

|  |             |
|--|-------------|
| <b>Foreword</b>  | <b>v</b>    |
| <b>Preface</b>   | <b>vii</b>  |
| <b>Acknowledgments</b>   | <b>xi</b>   |
| <b>About the Authors</b>   | <b>xiii</b> |
| <hr/> <b>1 Business View of Information Technology Applications</b>              | <b>1</b>    |
| Brief Contents   | 1           |
| What's in Store  | 1           |
| 1.1 Business Enterprise Organization, Its Functions, and Core Business Processes | 2           |
| <i>1.1.1 Core Business Processes</i>   | 3           |
| 1.2 Baldrige Business Excellence Framework (Optional Reading)                    | 6           |
| <i>1.2.1 Leadership</i>  | 7           |
| <i>1.2.2 Strategic Planning</i>  | 7           |
| <i>1.2.3 Customer Focus</i>  | 7           |
| <i>1.2.4 Measurement, Analysis, and Knowledge Management</i>                     | 7           |
| <i>1.2.5 Workforce Focus</i>   | 8           |
| <i>1.2.6 Process Management</i>  | 9           |
| <i>1.2.7 Results</i>   | 9           |
| 1.3 Key Purpose of using IT in Business  | 11          |
| 1.4 The Connected World: Characteristics of Internet-Ready IT Applications       | 12          |
| 1.5 Enterprise Applications (ERP/CRM, etc.) and Bespoke IT Applications          | 14          |
| 1.6 Information Users and Their Requirements                                     | 17          |
| Unsolved Exercises   | 18          |
| <hr/> <b>Case Study Briefs</b>   | <b>19</b>   |
| GoodLife HealthCare Group  | 19          |
| Introduction   | 19          |

|                           |  |           |
|---------------------------|--|-----------|
|                           | Business Segments                                    | 20        |
|                           | Organizational Structure                             | 20        |
|                           | Quality Management                                   | 20        |
|                           | Marketing  | 20        |
|                           | Alliance Management                                  | 21        |
|                           | Future Outlook                                       | 21        |
|                           | Information Technology at GoodLife Group             | 21        |
|                           | Human Capital Management & Training Management       | 21        |
| GoodFood Restaurants Inc. |  | 23        |
|                           | Introduction   | 23        |
|                           | Business Segments                                    | 23        |
|                           | Impeccable Processes and Standard Cuisine            | 23        |
|                           | Marketing  | 24        |
|                           | Supplier Management                                  | 24        |
|                           | Quality Management                                   | 24        |
|                           | Organization Structure                               | 24        |
|                           | Future Outlook                                       | 25        |
|                           | Information Technology at GoodFood                   | 25        |
| TenToTen Retail Stores    |  | 27        |
|                           | Introduction   | 27        |
|                           | Business Segments                                    | 27        |
|                           | Organizational Structure                             | 27        |
|                           | Marketing  | 28        |
|                           | Supplier Management                                  | 28        |
|                           | Quality Management                                   | 28        |
|                           | Future Outlook                                       | 28        |
|                           | Information Technology at TenToTen Stores            | 29        |
| <b>2</b>                  | <b>Types of Digital Data</b>                         | <b>31</b> |
|                           | Brief Contents                                       | 31        |
|                           | What's in Store                                      | 31        |
| 2.1                       | Introduction   | 31        |
| 2.2                       | Getting into "GoodLife" Database                     | 32        |
| 2.3                       | Getting to Know Structured Data                      | 33        |
|                           | <i>2.3.1 Characteristics of Structured Data</i>      | 33        |
|                           | <i>2.3.2 Where Does Structured Data Come From?</i>   | 34        |
|                           | <i>2.3.3 It's So Easy With Structured Data</i>       | 34        |
|                           | <i>2.3.4 Hassle-Free Retrieval</i>                   | 35        |
| 2.4                       | Getting to Know Unstructured Data                    | 36        |
|                           | <i>2.4.1 Where Does Unstructured Data Come From?</i> | 37        |
|                           | <i>2.4.2 A Myth Demystified</i>                      | 37        |
|                           | <i>2.4.3 How to Manage Unstructured Data?</i>        | 38        |
|                           | <i>2.4.4 How to Store Unstructured Data?</i>         | 39        |

|          |  |           |
|----------|--|-----------|
| 2.4.5    | <i>Solutions to Storage Challenges of Unstructured Data</i>      | 40        |
| 2.4.6    | <i>How to Extract Information from Stored Unstructured Data?</i> | 41        |
| 2.4.7    | <i>UIMA: A Possible Solution for Unstructured Data</i>           | 42        |
| 2.5      | Getting to Know Semi-Structured Data                             | 43        |
| 2.5.1    | <i>Where Does Semi-Structured Data Come From?</i>                | 45        |
| 2.5.2    | <i>How to Manage Semi-Structured Data?</i>                       | 47        |
| 2.5.3    | <i>How to Store Semi-Structured Data?</i>                        | 47        |
| 2.5.4    | <i>Modeling Semi-Structured Data (The OEM Way)</i>               | 48        |
| 2.5.5    | <i>How to Extract Information from Semi-Structured Data?</i>     | 49        |
| 2.5.6    | <i>XML: A Solution for Semi-Structured Data Management</i>       | 50        |
| 2.6      | Difference Between Semi-Structured and Structured Data           | 51        |
|          | Unsolved Exercises   | 56        |
| <b>3</b> | <b>Introduction to OLTP and OLAP</b>                             | <b>59</b> |
|          | Brief Contents   | 59        |
|          | What's in Store  | 59        |
| 3.1      | OLTP (On-Line Transaction Processing)                            | 59        |
| 3.1.1    | <i>Queries that an OLTP System can Process</i>                   | 60        |
| 3.1.2    | <i>Advantages of an OLTP System</i>                              | 61        |
| 3.1.3    | <i>Challenges of an OLTP System</i>                              | 61        |
| 3.1.4    | <i>The Queries that OLTP cannot Answer</i>                       | 61        |
| 3.2      | OLAP (On-Line Analytical Processing)                             | 62        |
| 3.2.1    | <i>One-Dimensional Data</i>                                      | 63        |
| 3.2.2    | <i>Two-Dimensional Data</i>                                      | 64        |
| 3.2.3    | <i>Three-Dimensional Data</i>                                    | 65        |
| 3.2.4    | <i>Should We Go Beyond the Third Dimension?</i>                  | 65        |
| 3.2.5    | <i>Queries that an OLAP System can Process</i>                   | 66        |
| 3.2.6    | <i>Advantages of an OLAP System</i>                              | 66        |
| 3.3      | Different OLAP Architectures                                     | 66        |
| 3.3.1    | <i>MOLAP (Multidimensional On-Line Analytical Processing)</i>    | 66        |
| 3.3.2    | <i>ROLAP (Relational On-Line Analytical Processing)</i>          | 67        |
| 3.3.3    | <i>HOLAP (Hybrid On-Line Analytical Processing)</i>              | 68        |
| 3.4      | OLTP and OLAP  | 68        |
| 3.5      | Data Models for OLTP and OLAP                                    | 70        |
| 3.5.1    | <i>Data Model for OLTP</i>                                       | 70        |
| 3.5.2    | <i>Data Model for OLAP</i>                                       | 70        |
| 3.6      | Role of OLAP Tools in the BI Architecture                        | 73        |
| 3.7      | Should OLAP be Performed Directly on Operational Databases?      | 74        |
| 3.8      | A Peek into the OLAP Operations on Multidimensional Data         | 74        |
| 3.8.1    | <i>Slice</i>   | 75        |
| 3.8.2    | <i>Dice</i>  | 75        |
| 3.8.3    | <i>Roll-Up</i>   | 75        |
| 3.8.4    | <i>Drill-Down</i>  | 76        |
| 3.8.5    | <i>Pivot</i>   | 76        |

|          |   |            |
|----------|---|------------|
|          | <i>3.8.6 Drill-Across</i>   | 77         |
|          | <i>3.8.7 Drill-Through</i>  | 77         |
| 3.9      | Leveraging ERP Data Using Analytics   | 77         |
|          | Solved Exercises  | 83         |
|          | Unsolved Exercises  | 86         |
| <b>4</b> | <b>Getting Started with Business Intelligence</b>   | <b>87</b>  |
|          | Brief Contents  | 87         |
|          | What's in Store   | 87         |
| 4.1      | Using Analytical Information for Decision Support   | 88         |
| 4.2      | Information Sources Before Dawn of BI?  | 88         |
| 4.3      | Definitions and Examples in Business Intelligence, Data Mining, Analytics, Machine Learning, Data Science | 89         |
| 4.4      | Looking at "Data" from Many Perspectives  | 95         |
|          | <i>4.4.1 Data Lifecycle Perspective</i>   | 96         |
|          | <i>4.4.2 Data Storage (Raw) for Processing</i>  | 97         |
|          | <i>4.4.3 Data Processing and Analysis Perspective</i>   | 97         |
|          | <i>4.4.4 Data from Business Decision Support Perspective</i>  | 100        |
|          | <i>4.4.5 Data Quality Management Aspects</i>  | 101        |
|          | <i>4.4.6 Related Technology Influences of Data</i>  | 103        |
| 4.5      | Business Intelligence (BI) Defined  | 104        |
|          | <i>4.5.1 Visibility into Enterprise Performance</i>   | 105        |
| 4.6      | Why BI? How Can You Achieve Your Stated Objectives?   | 106        |
| 4.7      | Some Important Questions About BI - Where, When and What  | 107        |
|          | <i>4.7.1 Where is BI being used?</i>  | 107        |
|          | <i>4.7.2 When should you use BI?</i>  | 107        |
|          | <i>4.7.3 What can BI deliver?</i>   | 107        |
| 4.8      | Evolution of BI and Role of DSS, EIS, MIS, and Digital Dashboards   | 107        |
|          | <i>4.8.1 Difference Between ERP (Enterprise Resource Planning) and BI</i>                                 | 109        |
|          | <i>4.8.2 Is Data Warehouse Synonymous with BI?</i>  | 109        |
| 4.9      | Need for BI at Virtually all Levels   | 110        |
| 4.10     | BI for Past, Present, and Future  | 110        |
| 4.11     | The BI Value Chain  | 111        |
| 4.12     | Introduction to Business Analytics  | 112        |
|          | Unsolved Exercises  | 116        |
| <b>5</b> | <b>BI Definitions and Concepts</b>  | <b>117</b> |
|          | Brief Contents  | 117        |
|          | What's in Store   | 117        |
| 5.1      | BI Component Framework  | 117        |
|          | <i>5.1.1 Business Layer</i>   | 118        |
|          | <i>5.1.2 Administration and Operation Layer</i>   | 120        |
|          | <i>5.1.3 Implementation Layer</i>   | 123        |

---

|     |  |     |
|-----|--|-----|
| 5.2 | Who is BI for?                                 | 126 |
|     | <i>5.2.1 BI for Management</i>                 | 127 |
|     | <i>5.2.2 Operational BI</i>                    | 127 |
|     | <i>5.2.3 BI for Process Improvement</i>        | 128 |
|     | <i>5.2.4 BI for Performance Improvement</i>    | 128 |
|     | <i>5.2.5 BI to Improve Customer Experience</i> | 128 |
| 5.3 | BI Users                                       | 129 |
|     | <i>5.3.1 Casual Users</i>                      | 130 |
|     | <i>5.3.2 Power Users</i>                       | 130 |
| 5.4 | Business Intelligence Applications             | 131 |
|     | <i>5.4.1 Technology Solutions</i>              | 131 |
|     | <i>5.4.2 Business Solutions</i>                | 133 |
| 5.5 | BI Roles and Responsibilities                  | 135 |
|     | <i>5.5.1 BI Program Team Roles</i>             | 135 |
|     | <i>5.5.2 BI Project Team Roles</i>             | 137 |
| 5.6 | Best Practices in BI/DW                        | 141 |
| 5.7 | The Complete BI Professional                   | 144 |
| 5.8 | Popular BI Tools                               | 146 |
|     | Unsolved Exercises                             | 148 |

---

|          |   |            |
|----------|---|------------|
| <b>6</b> | <b>Basics of Data Integration</b>                     | <b>151</b> |
|          | Brief Contents  | 151        |
|          | What's in Store                                       | 151        |
| 6.1      | Need for Data Warehouse                               | 152        |
| 6.2      | Definition of Data Warehouse                          | 154        |
| 6.3      | What is a Data Mart?                                  | 155        |
| 6.4      | What is then an ODS?                                  | 155        |
| 6.5      | Ralph Kimball's Approach vs. W.H. Inmon's Approach    | 155        |
| 6.6      | Goals of a Data Warehouse                             | 157        |
| 6.7      | What Constitutes a Data Warehouse?                    | 157        |
|          | <i>6.7.1 Data Sources</i>                             | 159        |
| 6.8      | Extract, Transform, Load                              | 159        |
|          | <i>6.8.1 Data Mapping</i>                             | 159        |
|          | <i>6.8.2 Data Staging</i>                             | 159        |
| 6.9      | What is Data Integration?                             | 164        |
|          | <i>6.9.1 Two Main Approaches to Data Integration</i>  | 164        |
|          | <i>6.9.2 Need and Advantages for Data Integration</i> | 167        |
|          | <i>6.9.3 Common Approaches of Data Integration</i>    | 167        |
| 6.10     | Data Integration Technologies                         | 170        |
| 6.11     | Data Quality  | 175        |
|          | <i>6.11.1 Why Data Quality Matters?</i>               | 175        |
|          | <i>6.11.2 What is Data Quality?</i>                   | 176        |
|          | <i>6.11.3 How Do We Maintain Data Quality?</i>        | 180        |
|          | <i>6.11.4 Key Areas of Study</i>                      | 181        |
|          | Unsolved Exercises                                    | 182        |

|                |   |            |
|----------------|---|------------|
| 6.12           | Data Profiling  | 182        |
|                | <i>6.12.1 The Context</i>                                 | 182        |
|                | <i>6.12.2 The Problem</i>                                 | 183        |
|                | <i>6.12.3 The Solution</i>                                | 183        |
|                | <i>6.12.4 What is Data Profiling?</i>                     | 184        |
|                | <i>6.12.5 When and How to Conduct Data Profiling?</i>     | 184        |
|                | Summary   | 187        |
|                | A Case Study from the Healthcare Domain                   | 187        |
|                | Solved Exercises  | 201        |
|                | Unsolved Exercises  | 203        |
| <hr/> <b>7</b> | <b>Multidimensional Data Modeling</b>                     | <b>205</b> |
|                | Brief Contents  | 205        |
|                | What's in Store   | 205        |
| 7.1            | Introduction  | 206        |
| 7.2            | Data Modeling Basics                                      | 207        |
|                | <i>7.2.1 Entity</i>                                       | 207        |
|                | <i>7.2.2 Attribute</i>                                    | 207        |
|                | <i>7.2.3 Cardinality of Relationship</i>                  | 207        |
| 7.3            | Types of Data Model                                       | 207        |
|                | <i>7.3.1 Conceptual Data Model</i>                        | 208        |
|                | <i>7.3.2 Logical Data Model</i>                           | 208        |
|                | <i>7.3.3 Physical Model</i>                               | 215        |
| 7.4            | Data Modeling Techniques                                  | 219        |
|                | <i>7.4.1 Normalization (Entity Relationship) Modeling</i> | 219        |
|                | <i>7.4.2 Dimensional Modeling</i>                         | 222        |
| 7.5            | Fact Table  | 225        |
|                | <i>7.5.1 Types of Fact</i>                                | 225        |
| 7.6            | Dimension Table   | 229        |
|                | <i>7.6.1 Dimension Hierarchies</i>                        | 229        |
|                | <i>7.6.2 Types of Dimension Tables</i>                    | 230        |
| 7.7            | Typical Dimensional Models                                | 237        |
|                | <i>7.7.1 Star Schema</i>                                  | 237        |
|                | <i>7.7.2 Snowflake Schema</i>                             | 239        |
|                | <i>7.7.3 Fact Constellation Schema</i>                    | 243        |
| 7.8            | Dimensional Modeling Life Cycle                           | 246        |
|                | <i>7.8.1 Requirements Gathering</i>                       | 247        |
|                | <i>7.8.2 Identify the Grain</i>                           | 249        |
|                | <i>7.8.3 Identify the Dimensions</i>                      | 250        |
|                | <i>7.8.4 Identify the Facts</i>                           | 250        |
|                | Designing the Dimensional Model                           | 251        |
|                | Solved Exercises  | 253        |
|                | Unsolved Exercises  | 255        |

---

|          |   |            |
|----------|---|------------|
| <b>8</b> | <b>Measures, Metrics, KPIs, and Performance Management</b>                  | <b>257</b> |
|          | Brief Contents  | 257        |
|          | What's in Store   | 257        |
| 8.1      | Understanding Measures and Performance                                      | 258        |
| 8.2      | Measurement System Terminology  | 258        |
| 8.3      | Navigating a Business Enterprise, Role of Metrics, and Metrics Supply Chain | 259        |
| 8.4      | "Fact-Based Decision Making" and KPIs                                       | 264        |
| 8.5      | KPI Usage in Companies  | 266        |
| 8.6      | Where do Business Metrics and KPIs Come From?                               | 267        |
| 8.7      | Connecting the Dots: Measures to Business Decisions and Beyond              | 268        |
|          | Summary   | 269        |
|          | Unsolved Exercises  | 270        |
| <b>9</b> | <b>Basics of Enterprise Reporting</b>                                       | <b>273</b> |
|          | Brief Contents  | 273        |
|          | What's in Store   | 273        |
| 9.1      | Reporting Perspectives Common to All Levels of Enterprise                   | 274        |
| 9.2      | Report Standardization and Presentation Practices                           | 275        |
|          | 9.2.1 <i>Common Report Layout Types</i>                                     | 277        |
|          | 9.2.2 <i>Report Delivery Formats</i>  | 280        |
| 9.3      | Enterprise Reporting Characteristics in OLAP World                          | 281        |
| 9.4      | Balanced Scorecard  | 282        |
|          | 9.4.1 <i>Four Perspectives of Balanced Scorecard</i>                        | 283        |
|          | 9.4.2 <i>Balanced Scorecard as Strategy Map</i>                             | 284        |
|          | 9.4.3 <i>Measurement System</i>   | 284        |
|          | 9.4.4 <i>Balanced Scorecard as a Management System</i>                      | 285        |
| 9.5      | Dashboards  | 290        |
|          | 9.5.1 <i>What are Dashboards?</i>   | 290        |
|          | 9.5.2 <i>Why Enterprises Need Dashboards?</i>                               | 291        |
|          | 9.5.3 <i>Types of Dashboard</i>   | 291        |
| 9.6      | How Do You Create Dashboards?   | 293        |
|          | 9.6.1 <i>Steps for Creating Dashboards</i>                                  | 293        |
|          | 9.6.2 <i>Tips For Creating Dashboard</i>                                    | 294        |
| 9.7      | Scorecards vs. Dashboards   | 296        |
|          | 9.7.1 <i>KPIs: On Dashboards as well as on Scorecards</i>                   | 297        |
|          | 9.7.2 <i>Indicators: On Dashboards as well as on Scorecards</i>             | 297        |
| 9.8      | The Buzz Behind Analysis...   | 298        |
|          | 9.8.1 <i>Funnel Analysis</i>  | 298        |
|          | 9.8.2 <i>Distribution Channel Analysis</i>                                  | 300        |
|          | 9.8.3 <i>Performance Analysis</i>   | 301        |
|          | Unsolved Exercises  | 307        |

|           |   |            |
|-----------|---|------------|
| <b>10</b> | <b>Understanding Statistics</b>   | <b>309</b> |
|           | Brief Contents  | 309        |
|           | What's in Store?  | 309        |
| 10.1      | Role of Statistics in Analytics   | 309        |
| 10.2      | Data, Data Description and Summarization                                  | 311        |
|           | <i>10.2.1 Getting to Describe "Categorical Data"</i>                      | 311        |
|           | <i>10.2.2 Getting to Describe "Numerical Data"</i>                        | 314        |
|           | <i>10.2.3 Association between Categorical Variables</i>                   | 316        |
|           | <i>10.2.4 Association between Quantitative Variables</i>                  | 317        |
| 10.3      | Statistical Tests   | 319        |
|           | <i>10.3.1 Paired and Unpaired Data Sets</i>                               | 319        |
|           | <i>10.3.2 Matched Pair Groups in Data Sets</i>                            | 319        |
|           | <i>10.3.3 Common Statistical Testing Scenarios</i>                        | 320        |
| 10.4      | Understanding Hypothesis and t-Test                                       | 321        |
|           | <i>10.4.1 The t-Test</i>  | 322        |
|           | <i>10.4.2 The p-Value</i>   | 322        |
|           | <i>10.4.3 Z-Test</i>  | 322        |
| 10.5      | Correlation Analysis  | 323        |
| 10.6      | Regression  | 324        |
| 10.7      | ANOVA   | 325        |
| 10.8      | The F-Test  | 325        |
| 10.9      | Time Series Analysis  | 327        |
| <b>11</b> | <b>Application of Analytics</b>   | <b>331</b> |
|           | Brief Contents  | 331        |
|           | What's in Store?  | 331        |
| 11.1      | Application of Analytics  | 331        |
|           | <i>11.1.1 Analytics in Business Support Functions</i>                     | 332        |
| 11.2      | Analytics in Industries   | 334        |
|           | <i>11.2.1 Analytics in Telecom</i>  | 334        |
|           | <i>11.2.2 Analytics in Retail</i>   | 335        |
|           | <i>11.2.3 Analytics in Healthcare (Hospitals or Healthcare Providers)</i> | 337        |
|           | <i>11.2.4 Analytical Application Development</i>                          | 338        |
| 11.3      | Widely Used Application of Analytics                                      | 339        |
|           | <i>11.3.1 Anatomy of Social Media Analytics</i>                           | 339        |
|           | <i>11.3.2 Anatomy of Recommendation Systems</i>                           | 342        |
|           | <i>11.3.3 Components of Recommendation Systems</i>                        | 343        |
| <b>12</b> | <b>Data Mining Algorithms</b>   | <b>347</b> |
|           | Brief Contents  | 347        |
|           | What's in Store?  | 347        |
| 12.1      | Association Rule Mining   | 347        |
|           | <i>12.1.1 Binary Representation</i>                                       | 349        |

|           |   |            |
|-----------|---|------------|
|           | <i>12.1.2 Itemset and Support Count</i>                         | 350        |
|           | <i>12.1.3 Implementation in R</i>                               | 351        |
| 12.2      | <i>k</i> -Means Clustering                                      | 355        |
|           | <i>12.2.1 Implementation in R</i>                               | 356        |
| 12.3      | Decision Tree   | 357        |
|           | <i>12.3.1 What is a Decision Tree?</i>                          | 360        |
|           | <i>12.3.2 Where is it Used?</i>                                 | 361        |
|           | <i>12.3.3 Advantages from Using a Decision Tree</i>             | 361        |
|           | <i>12.3.4 Disadvantages of Decision Trees</i>                   | 361        |
|           | <i>12.3.5 Decision Tree in R</i>                                | 361        |
|           | Unsolved Exercises  | 367        |
| <b>13</b> | <b>BI Road Ahead</b>  | <b>369</b> |
|           | Brief Contents  | 369        |
|           | What's in Store   | 369        |
| 13.1      | Understanding BI and Mobility                                   | 369        |
|           | <i>13.1.1 The Need for Business Intelligence on the Move</i>    | 370        |
|           | <i>13.1.2 BI Mobility Timeline</i>                              | 370        |
|           | <i>13.1.3 Data Security Concerns for Mobile BI</i>              | 373        |
| 13.2      | BI and Cloud Computing  | 373        |
|           | <i>13.2.1 What is Cloud Computing?</i>                          | 373        |
|           | <i>13.2.2 Why Cloud Computing?</i>                              | 375        |
|           | <i>13.2.3 Why Business Intelligence should be on the Cloud?</i> | 375        |
| 13.3      | Business Intelligence for ERP Systems                           | 377        |
|           | <i>13.3.1 Why BI in ERP?</i>                                    | 378        |
|           | <i>13.3.2 Benefits of BI in ERP</i>                             | 379        |
|           | <i>13.3.3 ERP Plus BI Equals More Value</i>                     | 379        |
| 13.4      | Social CRM and BI   | 380        |
|           | Unsolved Exercises  | 384        |
|           | <b>Glossary</b>   | <b>385</b> |
|           | <b>Index</b>  | <b>397</b> |



# 1



## Business View of Information Technology Applications

### BRIEF CONTENTS

|  |  |
|--|--|
| What's in Store  | The Connected World: Characteristics of Internet-ready IT Applications |
| Business Enterprise Organization, Its Functions, and Core Business Processes | Enterprise Applications (ERP/CRM, etc.) and Bespoke IT Applications    |
| Baldridge Business Excellence Framework<br>(Optional Reading)                | Information Users and Their Requirements                               |
| Key Purpose of Using IT in Business  | Unsolved Exercises   |

### WHAT'S IN STORE

Welcome to the “world of data”. Whatever profession you may choose for your career, there will be an element of Information Technology (IT) in your job. The competence to harness the power of data will be among the critical competencies that you will be required to develop in your career. This competency could potentially be one of the differentiators in your career. As already stated in the preface, the content of this book aims to provide basic but holistic knowledge needed to leverage the power of data using technologies, processes, and tools.

While designing this book, we have kept in mind the typical requirements of the undergraduate engineering/MCA/MBA students who are still part of the academic program in universities. Our attempt is to help you gain sound foundational knowledge of “Analytics” irrespective of the engineering discipline or special electives you have chosen. The best way to benefit from this book is to learn hands-on by completing all lab assignments, explore the sources of knowledge included in this book and study in detail one or two organizations which are leveraging analytics for their business benefits.

In this chapter, we intend to familiarize you with the complete context or the environment of a typical business enterprise which leverages IT for managing its business.

The topics included are:

- Business enterprise organization, its functions, and core business processes.
- Key purpose of using IT in business.
- The connected world: Characteristics of Internet-ready IT applications.
- Enterprise applications (ERP/CRM and Bespoke IT applications, systems, software, and network heterogeneity).
- Information users and their requirements.

No prior IT knowledge is assumed. This chapter is suggested as a “Must Read” for all readers as we intend to create the “big picture” where “Business Analytics” will be deployed and used. It will be a good idea to explore, collect, and delve into case studies of applications of Business Intelligence/Analytics/ Data Warehousing in the domain/discipline of your interest to develop your own mental model of the enterprise. Learning in teams by discussing the above topics across different disciplines will expand your interest and throw light on the ever increasing possibilities of leveraging data. This is the reason for referring to data as “corporate asset”.

One last noteworthy fact – As enterprises grow and expand their business, they acquire larger computers, change operating systems, database management systems, connect systems by digital network, and migrate to powerful database management systems. They never discard any data! Historical data is like the “finger print” or DNA of an enterprise and can potentially be a smart guidance system.

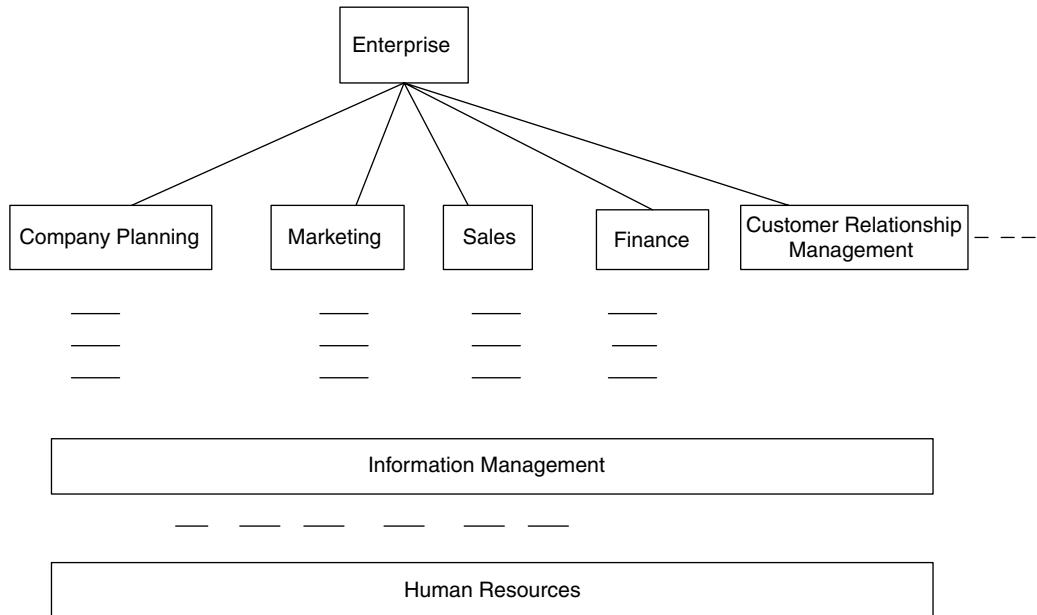
So, let's begin our exploration of the modern enterprise.

## 1.1 BUSINESS ENTERPRISE ORGANIZATION, ITS FUNCTIONS, AND CORE BUSINESS PROCESSES

---

No matter which industry domain you choose – retail, banking, manufacturing, transport, energy, etc. – you will find similarities in their organizational structure. In general, businesses serve their customers either with products or services or both. This immediately creates the need for functions such as product/service research, product manufacture or service delivery, quality assurance, purchase, distribution, sales, marketing, finance, and so on. As the product or service consumption increases, we will find functions like support, planning, training, etc. being added. The business will depend on its workforce to carry out all the functions identified so far, creating the need for human resources management, IT management, customer relationship management, partner management, and so on. As the business expands into multiple product/service lines and starts global operations, new requirements emerge, and each function realigns its own “internal services” to meet the new demands. You will also notice that some functions like HR, Training, Finance, Marketing, R&D, and IT management cut across various product/service lines and could be termed as “Support Units”. However, such product or service line functions that generate revenue could be termed as “Business Units”. Today, businesses, irrespective of their size, leverage IT as enabler to achieve market leadership or unique position. The organization of a typical business could be represented as shown in Figure 1.1.

Most businesses design standard enterprise-level processes for their core functions. The basic idea behind creating standard processes is to ensure repeatable, scalable, and predictable customer experience. Well-designed processes make customer experience independent of people and geography/location. Here a good example is the common dining experience in any McDonald's outlet anywhere in the world! You will come across processes that are contained in one department or involve multiple



**Figure 1.1** The organization of a typical enterprise.

departments. For example, a process like “Hire graduates from campuses” will be managed only by HR. But a process like “Order to Delivery” in a computer manufacturing company will involve functions of sales, order fulfillment, shipment, and finance. It is these processes that will be automated by the use of IT. We will focus on the automation of business processes using IT in subsequent sections.

### 1.1.1 Core Business Processes

Why do we need core business functions? The answer is for a smooth day-to-day functioning of the organization and for performing tasks efficiently and on time. Given below is the list of a few common core business functions.

- Sales and Marketing
- Product/Service Delivery
- Quality
- Product Development
- Accounting
- Technology
- Human Resource Management
- Supplier Management
- Legal and Compliance
- Corporate Planning
- Procurement (Purchases)

All the core business functions listed above might not be used in all organizations. There could be several factors for not using a core business function; for example, shortage of resources or in a small organization, a particular process can be easily managed manually, etc.

## Picture this...

An assumed organization “Imagination” is an IT products and services company. The number of employees working for it is roughly 550. Within a short time of its inception, the company has done quite well for itself. It has 32 clients spread across the globe. To manage its customers, “Imagination” has a **Sales and Marketing** function. As you will agree, in any business it is just not enough to have customers; you should be able to retain your customers and grow your customer base. “Imagination” has realized this and believes that the relationship with its customers’ needs to be managed. The aim is to have loyal customers coming back to the company with repeat business. Besides, they should also advocate about the company’s product and services on whichever forum they get invited to. Next in the line of the company’s core business processes is **Product/Service Delivery**. The company makes its revenue either by rendering service or by selling their products or both. And while they do the business, they want to be ahead of their competitors. Very likely, **Quality** will come into play and they would like it to be their key differentiator in the days to come.

“Imagination”, the company in our example, is into **Product Development** and is looking for innovative ways to develop innovative products. It could be the same proprietary product but developed using innovative ways which could be use of automation, reduced cost, better processes. The company is into business and they have customers, so it is not unusual for cash to come in for the products sold or the services rendered. The company needs to account for the cash flow and also file the tax returns, which brings us to another of the company’s core business function, **Accounting**. The Accounting function relies heavily on technology to be able to manage the transactions accurately and timely. So, **Technology** becomes another of its core business functions.

To manage its present employee base of 550 (the number is likely to double next year), the company has a core support function in **Human Resource Management (HR)**. The HR is responsible for hiring, training, and managing the employees.

“Imagination” has a few vendor partners, to manage relationship with them the **Supplier Management** process is in place. In the years to come, if the company grows to a big size, it might be worthwhile to consider **“Corporate Planning”** that looks at the allocation of resources (both tangible and non-tangible) and manpower to the various functions/units. The last business process but not the least is the **Procurement** process, very much relevant for an IT company such as “Imagination”. This process deals with the procurement of hardware/software licenses and other technical assets.

Let’s move to the next step and start thinking of generalizing or abstracting the commonalities between different businesses in different industries. What is common in the HR function, for example, relating to a bank, an electronics goods manufacturing plant, and a large retail chain like McDonald’s? They all hire new employees, train the employees in their functions, pay for the employees, evaluate employees’ on-the-job performance, maintain relationship with the suppliers and employees by way of programs such as “employee competency development”, “employee loyalty program”, etc., communicate strategic messages, reward employees, and so on. They also measure their own contribution to business. Think on similar lines about the other functions. For example, a bank may view home loan as a product, credit cards as product, etc. A large CT scanner manufacturer may view installation of equipment as a service associated with their product that generates additional revenue.

Business functions design core business processes that are practiced within the function or cut across several business and support functions. Refer to Table 1.1 for few of the common core business processes.

Can we represent all businesses in some common model? The answer is YES!

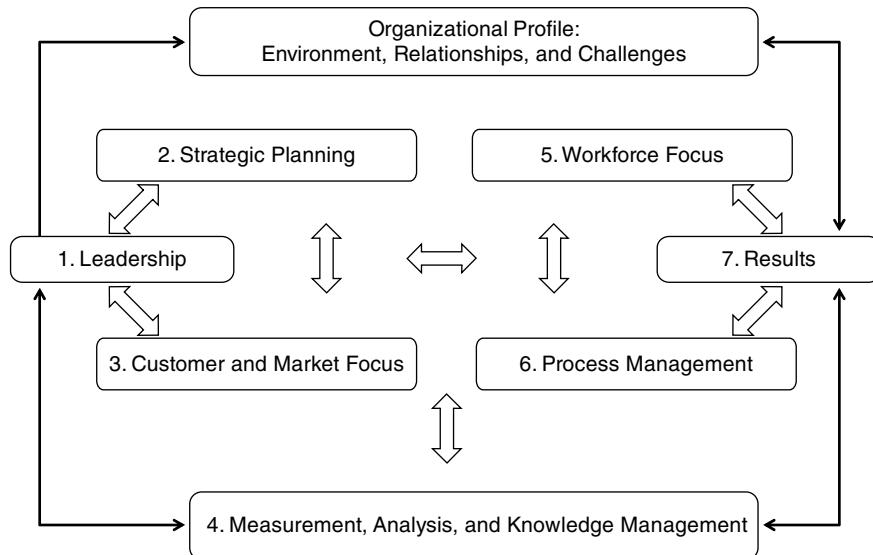
**Table 1.1** Few Core Business Processes

| <i>Resource Management</i>              |  |
|---|--|
| <b>Acquire to Retire (Fixed Assets)</b> | This end-to-end scenario includes processes such as Asset Requisition, Asset Acquisition, Asset utilization enhancement, Asset Depreciation, and Asset Retirement.   |
| <b>Hire to Retire</b>                   | An end-to-end process adopted by the HR unit/function of enterprises. It includes processes such as Recruitment, Hire, Compensation management, Performance management, Training, Promotion, Transfer, and Retire. |

| <i>Business Process Management</i>                                  |   |
|---|---|
| <i>Process</i>  | <i>Explanation</i>  |
| <b>Procure to Pay<br/>(Also referred to as Purchase to Payment)</b> | This end-to-end process encompasses all steps from purchasing goods from a supplier, to paying the supplier. The steps in the process include: determination of requirement → vendor selection (after comparison of quotations) → preparation of purchase order → release of purchase order → follow-up/amendments on purchase order → good received and inventory management → inspection of goods received → verification of invoice → issuance of payment advice   |
| <b>Idea to Offering</b>   | This entails research → concept, concept → commit, design → prototype, validate → ramp-up, monitor → improve, improve → end of life cycle   |
| <b>Market to Order<br/>(Also referred to as Market to prospect)</b> | This end-to-end process is as follows: research → market identification, market identification → plan, campaigning → lead, lead → order, account strategy → relationship management   |
| <b>Order to Cash</b>  | The basic activities of this core business process include: <ul style="list-style-type: none"> <li>• Create or maintain customer</li> <li>• Create or maintain material master</li> <li>• Create or maintain condition records for pricing</li> <li>• Enter and authorize sales order</li> <li>• Convert quote to order</li> <li>• Validate product configuration</li> <li>• Verify billing, shipping, payment details</li> <li>• Apply discounts</li> <li>• Review credit of customer</li> <li>• Verify stock</li> <li>• Plan delivery schedule</li> <li>• Pack and ship product</li> <li>• Billing and cash receipt</li> <li>• Contract to renewal</li> </ul> |
| <b>Quote to Cash</b>  | This process essentially is a two step process: <ul style="list-style-type: none"> <li>• Generation of quotations</li> <li>• Order to cash (described above)</li> </ul>   |
| <b>Issue to Resolution</b>  | The basic steps here are: detection of an issue → identification of the problem → development of solution → return/replace → closed loop feedback   |

We would like to present a simple but generic model from the “Malcolm Baldrige Performance Excellence Program”. This is a US national program based on the Baldrige Criteria for Performance Excellence. The Baldrige Criteria serves as a foundation for a performance management system. In this chapter, our focus is on the “Measurement, Analysis, and Knowledge Management” part of the Baldrige Criteria that helps you understand the strategic role of business analytics.

Figure 1.2 depicts the Malcolm Baldrige Criteria for Performance Excellence Framework. The Malcolm Baldrige Criteria for Performance Excellence Award was instituted in the USA to recognize businesses that achieve world-class quality and demonstrate excellence in business performance. It represents a common model to represent business enterprises. *Successful enterprises have a great leadership team. They focus on strategic planning (to deliver products or services or both). They study markets and customers intensely whom they serve. They set targets, measure achievements with rigor. They are highly focused on human capital management. They deploy processes (and automate using IT) and achieve planned business results. They repeat these steps every time consistently as well as constantly improve each of these components*



**Figure 1.2** Malcolm Baldrige Criteria for Performance Excellence framework.

## 1.2 BALDRIGE BUSINESS EXCELLENCE FRAMEWORK (OPTIONAL READING)

*Note: This section provides little deeper insight into the Baldrige business excellence framework. This is optional reading material.*

The Baldrige Criteria for Performance Excellence, depicted in Figure 1.2, are described below.

### 1. Leadership

2. Strategic Planning
3. Customer and Market Focus
4. Measurement, Analysis, and Knowledge Focus
5. Workforce Focus
6. Process Management
7. Results

We now discuss each one of them.

### 1.2.1 Leadership

The Leadership criterion describes how personal actions of the organization's senior leaders guide and sustain the organization and also how the organization fulfills its social, ethical, and legal responsibilities.

#### ***Senior Leadership***

Senior leadership includes the following:

- Guide the organization.
- Create a sustainable organization.
- Communicate and motivate employees.
- Focus on organizational performance.

#### ***Governance and Social Responsibility***

Corporate Governance and Corporate's social responsibility focus on

- Addressing organizational governance
  - a. Transparency in operations and disclosure practices.
  - b. Accountability for management's actions.
  - c. Evaluating the performance of CEO and members of governance board.
- Promoting legal and ethical behavior.
- Support of key communities.

### 1.2.2 Strategic Planning

The Strategic Planning criterion examines how the organization develops its strategic objectives and how these objectives are implemented.

#### ***Strategy Development***

The strategy development process of Strategic Planning includes the following:

- Determination of key strategic objectives with timelines.
- Balancing the needs of all stakeholders.
- Assessing organization's strengths, weaknesses, opportunities, and threats.
- Risk assessment.
- Changes in technology trends, markets, competition, and regulatory environment.

#### ***Strategy Deployment***

The strategy deployment process of Strategic Planning includes the following:

- Action plan development and deployment
  - a. Allocating resources for accomplishment of the action plan.
  - b. Allocating human resources to accomplish short-term or long-term goals.
  - c. Developing performance measures or indicators for tracking the effectiveness and achievement of the action plan.
- Performance projection: Addressing current gaps in performance with respect to competition.

### 1.2.3 Customer Focus

The Customer Focus criterion determines customer requirements, builds customer relationships, and uses customer feedback to improve and identify new opportunities for innovation.

#### ***Customer and Market Knowledge***

- Identification and determination of target customers and market segments.
- Using feedback to become more focused and satisfy customer needs.

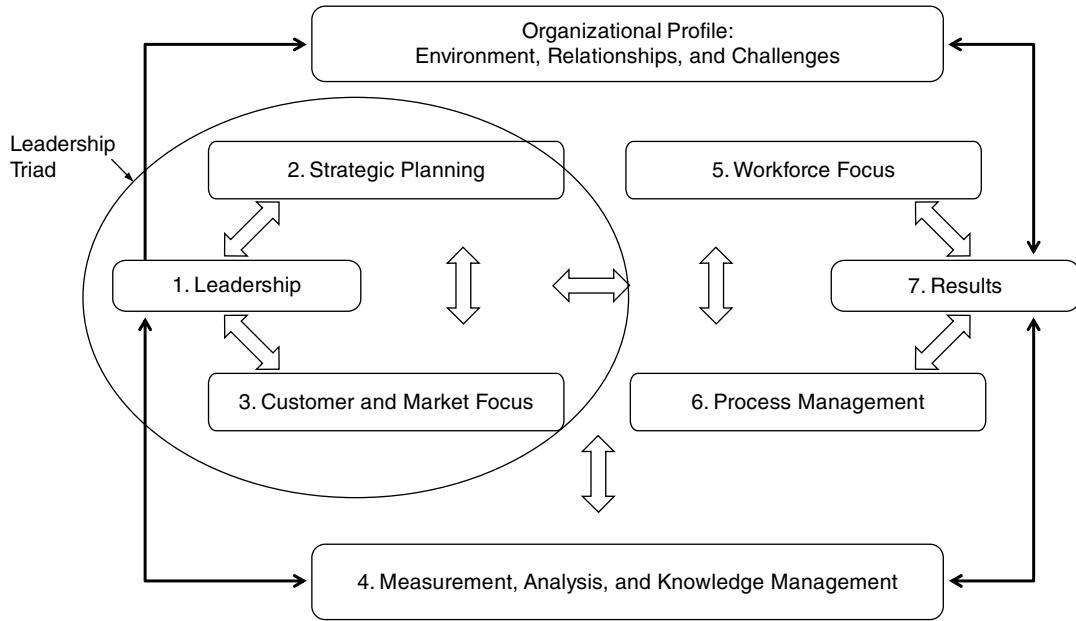
#### ***Customer Relationship and Satisfaction***

- Acquire new customers, meet, and exceed customer expectations.
- Implement processes for managing customer complaints effectively and efficiently.

**Leadership, Strategic Planning, and Customer and Market Focus constitute the Leadership Triad shown in Figure 1.3.**

### 1.2.4 Measurement, Analysis, and Knowledge Management

The Measurement, Analysis, and Knowledge Management criterion determines how the organization selects, gathers, analyzes, and improves its data, information, and knowledge assets and how it manages its information technology.



**Figure 1.3** The Leadership Triad.

### ***Measurement, Analysis, and Improvement of Organizational Performance***

- **Performance Measurement**
  - a. Collecting, aligning, and integrating data and information.
  - b. Tracking the overall performance of the organization and the progress relative to strategic objectives and action plans.
- **Performance Analysis, Review, and Improvement**
  - a. Assessing the organization's success, competitive performance, and progress on strategic objectives.
  - b. Systematic evaluation and improvement of value creation processes.

### ***Measurement of Information Resources, Information Technology, and Knowledge***

- **Management of Information Resources**
  - a. Making needed data and information available and providing access to workforce.
  - b. Keeping data and information availability mechanisms, including software and

hardware systems, in tune with business needs and technological changes in the operating environment.

- **Data, Information, and Knowledge Management**

Ensuring the following properties of the organizational data, information, and knowledge:

- a. Accuracy
- b. Integrity and reliability
- c. Security and confidentiality

### **1.2.5 Workforce Focus**

The Workforce Focus criterion examines the ability to assess the workforce capability and builds a workforce environment conducive to high performance.

#### ***Workforce Environment***

- Focusing on maintaining a good working environment for the employees.
- Recruiting, hiring, and retaining new employees.
- Managing and organizing the workforce.
- Ensuring and improving workplace health, safety, and security.

### **Workforce Engagement**

- Fostering an organizational culture conducive to high performance and a motivated workforce.
- Focusing on the training and education support in the achievement of overall objectives, including building employee knowledge, skills, and contributing to high performance.
- Reinforcing new knowledge and skills on job.
- Developing personal leadership attributes.

### **1.2.6 Process Management**

The Process Management criterion examines how the organization manages and improves its work systems and work processes to deliver customer satisfaction and achieve organizational success and sustainability.

### **Work System Design**

- Determines core competencies of the organization.
- Identifies key work processes.
- Determines key work process requirements incorporating inputs from customers.

- Designs and innovates work processes to meet key requirements.

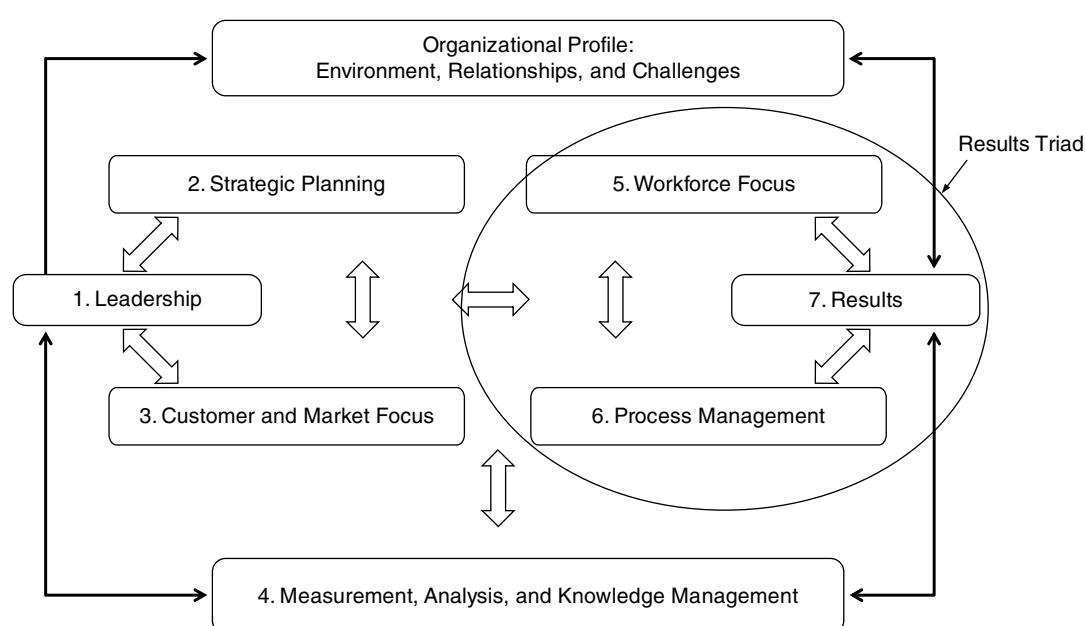
### **Work Process Management and Improvement**

- Implementing the work processes to ensure that they meet design requirements.
- Using key performance measures or indicators for control and improvement of work processes.
- Preventing defects, service errors, and re-work.

### **1.2.7 Results**

The Business Results criterion examines the organization's performance and improvement in all the key areas – product and process outcomes, customer focused outcomes, human resource outcomes, leadership outcomes, and financial outcomes. The performance level is examined against that of the competitors and other organizations with similar product offering.

**Workforce Focus, Process Management, and Results represent the Results Triad as shown in Figure 1.4.**



**Figure 1.4** The Results Triad.



## Remind Me

- The Malcolm Baldrige Criteria for Performance Excellence is a framework that any organization can use to improve its overall performance and achieve world-class quality.
- The criteria for performance excellence framework are:
  - a. Leadership
  - b. Strategic Planning
  - c. Customer and Market Focus
  - d. Measurement, Analysis, and Knowledge Focus
  - e. Business Focus
  - f. Process Management
  - g. Results
- The **Leadership** criterion describes how personal actions of the organization's senior leaders guide and sustain the organization and also how the organization fulfils its social, ethical, and legal responsibilities.
- The **Strategic Planning** criterion examines how the organization develops strategic objective and how these objectives are implemented.
- The **Customer Focus** criterion determines customer requirements, builds customer relationships and use customer feedback to improve and identify new opportunities for innovation.

- The Measurement, Analysis, and Knowledge Management criterion determines how the organization selects, gathers, analyzes, and improves its data, information, and knowledge assets and how it manages its information technology.
- The Workforce Focus criterion examines the ability to assess workforce capability and build a workforce environment conducive to high performance.
- The Process Management criterion examines how the organization manages and improves its work systems and work processes to deliver customer satisfaction and achieve organizational success and sustainability.
- The Results criterion examines the organization's performance and improvement in all the key areas.
- Leadership, Strategic Planning, and Customer and Market Focus constitute the Leadership Triad.
- Workforce Focus, Process Management, and Results represent the Results Triad.
- The horizontal arrow in the center of the framework (Figures 1.3 and 1.4) links the Leadership Triad to the Results Triad, a linkage critical to organizational success.



## Point Me (Books)

- *Comparing ISO 9000, Malcolm Baldrige, and the SEI CMM for Software: A Reference and*

*Selection Guide* [Paperback], Michael O Tingey.



### *Test Me Exercises*

1. \_\_\_\_\_ Baldrige Criteria serves as a foundation for the performance management system.
2. The \_\_\_\_\_ Baldrige criterion assesses workforce capability and capacity needs and builds a workforce environment conducive to high performance.
3. The Customer and Market Focus Baldrige criterion examines which of the following?
  - a. The requirements, needs, expectations of the customers
  - b. Customer satisfaction and loyalty
  - c. Customer retention
  - d. All the options are valid

#### **Solution:**

1. Measurement, Analysis, and Knowledge Management
2. Workforce Focus
3. All the options are valid.

## **1.3 KEY PURPOSE OF USING IT IN BUSINESS**

At the outset we will start with the idea that IT is used to automate business processes of departments/functions of the enterprise or core business processes of the enterprise that span across functions. The use of IT to innovate new ways of doing business results in huge benefits for the enterprise and gives it leadership position in the market.

All of us have used software like email, Google search tool, word processor, spreadsheet and presentation tools that enhance personal productivity. These are “office productivity/office automation” IT applications. Some office productivity applications may focus on team productivity and these are termed as “Groupware”. Intra-company email, team project management software, intranet portals, and knowledge management systems are typical examples of Groupware.

Consider an IT application like “Payroll processing”. We could consider payroll as a key IT application for HR function. The core purpose of the payroll IT application is to calculate the salary payable to all employees, generate pay slips, and inform the bank for funds transfer to the respective employee accounts. The same work can definitely be done manually. But the manual system works well when the number of employees is small and salary computation is simple. When you have, say, 10,000 employees to pay and the salary calculation needs to take care of several data pieces like leave, overtime, promotions, partial working, and so on, and the entire work needs to be done very quickly, you will face challenges. There could be computational errors and wrong assumptions that would lead to delay and dissatisfaction. However, when the salary calculation is automated, the benefits gained will be speed, accuracy, and the ability to repeat the process any number of times. These IT applications also help in fast information retrieval and generate statutory reports immediately. Such a type of application could be termed as **Departmental IT Applications**. A payroll IT application generates computed data, and the volume of such data increases with every successive payroll run. This kind of historical data is very useful in business analytics.

You may have come across some common IT applications like train/bus/ airline ticket reservation systems. These are called **On-line Transaction Processing Systems**. On an on-line train ticket reservation system, you will never face the problem of duplicate seat allocation (that is, allocation of the same seat to more than one passenger) and more allocation than the capacity, assuming that the data entered into the system is correct. The online transaction processing systems also generate large volumes of data, typically stored in RDBMS (relational database management system). You can imagine the volume of data which gets accumulated over one year, say, for all trains for all passengers who would have travelled! Again, this type of historical data is very useful for business analytics. When you enter your personal details to register yourself in a website, the data you provide will be stored in similar ways. This category of IT applications too provides benefits like speed, accuracy, search capability, status inquiry, easy update ability, and so on.

Let's turn our attention to newer types of applications such as the Amazon book store. This book store, through the combination of the Internet technology and book store software management, has completely changed the way you shop for a book. Here there is really no book store(no physical book store) that you could visit, stand in front of the book collections in an aisle, browse through the catalog or search the book shelves, pick a book, get it billed, and carry it home, as you do in a normal book store. Applications like the one used at the Amazon book store are called **Business Process/Model Innovation** applications. The ability to transfer money across the globe to any bank account from home is a new innovation. Lastly, let's look at the IT applications that help us in decision making. These applications provide facts that enable us make objective decisions. Suppose you would like to buy a new two-wheeler. One option for you is to visit showrooms of all brands, test-drive different vehicles, and collect data sheets of various options. Another option is to feed in an IT system inputs like your budget, expected mileage, availability, etc. and the system would present the best options as a comparison. Certainly, the second option would help you make informed decisions faster. This is possibly one of the best possible ways to use IT. The IT applications which help in decision-making are called **Decision Support** applications. These IT applications need to gather data from various sources, fit them into the user preferences, and present recommendations with the ability to show more details when the user needs it. These applications are very useful in business analytics.

Summarizing this topic, we can say IT applications are of various categories. Office productivity/office automation IT applications, Departmental IT applications, Enterprise level on-line transaction processing applications, Business Process or Model Innovation applications, and Decision Support applications. All these generate data of varying quality. The data is likely to be in different formats. These applications may be developed using different programming languages like C, Pascal, C++, Java, or COBOL and run on different operating systems. They may even use different vendors for the RDBMS package like MS SQL, Oracle, Sybase, or DB2. We will examine these issues in subsequent sections.

## 1.4 THE CONNECTED WORLD: CHARACTERISTICS OF INTERNET-READY IT APPLICATIONS

---

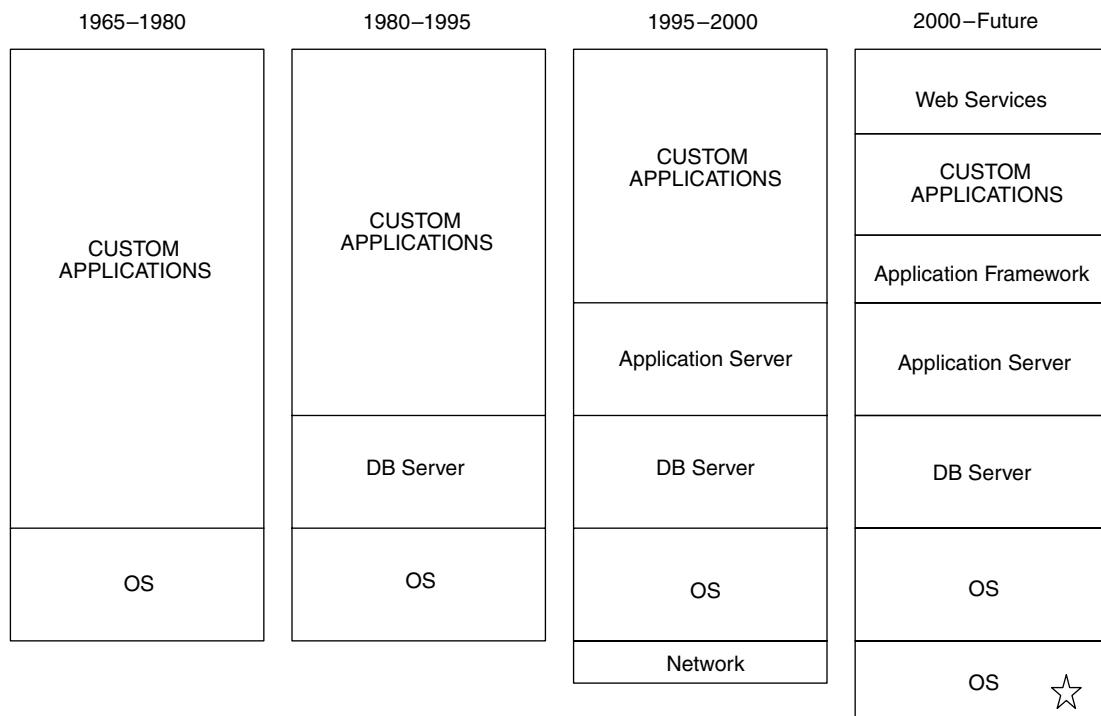
Computer systems have evolved from centralized mainframes with connected terminals to systems connected in Local Area Network (LAN) to distributed computing systems on the Internet. A large enterprise may have all these generations of computing resources.

We live in the connected world today. Computer networks are the basic foundation platform on which the entire IT infrastructure is built. From the enterprise perspective, majority of users access the enterprise and departmental applications either over LAN or through Internet in a secure fashion. Most

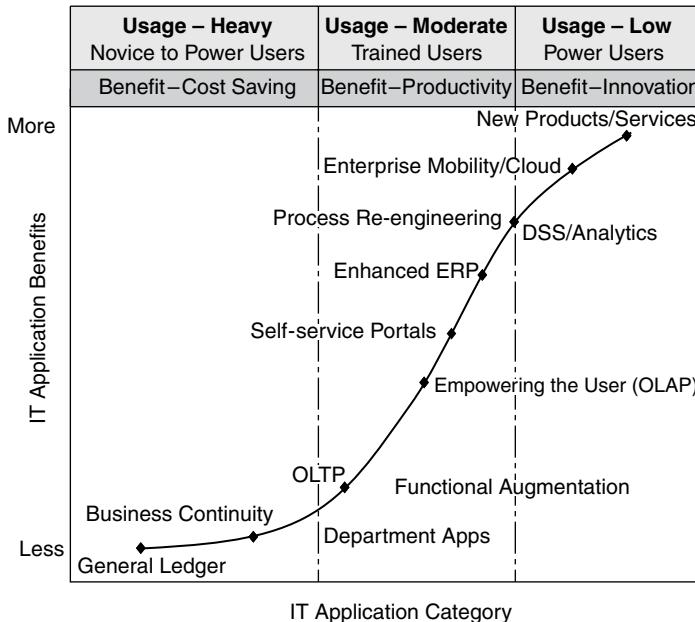
individual-centric data is stored in laptops or tablet computers. All enterprise data that is critical to the business is invariably stored on server systems spread across several locations with built-in mechanisms for disaster recovery in the event of failure of any server. These servers are secured with several hardware/software security solutions so that unauthorized access and theft of the enterprise data is prevented. Not everyone can install any software or copy any data on to these secure servers. Special software distribution and installation as well as anti-virus solutions are used to protect the servers from malicious hackers.

Refer to Figure 1.5. In the mainframe computing world, the system software consisted of an operating system and IT applications. Today, the scenario is entirely different. You will see several layers of system software that run on servers. These include operating system, network operating system, RDBMS software, application server software, middleware software products, and IT applications. In a typical enterprise with globally spread users, the IT infrastructure may consist of several data networks, routers, multi-vendor servers, multi-vendor OS, security servers, multiple RDBMS products, different types of application servers, and specific IT applications. Most users access enterprise data over LAN or secure Internet (VPN).

Enterprises strive to provide “single window” or “one-stop” access to several applications using techniques like “single sign-on”. In “single sign-on” once you have successfully logged onto the machine’s OS using a username and password, from that point forward any access to any application will automatically use the same username and password. There is a great emphasis on common look and feel or “user experience”. Gone are the days of command-driven user interface. It’s all about context sensitive menus, online help, navigation using touch screen/mouse, high resolution graphical displays. The “user experience” is a key driver for bringing computing facilities close to all types of users. Refer to Figure 1.6.



**Figure 1.5** Technology Centric Applications – From 1965 to 2000 to future.



**Figure 1.6** IT applications: Cost focus, productivity focus, and opportunity focus.

Internet-ready applications do have several advanced common capabilities like:

- Support large number of users of different interests and abilities.
- Provides display on multiple devices and formats.
- Deployed on large secure servers through license management software.
- Support single sign-on and support special authentication and authorization requirements.
- Ability to run on any operating system.
- Ability to use/connect to any RDBMS for data storage.
- Could be implemented in multiple programming languages or combinations as well.
- Leverages enterprise storage capabilities and back-up systems.
- Supports extensive connectivity to different types of networks and Internet services.

In the context of our topic of interest, that is, business analytics, it's important to know that the data we are looking for may exist on any or many servers or different RDBMS products in different locations connected over different networks and very likely will be in different formats!!

## 1.5 ENTERPRISE APPLICATIONS (ERP/CRM, ETC.) AND BESPOKE IT APPLICATIONS

Large businesses have hundreds of business processes and frontier areas to be enabled by IT for competitive advantage. This results in several IT applications being deployed in the enterprise. There are several ways in which we could look at the classification of these IT applications. We have already seen in the previous section a classification based on function and purpose such as office automation, business innovation, decision support, etc. We could also look at classifying IT applications based on the hardware platform they run like mainframe, open system, or Internet (distributed servers).

Another key classification parameter for IT applications is the application users. It's common to look at applications as customer facing, partner/supplier facing, and employee facing applications. This classification has also resulted in bringing focus on core processes that enable the enterprise to serve the customer. Customer relationship management, supply chain management, human capital management, financial accounting, customer order processing, customer order fulfillment, product distribution management, procurement management, corporate performance management, business planning, inventory management, innovation management, and service or repair order management are some of the commonly chosen areas for IT enablement.

When it comes to IT enablement of any chosen business area, there are four different approaches enterprises typically follow. Each approach has its own merits and associated benefits as well as investments. An enterprise may choose to develop and implement an IT application with its dedicated team, as they are the ones who understand business process steps very well. Such enterprises will invest in hardware, software, development tools and testing tools, and own the entire software lifecycle responsibility. This is termed as **bespoke application development**. Alternately, an enterprise may design an IT application and implement the tested application with its team but outsource the software development part. Some enterprises may choose to procure licenses for globally standard software and implement it with some customization. This approach is faster, brings global processes experience in the package but may not be entirely flexible to meet enterprise processes. The new trend is to entirely outsource the processes so that the chosen partner will take care of hardware, software, tools, development and maintenance team, and so on. It is not uncommon to find an enterprise that uses a hybrid model leveraging all these approaches.

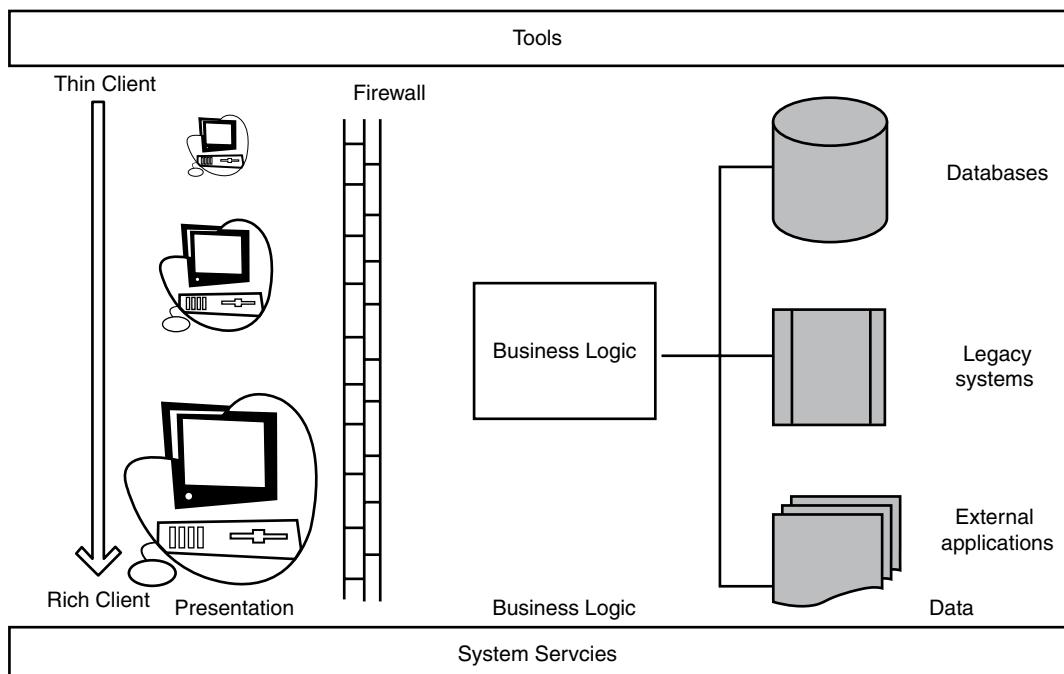
Enterprise Resource Planning (ERP) is the critical business IT applications suite that many enterprises choose to procure, customize, and implement. The most attractive element of ERP is the integrated nature of the related applications. Let's understand the usefulness of integrated applications. Consider a large enterprise that has several functions to be enabled by IT applications, like human capital management, purchase management, inventory management, and so on. One approach could be to identify each of the business areas through business analysis and follow the software engineering lifecycle to develop IT application for each of the areas. Of course, different teams will work on different business areas. Refer to Figure 1.7. All IT applications will invariably have application-specific user interface, business logic associated with the areas of business operation, and a data store that holds master and transaction data of the applications. It is not hard to imagine that the business processes are not isolated. This approach results in many islands of information and lots of redundant data being stored in each application. ERP prevents this by designing common integrated data stores and applications that can smoothly exchange data across them.

It is important to understand that enterprises maintain different IT infrastructure platforms to preserve continuity of IT applications, their availability and performance. It is a common scenario in enterprises to have three or more IT application-related infrastructures. Typically an IT application development infrastructure will have smaller capacity servers but powerful software development and testing tools. The development environment will also have team project management, communication, and knowledge repository management software. Enterprises dedicate independent IT infrastructure for IT application software testing. The test set-up will have more facilities for testing functionality, adherence to standards, performance testing, and user experience testing. Production systems of enterprises have large capacity robust IT infrastructure. Much like data centres, these will have large memory, storage, administration tools supported by service desks, security management systems, and so on. The production systems are hardly visible to the users but they can get their tasks accomplished in a smooth fashion. Enterprises take utmost care to protect the production environment from hackers; they migrate applications using tools from test environment to production environment with facility to roll-back in the event of malfunctioning of IT applications.

From the several ideas about enterprise IT applications discussed above, you can start visualizing that:

- All IT applications required to run an enterprise will never be created at the same time. There will be several generations of applications developed or purchased over several years.
- All IT applications need maintenance as business rules and business environment change, and each IT application will be at a different stage of maintenance at any given point of time. Enterprises migrate from small IT applications to more robust integrated suites as businesses grow.
- Enterprises choose the optimum hardware, software, OS, RDBMS, network and programming language available at the time of making build or buy decision. This results in heterogeneous hardware and multi-vendor software products in the same enterprise. Also, enterprises may acquire other businesses to expand their growth in non-linear ways, and this may again result in multiple technologies to co-exist in the same enterprise.
- The same way, enterprises may have ERP, home-grown applications, partially outsourced to fully out-sourced IT application development, etc.
- Enterprises may loosely bring together several applications in an enterprise portal but may have limitations in terms of data exchange across IT applications.
- Enterprises may design new applications to combine data from several critical data sources for the management information system (MIS).

Again we are emphasizing the data coming from different IT applications. Sometimes even though data is stored in a purchased application, it may be stored in a proprietary form and may not be accessible at all. You need to know the RDBMS schema in order to understand the organization of data and



**Figure 1.7** A typical enterprise application architecture.

design mechanisms to access such data. Data in different applications could be in different formats, duplicated or inconsistent in many ways. Unless we can get error-free data into one place, it will be difficult to use in a business analytics application that is of value to users.

## 1.6 INFORMATION USERS AND THEIR REQUIREMENTS

---

First of all, IT applications have matured to such a state today that it can typically be used by anyone. Considering the requirements of the enterprise to deliver right information to the right authorized user at the right time, users of IT applications may need to be given lots of attention. We have already indicated that “data” is really an asset of any enterprise and is a key ingredient in the recipe for a successful business. IT application users in the Internet age can be a whole lot of people roughly categorized into the following groups:

- Employees, partners, suppliers, customers, investors, analysts, prospective employees, and general public interested in the business affairs of the enterprise.
- Office goers, mobile users, home office users, remote securely connected users, casual visitors to website, and digital users who conduct transactions using the Internet.
- Business leaders, decision makers, operations managers, project managers, junior executives, and trainees who have a hierarchy.
- Role-based users who have access to certain category of IT applications, certain level of classified information, access to specific systems, and even specific operations they are allowed to perform.
- Securely connected users who may be allowed to access specific servers from a specific location during specified hours.
- Administrative users, who configure the IT environment, manage users' access control, execute anti-virus programs, perform anti-theft checks, install updates/upgrades, back-up enterprise data, and restore in the event of data corruption.
- Users who have permission to read or update or have full control over the enterprise information.
- Knowledge workers/analytical users who discover new patterns in the enterprise data to help businesses make innovative moves in the market place for competitive advantage.
- Multi-device access users who sometimes work in the office, move in the field, use different devices ranging from desktop systems to hand held smartphones to connect to the enterprise IT applications.

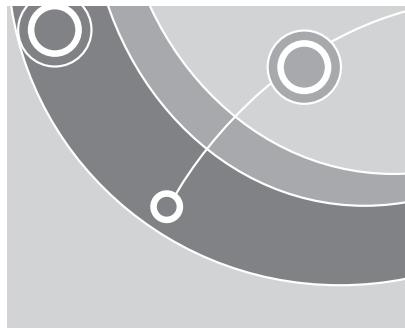
No matter how users access the enterprise data from various locations, they have certain common expectations. Some of these include:

- Smooth authentication and access to authorized resources.
- Availability of IT applications 24 × 7, 365 days.
- Speed of information delivery without having to wait long for response from systems.
- Ease-of-navigation and simple “user experience” with personalization capabilities.
- Secure delivery and transport of data to and from the IT applications.
- Secure transaction completion and roll-back especially involving monetary transactions.
- Consistent, accurate information, and ability to recover/restart from faults.
- Anytime, anywhere, any device-centric information delivery.

## UNSOLVED EXERCISES

---

1. Think of an industry of your choice such as retail, manufacturing, telecommunications, finance, etc. Identify the core business processes for the chosen industry.
2. Briefly explain the information users and their requirements.
3. Explain any Department IT application of your choice.
4. Explain the On-Line Transaction Processing system with a suitable example.
5. Explain any Decision Support application that you have come across.
6. What is your understanding of the connected world? Explain.
7. What is your understanding of business analytics? Explain.
8. Think about your college/university. What according to you are the applications that can be automated? Explain giving suitable examples.
9. What are some of the automated IT applications that are a great help to your college/university in its day-to-day functioning? Explain.
10. Can you think of one manual process/function that you wish were automated in your school/college/university?



# Case Study Briefs

This section presents three case briefs:

- GoodLife HealthCare Group
- GoodFood Restaurants Inc.
- TenToTen Retail Stores

Several examples/illustrations/assignments in the chapters to follow (Chapter 2–10) will refer to these case briefs. It will help to revisit these case briefs as and when you read the chapters that refer to them.

## GOODLIFE HEALTHCARE GROUP

### INTRODUCTION

---

**“GoodLife HealthCare Group”** is one of India’s leading healthcare groups. The group began its operations in the year 2000 in a small town off the south-east coast of India, with just one tiny hospital building with 25 beds. Today, the group owns 20 healthcare centers across all the major cities of India. The group has witnessed some major successes and attributes it to its focus on assembly line operations and standardizations. The group believes in making a “Dent in Global Healthcare”. A few of its major milestones are as listed below in chronological sequence:

- Year 2000 – the birth of the GoodLife HealthCare Group. Functioning initially from a tiny hospital building with 25 beds.
- Year 2002 – built a low cost hospital with 200 beds in India.
- Year 2004 – gained foothold in other cities of India.
- Year 2005 – the total number of healthcare centres owned by the group touched the 20 mark.
- The next five years saw the group’s dominance in the form of it setting up a GoodLife HealthCare Research Institute to conduct research in molecular biology and genetic disorders.
- Year 2010 witnessed the group bag the award for the “Best HealthCare Organization of the Decade”.

## BUSINESS SEGMENTS

---

GoodLife HealthCare offers the following facilities:

- Emergency Care 24 × 7.
- Support Groups.
- Support and help through call centers.

The healthcare group also specializes in orthopedic surgeries. The group has always leveraged IT to offer the best possible and affordable services to their patients. It has never hesitated in spending money on procuring the best possible machines, medicines, and facilities to provide the finest comfort to the patients. The doctors, surgeons, nurses and paramedical staff are always on the lookout for groundbreaking research in the field of medicine, therapy and treatment. Year 2005 saw the group establish its own Research Institute to do pioneering work in the field of medicine.

## ORGANIZATIONAL STRUCTURE

---

GoodLife HealthCare Group has a Board of Company's Directors at its helm. They are four in all – each being an exceptional leader from the healthcare industry. The group lives by the norm that a disciplined training is the key to success. The senior doctors and paramedical staff are very hands-on. They walk the talk at all times. Strategic decisions are taken by the company's board after cautious consultation, thorough planning and review. The strategic decisions are then conveyed to all the stakeholders such as the shareholders, vendor partners, employees, external consultants, etc. The group has acute focus on the quality of service being offered.

## QUALITY MANAGEMENT

---

The healthcare group spends a great deal of time and effort in ensuring that its people are the best. They believe that the nursing and paramedical staff constitute the backbone of the organization. They have a clearly laid out process for recruitment, training and on-the-job monitoring. The organization has its own curriculum and a grueling training program that all the new hires have to religiously undertake. Senior doctors act as mentors of the junior doctors. It is their responsibility to groom their junior partners. Every employee is visibly aware of the organization's philosophy and the way of life prevalent in the organization.

## MARKETING

---

GoodLife HealthCare Group had long realized the power of marketing. They have utilized practically all channels to best sell their services. They regularly advertise in the newspapers and magazines and more so when they introduce a new therapy or treatment. They have huge hoardings speaking about their facilities. They advertise on television with campaigns that ensure that viewers cannot help but sit through it. They have had the top saleable sportsperson to endorse their products. Lastly the group understands that "*word of mouth*" is very powerful and the best advertisers are the patient themselves who have been treated at one of the group's facilities.

## ALLIANCE MANAGEMENT

---

Over the years the GoodLife HealthCare Group has established a good alliance with specialist doctors, consultants, surgeons and physiotherapists, etc. The group has also built a strong network of responsible and highly dependable supplier partners. There is a very transparent system of communication to all its supplier partners. All its vendor partners are aware of the values and organization's philosophy that defines the healthcare group. The group believes in a win-win strategy for all. It is with the help and support from these vendor partners that the group is able to stock just the required amount of inventory and is able to procure the emergency inventory supplies at a very short notice. The group focuses only on its core business processes and out-sources the remaining processes to remain in control and steer ahead of competition.

## FUTURE OUTLOOK

---

GoodLife HealthCare Group is looking at expansion in other countries of the world too. They are also looking at growing in the existing markets. They are 27,000 employees today and are looking at growing to 60,000 in the next 5 years. The group already has a dedicated wing for the treatment of bone deformities. It aspires to set up a chemist store within its premises to make it convenient for their patients. They would like to set up an artificial limb center for the production of artificial limbs and rehabilitation of patients of orthopedic surgeries. The GoodLife group realizes its social obligation too and is looking forward to setting up a free hospital with 250 beds in a couple of year's time.

## INFORMATION TECHNOLOGY AT GOODLIFE GROUP

---

### Web Presence

- GoodLife group has excellent web presence: leveraging website, social networking and mobile device banner ads.
- Leverages Internet technology for surveys, targeted mailers.
- Self-help portal for on-line registration for treatment of ailments.

### Front Office Management

- Patient relationship management.
- Alliance Management.
- Registration and discharge of patients.
- Billing.
- Help Desk.

## HUMAN CAPITAL MANAGEMENT & TRAINING MANAGEMENT

---

- Employee satisfaction surveys.
- Employee retention program management.
- Employee training and development program management.

## Data Center Hosted Application

- Finance and Accounting, Corporate performance.
- Inventory Management.
- Suppliers and Purchase.
- Marketing and Campaign Management.
- Funnel and Channel analysis application.

## Personal Productivity

- Email, web access, PDA connect.
- Suggestions.
- Quick surveys.
- Feedback from communities of patients and also the communities of specialist doctors, consultants and physiotherapists.

## GOODFOOD RESTAURANTS INC.

### INTRODUCTION

---

*GoodFood Restaurants* is a very well known concept restaurants chain head quartered in the USA. It started its operations in the year 1999 in the USA, the UK, and Australia. The restaurant chain is led by a team of five experienced professionals; their leaders are amongst the best in the hospitality industry. The new group has had several significant triumphs and is an established benchmark in the industry for its “Quality of Dining Experience”. Some of the major milestones in the growth of GoodFood Restaurants’s business include:

- Year 1999 – Launch of 8 concept restaurants.
- Year 2001 – Serves over 100 popular standard items across the chain.
- Year 2002 – Leveraging IT in 5 strategic functions, globally.
- Year 2003 – Awarded the “Global Dining Experience Award”.
- Year 2005 – Touched 100 restaurants with IPO.
- Year 2006 – Set-up “International Master Chefs School”.
- Year 2009 – Completed expansion to 10 countries.

### BUSINESS SEGMENTS

---

Serving only Continental, Chinese, and Italian cuisines during its initial launch days, it is a different world today at GoodFood Restaurants. The dynamic management team was always listening to the voice of the customers and made several quick moves to expand its business. They were the first to introduce “Dial your Dinner” for take-away service. GoodFood delivered to customers’ expectations providing “fast turnaround time” for orders especially during the weekend evenings when almost all from today’s workforce desired a quick pack home. No complaints were received about wrong items, short supplied items, or any customization requests. Take-away segment is a separate line with pre-defined menu, service charges, and multi-channel approach to customer order processing, delivery, feedback collection. IT applications track accurately all transactions and monitor the delivery carefully to ensure customer satisfaction. Its dine-in restaurants are a great crowd puller and boast of massive reservations for theme dinners that are advertised. GoodFood has leveraged technology to ensure that customers are able to make their choice of restaurant, menu, table, etc. prior to coming into the restaurant. This is the main cash cow for the chain. The recent entry into catering in cruise liners has given a boost to GoodFood’s business. Looking at the sharp peaking growth during holiday seasons, GoodFood has established specialized processes and has leveraged IT to ensure zero-defects catering across several ports in the world.

### IMPECCABLE PROCESSES AND STANDARD CUISINE

---

GoodFood invested a lot into standardization of recipe, equipment, and training of chefs. The automated kitchens, access to standard processes, continuous tracking of quality has helped the restaurant chain to consistently deliver “Dining Experience” across the chain. Customers would never be caught

by “surprise” and the stewards on the floor were empowered to make decisions to delight the customer. The team of chefs constantly exchange best practices, pit falls, innovations, and improvements. A lot of data has been collected over years that serve as a goldmine for new chefs entering into the system. Master Chefs have developed their own innovations that have received global awards and recognition.

## **MARKETING**

---

GoodFood systematically invests in marketing its brand, listens to customers, and develops campaigns to attract different segments of customers. The campaigns are known for personalization, fast and impactful briefing, and demonstrating value for money. In the year 2005, the Management changed its corporate priorities. The order became (a) *customer delight*, (b) *return on assets*, and (c) *market share*. All regions began reposting in standard formats and reviews were fact-based. All investment decisions were based on data. The chain effectively leveraged customers as their brand ambassadors, globally. Every function made innovations to enhance customer delight and maximize return on assets that led to the realization of the third goal (that of increasing their market share).

## **SUPPLIER MANAGEMENT**

---

GoodFood has continually relied on loyal and collaborative suppliers to help achieve success in their endeavors. They have consciously kept the number of supplier’s low, automated the purchasing processes, and enhanced trust, to the extent of sometimes enjoying access to each other’s accounting records. This has helped GoodFood to maintain low levels of inventory, get the best available ingredients, replenish items quickly, plan lead times collaboratively, and alert each other of any changed situations. The team even benchmarked inventory management practices of auto spare parts dealers and drew inspiration from their innovative practices. The IT infrastructure management and mission-critical IT applications production run were all outsourced with standard service level agreements. The teams handled global expansion projects and delivered operating facilities ahead of schedule.

## **QUALITY MANAGEMENT**

---

The Management team at GoodFood is credited with conceptualizing and developing a “Leadership through Quality” program within 3 years of inception. The program has achieved global recognition several times. They are regarded “a quality company”. The focus has always been on providing both internal and external customers with products/services that duly meet their requirements. Quality is the prerogative of every employee. All functions have pre-defined performance standards, quality targets, rigor in data collection, and systematic analysis of deviation and tracking of corrective actions. Process metrics data represents a large body of knowledge at GoodFood.

## **ORGANIZATION STRUCTURE**

---

The head quarter focuses on strategic areas and collaborates with country operations by empowering the country management. Finance constantly provides the much-needed fuel for growth and shares the

corporate performance with investors. The planning and competition watch were integrated into the Finance function. IPO project was again led by Finance in collaboration with Marketing. Human capital management became a role model for hiring, empowerment, training, and internationalization of operations. Flat organization structure helped in enhanced transparency, sharing of knowledge, and collaborative workforce development. Management persistently leverages technology for fast decision making and remaining on-the-same-page with global leaders. Standard formats, frequency of information sharing, and objective communication have helped GoodFood retain its competitive edge and innovate constantly.

## FUTURE OUTLOOK

---

GoodFood is a socially conscious enterprise that has networked with communities to reduce food wastage, recycle waste to ensure safety and hygiene, etc. Given its firm belief for equal opportunity employment, global accounting standards, human capital management excellence, GoodFood is likely to acquire restaurants in theme parks, high-end resorts, and flight kitchen operations management.

## INFORMATION TECHNOLOGY AT GOODFOOD

---

### Web Presence

- GoodFood has excellent web presence: leveraging website, social networking and mobile device banner advertisements.
- Leverages Internet technology for surveys, targeted mailers, personalized invites, loyalty program management.
- Self-help portal for reservation, choice of seating tables, pre-ordering selections.

### Front Office Management

- Wireless Kitchen Order Tickets, guest relationship management, loyalty programs.
- POS, Billing, Charging, Promotion points redemption.
- Messaging, email, ambience control including piped-in music, video, news, alerts, weather.

### Team Collaboration

- Intranet, Scheduling, Search.
- Collaboration.

### Data Center Hosted Application

- Finance and Accounting, Corporate performance.
- Human Capital Management and Training Management.
- Inventory Management.
- Suppliers and Purchase.
- Marketing and Campaign Management.
- Menu, recipe, and global (product) descriptions.

## Personal Productivity

- Email, web access, PDA connect.
- Networking printing.
- Suggestion.
- Quick surveys.

## TENTOTEN RETAIL STORES

### INTRODUCTION

---

**TenToTen Retail Stores** is one of the world's leading distribution groups. The group began its operations in the year 1990 with just one grocery store setup titled "*Traditional Supermarkets*". Headquartered in the USA, the group now also operates out of the UK. Currently, it has four major grocery store setups: "Hypermarkets & Supermarkets", "Traditional Supermarket", "Dollar Store", and "Super Warehouse". In the 20 years since its inception, the group owns close to 10,000 stores either company operated or franchises. The group has witnessed some major successes and attributes them to its focus on enhancing the "quality of customer experience". The group believes in becoming "Better and not just Bigger". A few of its major milestones are as listed below in chronological sequence:

- Year 1990 – the birth of the TenToTen Retail Stores. Initially operated only out of the USA.
- Year 2000 – awarded the honour of "Best Managed Small-Sized Business".
- Year 2000 – gained foothold in the UK.
- Year 2005 – the total number of stores owned by the group touched the 5000 mark.
- The next five years saw the group's dominance in the form of it acquiring another 5000 stores.
- Year 2010 witnessed the group bag the award for the "Best Employer of the Decade".

### BUSINESS SEGMENTS

---

The grocery Stores initially used to stock-up only grocery items such as confectionaries, food items such as grains, pulses, breads (Indian and Italian), fresh fruits, fresh vegetables, tinned food, dairy products, etc. Slowly and gradually, it started stock-piling garments for men, women, kids, and infants; electronic gizmos; video games; CDs/DVDs; gardening appliances; cosmetics; etc. Today, it also has fresh vegetable and fruit juice corners and a couple of fast food joints within each of its setup. The group has always believed in leveraging information technology to serve its customers better. They maintain a list of their premium customers and accord special discounts to their premium community. They have constantly paid heed to the customers' demands and desires through their "Hear Me – Voice of Consumer" service program. The group has a team of dedicated analysts who work round the clock to provide "Competitive Intelligence", "Customer Intelligence", etc. to help run the business smoothly and efficiently.

### ORGANIZATIONAL STRUCTURE

---

TenToTen Stores has a Board of Directors at its helm. They are six in all, each being an exceptional leader from the retail industry. The group lives by the norm that the top leaders will mentor their middle level, which in turn will mentor their operational managers. Strategic decisions are taken by the company's board after careful consultation, meticulous planning, and review. The strategic decisions are then conveyed to all the stakeholders such as the shareholders, vendor partners, employees, external consultants, customers, etc. The group has acute customer and market focus. Before launching a new product, the senior executives conduct a detailed study of the market and the various customer segments. The group also uses the services of a few external consultants to keep them abreast of competition.

## MARKETING

---

TenToTen Stores had realized the power of marketing while taking its baby steps into the business world. They have utilized practically all channels to best sell their products. They regularly advertise in the newspapers and magazines and more so when they introduce a new product or announce a promotional scheme. They have huge hoardings speaking about their products and the shopping experience that awaits the customers. They advertise on television with campaigns that ensure that viewers cannot help but sit through it. They have had the top saleable cinestars and sportsperson to endorse their products. Lastly the group understands that "*word of mouth*" is very powerful and the best advertisers are the customers themselves. A lot has gone into making customers their brand ambassadors.

## SUPPLIER MANAGEMENT

---

TenToTen Stores has built a strong network of trustworthy and highly reliable supplier partners. Their's is a very transparent system of communication to all the supplier partners. All their vendor partners are aware of the values and principles that define this group. The group believes in a win-win strategy for all. The group focuses only on its core business processes and out-sources the remaining processes to remain in control and steer ahead of competition.

## QUALITY MANAGEMENT

---

TenToTen Stores has quality processes in place to assess the quality of products. They have clearly laid out processes to house the inventory – how much and where. They have clear guidelines on how to stock the various products – market basket. The products are classified very well and are stocked in separate sections such as "Dairy Products", "Electronic Goods", "Traveller's accessories", "Home Appliances", etc. Clear guidelines also exist for on-boarding employees before releasing them to the shop floor. Clearly, "Quality is the differentiator". The group also hires services of external agencies to garner customer feedback and employee satisfaction feedback.

## FUTURE OUTLOOK

---

The group has employed several analysts, whose major tasks involve studying the markets, studying the customers buying behavior, the customer's demographics, when to announce discounts, how much discount to offer, which customer segment to target, etc. They are looking at growing in the existing markets as well as expansion to new territories such as Middle East. They have 57,000+ employees today and are looking at growing to 1,00,000 employees in the next 5 years.

TenToTen Stores is also aware of its social responsibilities and is constantly looking at means to conserve power/energy, reducing the use of polythene/plastic bags, etc. They have started the practice of using recycled paper bags. Several tourists and travellers also visit the stores owned by TenToTen Stores. The group wants to make it easy for them to make their purchases by allowing the "Money/Travellers Cheques Exchange" at their stores. Today is an era of digital consumers; TenToTen Stores will be focusing on creating brand ambassadors in the digital space too.

## INFORMATION TECHNOLOGY AT TENTOTEN STORES

---

### Web Presence

- TenToTen group has excellent web presence: leveraging website, social networking, and mobile device banner ads.
- Leverages Internet technology for surveys, targeted mailers, personalized invites, loyalty program management.
- Self-help portal for on-line purchases.

### Front Office Management

- Customer relationship management.
- POS, Billing, Charging, Promotion points redemption.
- Customer Help Desk.

### Human Capital Management and Training Management

- Employee satisfaction surveys.
- Employee retention program management.
- Employee training and development program management.

### Data Center Hosted Application

- Finance and Accounting, Corporate performance.
- Inventory Management.
- Suppliers and Purchase.
- Marketing and Campaign Management.
- Funnel and Channel analysis application.
- Voice of Consumer application – to collect feedback.

### Personal Productivity

- Email, web access, PDA connect.
- Suggestions.
- Quick surveys.



# 2



## Types of Digital Data

---

### BRIEF CONTENTS

|                                  |                                      |
|----------------------------------|--------------------------------------|
| What's in Store                  | Getting to Know Unstructured Data    |
| Introduction                     | Getting To Know Semi-Structured Data |
| Getting into “GoodLife” Database | Difference Between Semi-Structured   |
| Getting to Know Structured Data  | and Structured Data                  |
|                                  | Unsolved Exercises                   |

---

### WHAT'S IN STORE

Today, data undoubtedly is an invaluable asset of any enterprise (big or small). Even though professionals work with data all the time, the understanding, management and analysis of data from heterogeneous sources remains a serious challenge.

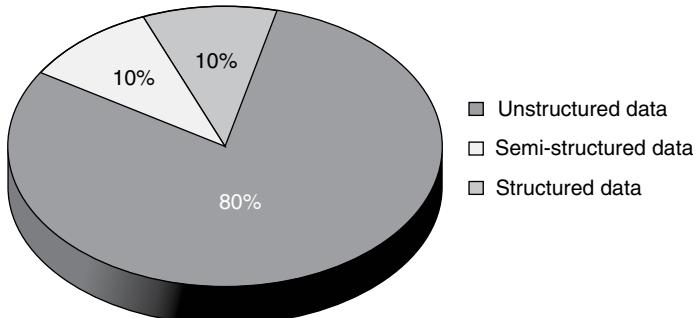
This chapter is a “Must Read” for first-time learners interested in understanding the role of data in business intelligence. In this chapter, we will introduce you to the various formats of digital data (structured, semi-structured and unstructured data), data storage mechanism, data access methods, management of data, the process of extracting desired information from data, challenges posed by various formats of data, etc.

We suggest you refer to some of the learning resources suggested at the end of this chapter and also complete the “Test Me” exercises. You will get deeper knowledge by interacting with people who have shared their project experiences in blogs. We suggest you make your own notes/bookmarks while reading through the chapter.

---

### 2.1 INTRODUCTION

Data growth has seen exponential acceleration since the advent of the computer and Internet. In fact, the computer and Internet duo has imparted the digital form to data. Digital data can be classified into three forms:



**Figure 2.1** Distribution of digital data in three forms.

- Unstructured.
- Semi-structured.
- Structured.

Usually, data is in the unstructured format which makes extracting information from it difficult. According to Merrill Lynch, 80–90% of business data is either unstructured or semi-structured. Gartner also estimates that unstructured data constitutes 80% of the whole enterprise data. Here is a percent distribution of the three forms of data as shown in Figure 2.1. A detailed explanation of these forms will follow subsequently.

- **Unstructured data:** This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program. About 80–90% data of an organization is in this format; for example, memos, chat rooms, PowerPoint presentations, images, videos, letters, researches, white papers, body of an email, etc.
- **Semi-structured data:** This is the data which does not conform to a data model but has some structure. However, it is not in a form which can be used easily by a computer program; for example, emails, XML, markup languages like HTML, etc. Metadata for this data is available but is not sufficient.
- **Structured data:** This is the data which is in an organized form (e.g., in rows and columns) and can be easily used by a computer program. Relationships exist between entities of data, such as classes and their objects. Data stored in databases is an example of structured data.

## 2.2 GETTING INTO “GOODLIFE” DATABASE

---

Refer to the case brief on “GoodLife HealthCare Organization”. Everyday “GoodLife” witnesses enormous amounts of data being exchanged in the following forms:

- Doctors’ or nurses’ notes in an electronic report.
- Emails sharing information about consultations or investigations.
- Surveillance system reports.
- Narrative portions of electronic medical records.
- Investigative reports.
- Chat rooms.

| GoodLife Healthcare<br>Patient Index Card |    |                |    |
|---|----|----------------|----|
| Patient ID                                | <> | Date           | <> |
| Nurse Name                                | <> |                |    |
| Patient Name                              | <> | Patient Age    | <> |
| Body Temperature                          | <> | Blood Pressure | <> |

**Figure 2.2** A snapshot of structured data.

“GoodLife” maintains a database which stores data only in a structured format. However, the organization also has unstructured and semi-structured data in abundance.

Let us try to understand the following aspects of GoodLife data: Where is each type of its data present? How is it stored? How is the desired information extracted from it? How important is the information provided by it? How can this information augment public health and healthcare services?

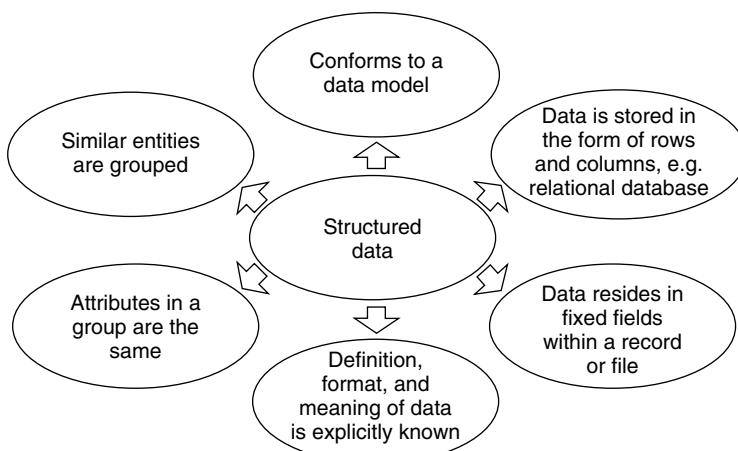
## 2.3 GETTING TO KNOW STRUCTURED DATA

Let us start with an example of structured data. The patient index card shown in Figure 2.2 is in a structured form. All the fields in the patient index card are also structured fields.

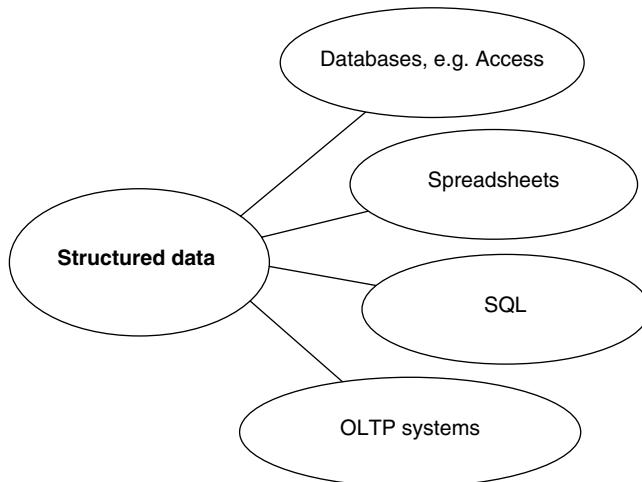
“GoodLife” nurses make electronic records for every patient who visits the hospital. These records are stored in a relational database. For example, nurse Anne records the body temperature and blood pressure of a patient, Dexter, and enters them in the hospital database. Dr. Brian, who is treating Dexter, searches the database to know his body temperature. Dr. Brian is able to locate the desired information easily because the hospital data is structured and is stored in a relational database.

### 2.3.1 Characteristics of Structured Data

Structured data is organized in semantic chunks (entities) with similar entities grouped together to form relations or classes. Entities in the same group have the same descriptions, i.e. attributes.



**Figure 2.3** Characteristics of structured data.



**Figure 2.4** Sources of structured data.

Descriptions for all entities in a group (schema)

- have the same defined format,
- have a predefined length,
- and follow the same order.

Figure 2.3 depicts the characteristics of structured data.

### 2.3.2 Where Does Structured Data Come From?

Data coming from databases such as Access, OLTP systems, SQL as well spreadsheets such as Excel, etc. are all in the structured format. Figure 2.4 depicts the sources of structured data.

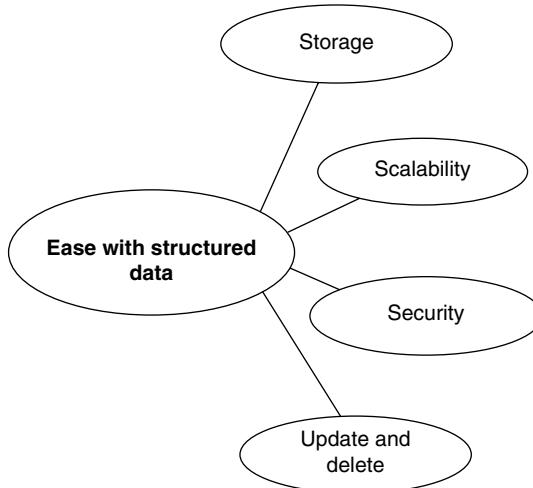
To summarize, structured data

- Consists of fully described data sets.
- Has clearly defined categories and sub-categories.
- Is placed neatly in rows and columns.
- Goes into the records and hence the database is regulated by a well-defined structure.
- Can be indexed easily either by the DBMS itself or manually.

### 2.3.3 It's So Easy With Structured Data

Working with structured data is easy when it comes to storage, scalability, security, and update and delete operations (Figure 2.5):

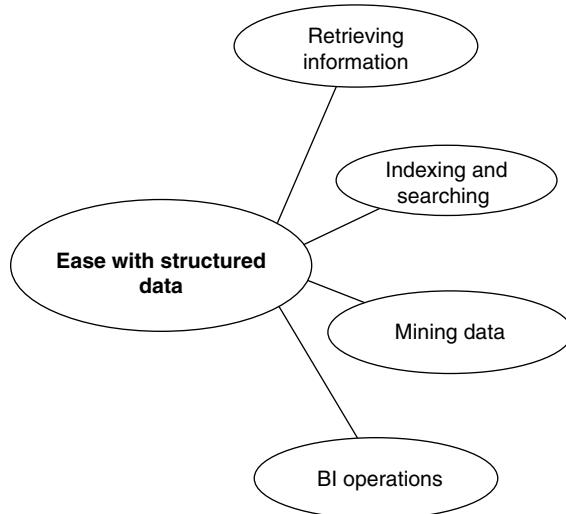
- **Storage:** Both defined and user-defined data types help with the storage of structured data.
- **Scalability:** Scalability is not generally an issue with increase in data.
- **Security:** Ensuring security is easy.
- **Update and Delete:** Updating, deleting, etc. is easy due to structured form.

**Figure 2.5** Ease with structured data.

### 2.3.4 Hassle-Free Retrieval

You won't get headache while retrieving desired information from structured data thanks to its following features (Figure 2.6):

- **Retrieving information:** A well-defined structure helps in easy retrieval of data.
- **Indexing and searching:** Data can be indexed based not only on a text string but also on other attributes. This enables streamlined search.
- **Mining data:** Structured data can be easily mined and knowledge can be extracted from it.
- **BI operations:** BI works extremely well with structured data. Hence data mining, warehousing, etc. can be easily undertaken.

**Figure 2.6** Ease of retrieval of structured data.

## 2.4 GETTING TO KNOW UNSTRUCTURED DATA

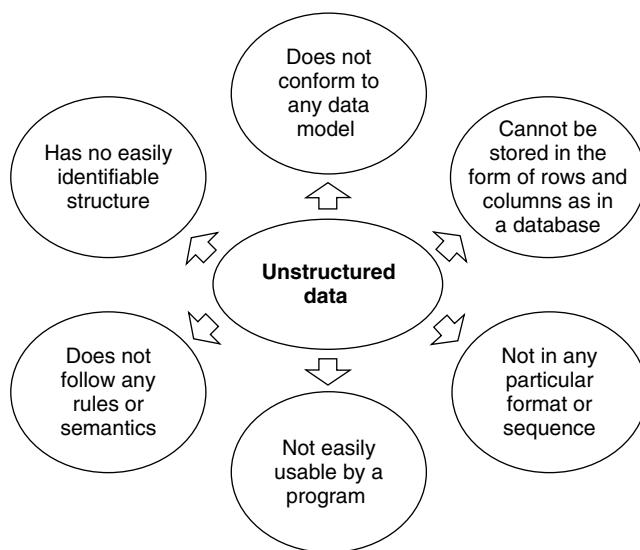
### Picture this...

Dr. Ben, Dr. Stanley, and Dr. Mark work at the medical facility of “GoodLife”. Over the past few days, Dr. Ben and Dr. Stanley had been exchanging long emails about a particular case of gastro-intestinal problem. Dr. Stanley has chanced upon a particular combination of drugs that has successfully cured gastro-intestinal disorders in his patients. He has written an email about this combination of drugs to Dr. Ben.

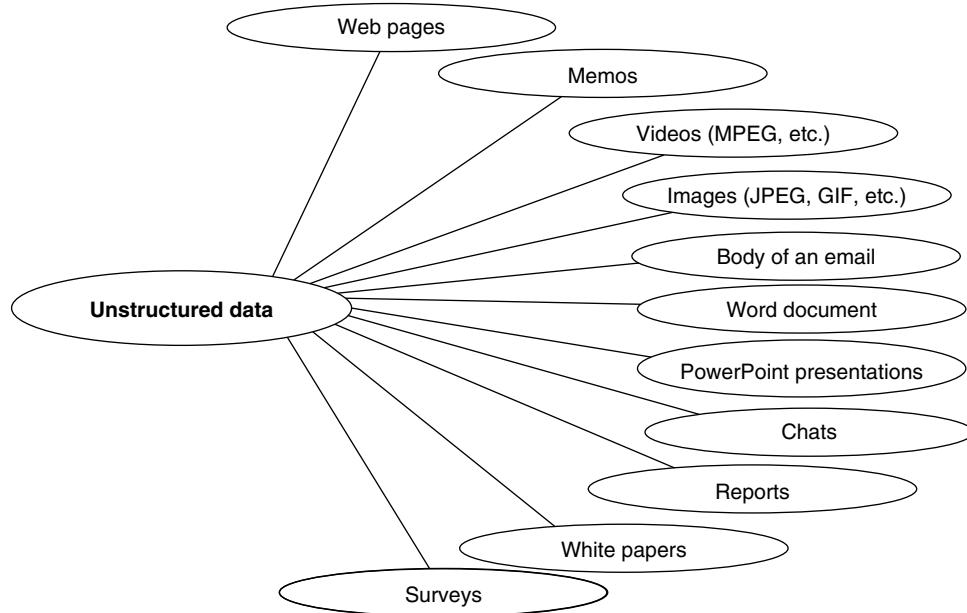
Dr. Mark has a patient in the “GoodLife” emergency unit with quite a similar case of gastro-intestinal disorder whose cure Dr. Stanley has chanced upon. Dr. Mark has already tried regular drugs but with no positive results so far. He quickly searches the organization’s database for answers, but with no luck. The information he wants is tucked away in the email conversation between two other “GoodLife” doctors, Dr. Ben and Dr. Stanley. Dr. Mark would have accessed the solution with few mouse clicks had the storage and analysis of unstructured data been undertaken by “GoodLife”.

As is the case at “GoodLife”, 80–85% of data in any organization is unstructured and is growing at an alarming rate. An enormous amount of knowledge is buried in this data. In the above scenario, Dr. Stanley’s email to Dr. Ben had not been successfully updated into the medical system database as it fell in the unstructured format.

Unstructured data, thus, is the one which cannot be stored in the form of rows and columns as in a database and does not conform to any data model, i.e. it is difficult to determine the meaning of the data. It does not follow any rules or semantics. It can be of any type and is hence unpredictable. The characteristics of unstructured data are depicted in Figure 2.7.



**Figure 2.7** Characteristics of unstructured data.



**Figure 2.8** Sources of unstructured data.

### 2.4.1 Where Does Unstructured Data Come From?

Broadly speaking, anything in a non-database form is unstructured data. It can be classified into two broad categories:

- **Bitmap objects:** For example, image, video, or audio files.
- **Textual objects:** For example, Microsoft Word documents, emails, or Microsoft Excel spreadsheets.

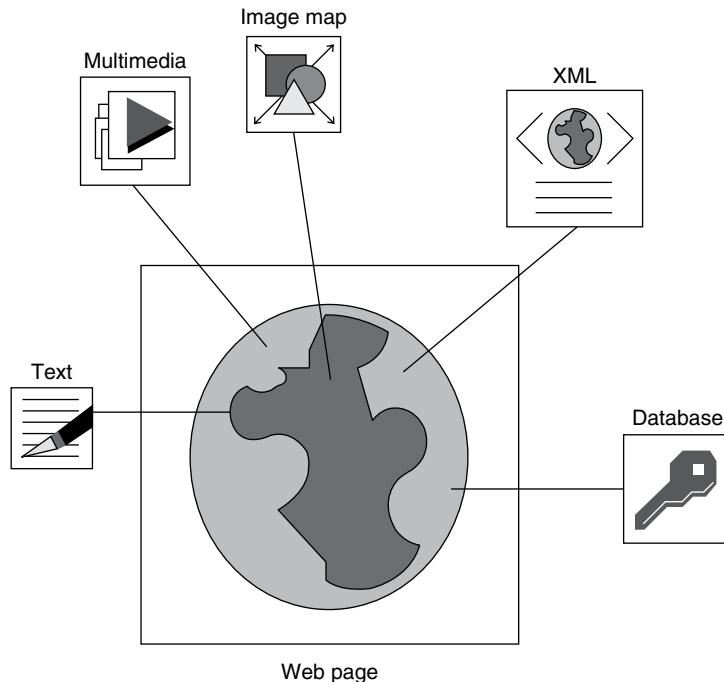
Refer to Figure 2.8. Let us take the above example of the email communication between Dr. Ben and Dr. Stanley. Even though email messages like the ones exchanged by Dr. Ben and Dr. Stanley are organized in databases such as Microsoft Exchange or Lotus Notes, the body of the email is essentially raw data, i.e. free form text without any structure.

A lot of unstructured data is also **noisy text** such as chats, emails and SMS texts. The language of noisy text differs significantly from the standard form of language.

### 2.4.2 A Myth Demystified

Web pages are said to be unstructured data even though they are defined by HTML, a markup language which has a rich structure. HTML is solely used for rendering and presentations. The tagged elements do not capture the meaning of the data that the HTML page contains. This makes it difficult to automatically process the information in the HTML page.

Another characteristic that makes web pages unstructured data is that they usually carry links and references to external unstructured content such as images, XML files, etc. Figure 2.9 is the pictorial representation of a typical web page.



**Figure 2.9** A typical web page.

### 2.4.3 How to Manage Unstructured Data?

Let us look at a few generic tasks to be performed to enable storage and search of unstructured data:

- **Indexing:** Let us go back to our understanding of the Relational Database Management System (RDBMS). In this system, data is indexed to enable faster search and retrieval. On the basis of some value in the data, index is defined which is nothing but an identifier and represents the large record in the data set. In the absence of an index, the whole data set/document will be scanned for retrieving the desired information.

In the case of unstructured data too, indexing helps in searching and retrieval. Based on text or some other attributes, e.g. file name, the unstructured data is indexed. Indexing in unstructured data is difficult because neither does this data have any pre-defined attributes nor does it follow any pattern or naming conventions. Text can be indexed based on a text string but in case of non-text based files, e.g. audio/video, etc., indexing depends on file names. This becomes a hindrance when naming conventions are not being followed.

- **Tags/Metadata:** Using metadata, data in a document, etc. can be tagged. This enables search and retrieval. But in unstructured data, this is difficult as little or no metadata is available. Structure of data has to be determined which is very difficult as the data itself has no particular format and is coming from more than one source.
- **Classification/Taxonomy:** Taxonomy is classifying data on the basis of the relationships that exist between data. Data can be arranged in groups and placed in hierarchies based on the

taxonomy prevalent in an organization. However, classifying unstructured data is difficult as identifying relationships between data is not an easy task. In the absence of any structure or metadata or schema, identifying accurate relationships and classifying is not easy. Since the data is unstructured, naming conventions or standards are not consistent across an organization, thus making it difficult to classify data.

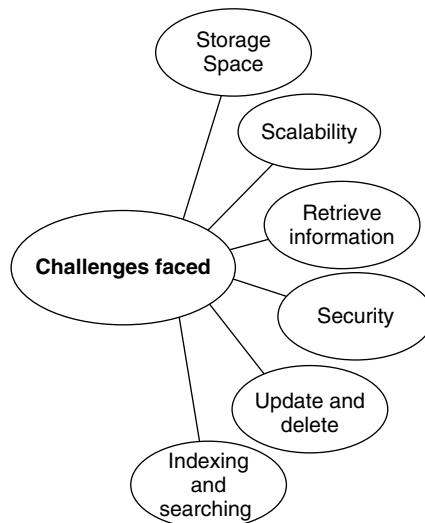
- **CAS (Content Addressable Storage):** It stores data based on their metadata. It assigns a unique name to every object stored in it. The object is retrieved based on its content and not its location. It is used extensively to store emails, etc.

Data management, however, does not end with the performance of above-mentioned tasks. This is merely the beginning. Provisions now need to be made to store this data as well. So the next question that comes to mind is: How to store this unstructured data?

#### 2.4.4 How to Store Unstructured Data?

The challenges faced while storing unstructured data are depicted in Figure 2.10 and listed below.

- **Storage space:** It is difficult to store and manage unstructured data. A lot of space is required to store such data. It is difficult to store images, videos, audios, etc.
- **Scalability:** As the data grows, scalability becomes an issue and the cost of storing such data grows.
- **Retrieve information:** Even if unstructured data is stored, it is difficult to retrieve and recover from it.
- **Security:** Ensuring security is difficult due to varied sources of data, e.g. emails, web pages, etc.
- **Update and delete:** Updating and deleting unstructured data are very difficult as retrieval is difficult due to no clear structure.
- **Indexing and searching:** Indexing unstructured data is difficult and error-prone as the structure is not clear and attributes are not pre-defined. As a result, the search results are not very accurate. Indexing becomes all the more difficult as the volume of data grows.



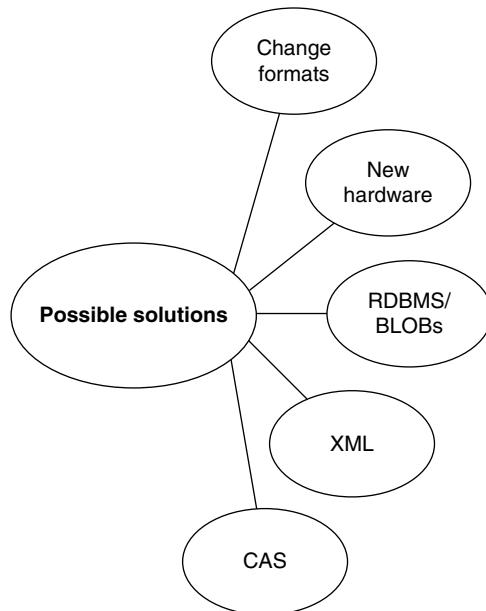
**Figure 2.10** Challenges faced while storing unstructured data.

## 2.4.5 Solutions to Storage Challenges of Unstructured Data

Now that we understand the challenges in storing unstructured data, let us look at a few possible solutions depicted in Figure 2.11 and described below:

- **Changing format:** Unstructured data may be converted to formats which are easily managed, stored and searched. For example, IBM is working on providing a solution which will convert audio, video, etc. to text.
- **Developing new hardware:** New hardware needs to be developed to support unstructured data. It may either complement the existing storage devices or may be a stand-alone for unstructured data.
- **Storing in RDBMS/BLOBs:** Unstructured data may be stored in relational databases which support BLOBs (Binary Large Objects). While unstructured data such as video or image file cannot be stored fairly neatly into a relational column, there is no such problem when it comes to storing its metadata, such as the date and time of its creation, the owner or author of the data, etc.
- **Storing in XML format:** Unstructured data may be stored in XML format which tries to give some structure to it by using tags and elements.
- **CAS (Content Addressable Storage):** It organizes files based on their metadata and assigns a unique name to every object stored in it. The object is retrieved based on its content and not its location. It is used extensively to store emails, etc.

XML or eXtensible Markup Language will be explained in detail in the “Semi-structured Data” section. We are now at a juncture where we have successfully made provisions to store as well as manage data that comes in unstructured format. But would merely the presence of this data in a database be sufficient to enable Dr. Mark give adequate medical treatment to his patient at the right time? The answer is “no”, because management and storage of unstructured data is not enough; we need to extract information from this data as well.



**Figure 2.11** Possible solutions for storing unstructured data.

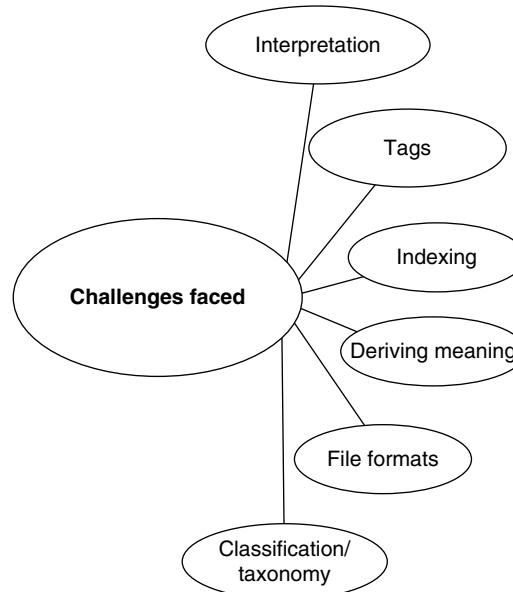
## 2.4.6 How to Extract Information from Stored Unstructured Data?

Let us again start off by looking at some of the challenges faced while extracting information from unstructured data. These challenges are depicted in Figure 2.12 and described below:

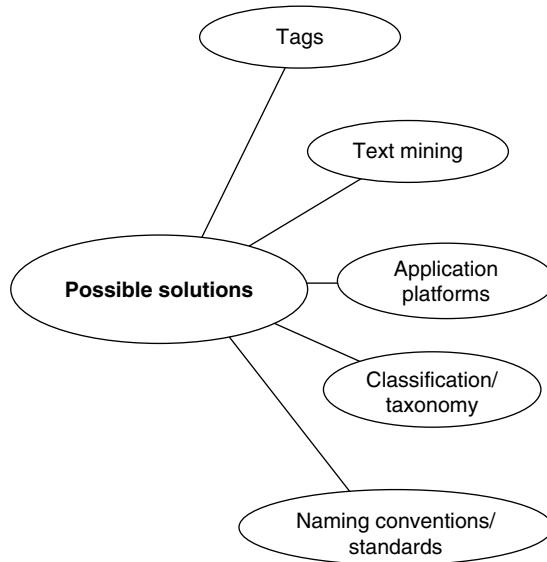
- **Interpretation:** Unstructured data is not easily interpreted by conventional search algorithms.
- **Classification/Taxonomy:** Different naming conventions followed across the organization make it difficult to classify data.
- **Indexing:** Designing algorithms to understand the meaning of the documents and then tagging or indexing them accordingly is difficult.
- **Deriving meaning:** Computer programs cannot automatically derive meaning/structure from unstructured data.
- **File formats:** Increasing number of file formats makes it difficult to interpret data.
- **Tags:** As the data grows, it is not possible to put tags manually.

The possible solutions to the challenges just mentioned are depicted in Figure 2.13 and described below:

- **Tags:** Unstructured data can be stored in a virtual repository and be automatically tagged. For example, Documentum provides this type of solution.
- **Text mining:** Text mining tools help in grouping as well as classifying unstructured data and assist in analyzing by considering grammar, context, synonyms, etc.
- **Application platforms:** Application platforms like XOLAP help extract information from email and XML-based documents.
- **Classification/Taxonomy:** Taxonomies within the organization can be managed automatically to organize data in hierarchical structures.



**Figure 2.12** Challenges faced while extracting information from stored unstructured data.



**Figure 2.13** Possible solutions for extracting information from stored unstructured data.

- **Naming conventions/standards:** Following naming conventions or standards across an organization can greatly improve storage, retrieval, index, and search.

#### 2.4.7 UIMA: A Possible Solution for Unstructured Data

UIMA (Unstructured Information Management Architecture) is an open source platform from IBM which integrates different kinds of analysis engines to provide a complete solution for knowledge discovery from unstructured data. In UIMA (depicted in Figure 2.14), the analysis engines enable integration and analysis of unstructured information and bridge the gap between structured and unstructured data.

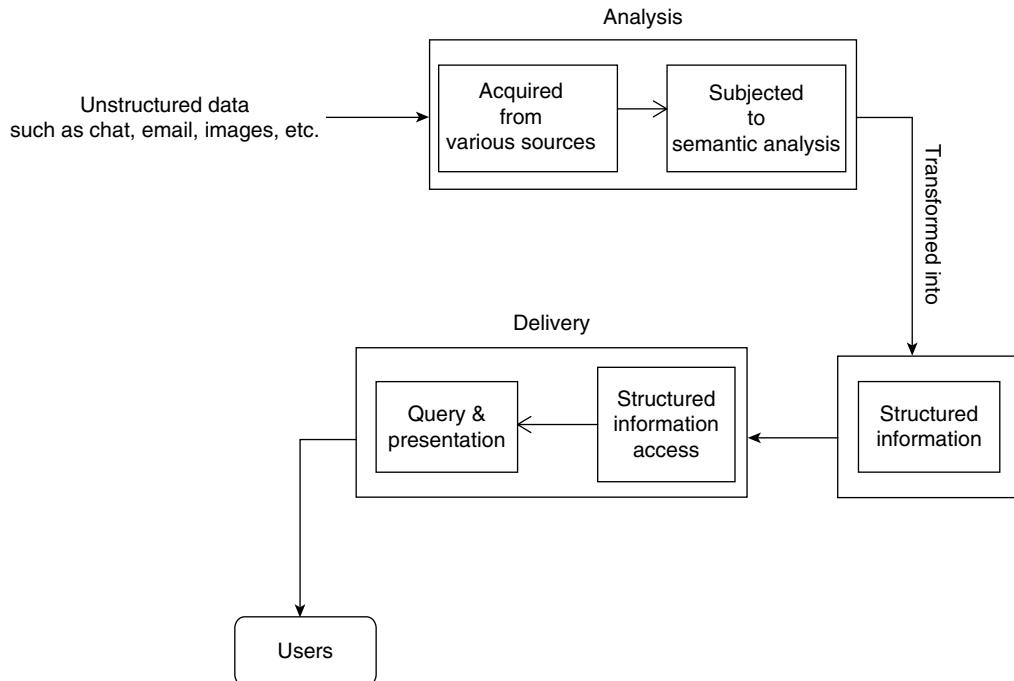
UIMA stores information in a structured format. The structured resources can be then mined, searched, and put to other uses. The information obtained from structured sources is also used for subsequent analysis of unstructured data. Various analysis engines analyze unstructured data in different ways such as:

- Breaking up of documents into separate words.
- Grouping and classifying according to taxonomy.
- Detecting parts of speech, grammar, and synonyms.
- Detecting events and times.
- Detecting relationships between various elements.

For more information refer to:

<http://www.research.ibm.com/UIMA/UIMA%20Architecture%20Highlights.html>

At this point we have discussed the management, storage, and analysis of unstructured data. But we are yet to deal with another prevalent digital data format, namely, semi-structured data.



**Figure 2.14** Unstructured Information Management Architecture (UIMA).

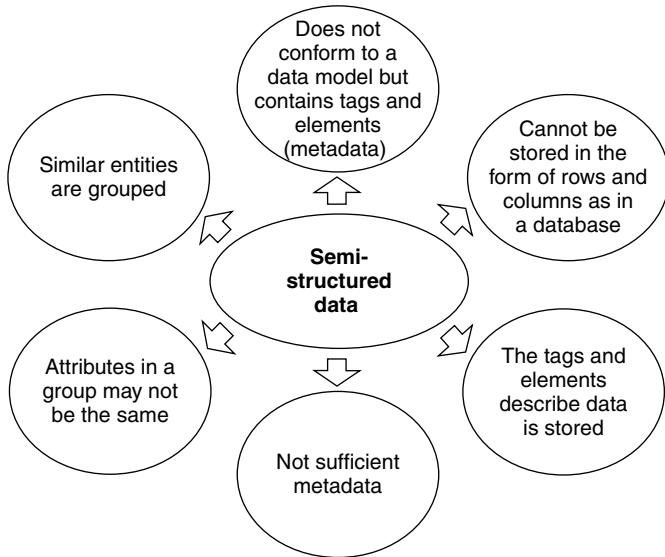
## 2.5 GETTING TO KNOW SEMI-STRUCTURED DATA

### Picture this...

Dr. Marianne of “GoodLife HealthCare” organization usually gets a blood test done for migraine patients visiting her. It is her observation that patients diagnosed with migraine usually have a high platelet count. She makes a note of this in the diagnosis and conclusion section in the blood test report of patients. One day another “GoodLife” doctor, Dr. B. Brian, searches the database when he is unable to find the cause of migraine in one of his patients, but with no luck! The answer he is looking for is nestled in the vast hoards of data.

As depicted in Figure 2.1, only about 10% of data in any organization is semi-structured. Still it is important to understand, manage, and analyze this semi-structured data coming from heterogeneous sources. In the above case of migraine patients, Dr. Marianne’s blood test reports on patients were not successfully updated into the medical system database as they were in the semi-structured format.

Semi-structured data, as depicted in Figure 2.15, does not conform to any data model, i.e. it is difficult to determine the meaning of this data. Also, this data cannot be stored in rows and columns as in a database. Semi-structured data, however, has tags and markers which help group the data and describe how the data is stored, giving some metadata, but they are not sufficient for management and automation of data. In semi-structured data, similar entities are grouped and organized in a hierarchy. The attributes or the properties within a group may or may not be the same.



**Figure 2.15** Characteristics of semi-structured data.

For example, two addresses may or may not contain the same number of properties/attributes:

```

Address 1
<house number><street name><area name><city>
Address 2
<house number><street name><city>

```

On the other hand, an email follows a standard format such as

|          |                                 |
|----------|---------------------------------|
| To:      | <Name>                          |
| From:    | <Name>                          |
| Subject: | <Text>                          |
| CC:      | <Name>                          |
| Body:    | <Text, Graphics, Images, etc. > |

Though the above email tags give us some metadata, the body of the email contains no format. Neither does it convey the meaning of the data it contains.

Now, consider the blood test report prepared by Dr. Marianne as shown in Figure 2.16. The blood test report is semi-structured. It has structured fields like Date, Department, Patient Name, etc. and unstructured fields like Diagnosis, Conclusion, etc.

Another example of semi-structured data are web pages. These pages have content embedded within HTML and often have some degree of metadata within tags. This automatically implies certain details about the data being presented.

Remember, there is a very fine line between unstructured and semi-structured data!!! Email, XML, TCP/IP packets, zipped files, etc. are semi-structured data as all have certain amount of metadata.

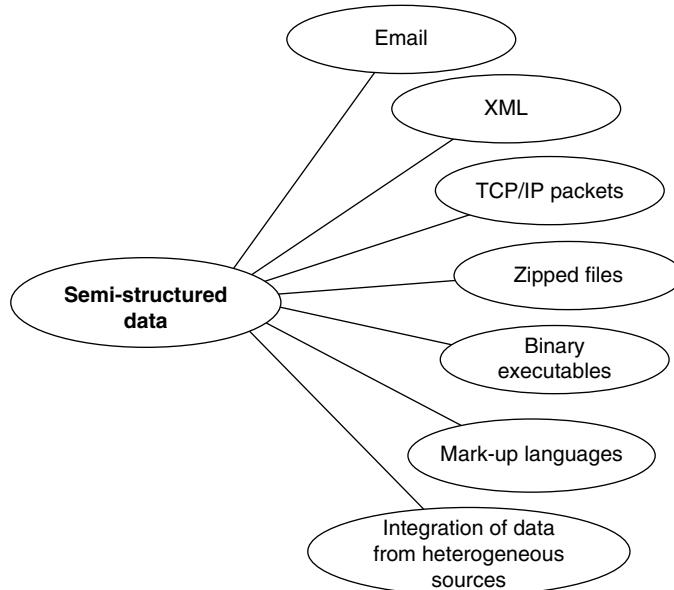
| ABC Healthcare<br>Blood Test Report |    |                  |    |
|-------------------------------------|----|------------------|----|
| Date                                | <> |                  |    |
| Department                          | <> | Attending Doctor | <> |
| Patient Name                        | <> | Patient Age      | <> |
| Hemoglobin content                  | <> |                  |    |
| RBC count                           | <> |                  |    |
| WBC count                           | <> |                  |    |
| Platelet count                      | <> |                  |    |
| Diagnosis <notes>                   |    |                  |    |
| Conclusion <notes>                  |    |                  |    |

**Figure 2.16** The blood test report, an example of semi-structured data.

### 2.5.1 Where Does Semi-Structured Data Come From?

The sources of semi-structured data are depicted in Figure 2.17. Characteristics of semi-structured data are summarized below:

- It is organized into semantic entities.
- Similar entities are grouped together.
- Entities in the same group may not have same attributes.



**Figure 2.17** Sources of semi-structured data.

- The order of attributes is not necessarily important.
- Not always all attributes are required.
- Size of the same attributes in a group may differ.
- Type of the same attributes in a group may differ.

Let us see with an example how entities can have different set of attributes. The attributes may also have different data types and most certainly can have different sizes. For example, names and emails of different people can be stored in more than one way as shown below:

***One way is:***

Name: Patrick Wood

Email: ptw@dcs.abc.ac.uk, p.wood@ymail.uk

***Another way is:***

First name: Mark

Last name: Taylor

Email: MarkT@dcs.ymail.ac.uk

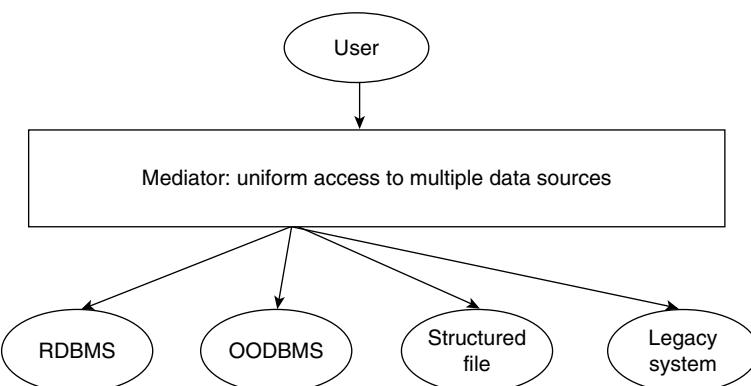
***Yet another way is:***

Name: Alex Bourdoo

Email: AlexBourdoo@dcs.ymail.ac.uk

Integration of data from heterogeneous sources (depicted in Figure 2.18) leads to the data being semi-structured. Because it is likely that data from one source may not have adequate structure while others may have information which is not required or the required information missing from them.

The problems arising because of the nature of semi-structured data are evident in Dr. Brian's failure to deliver good healthcare to his patient. The reason behind his failure is that the information he seeks is in the semi-structured format of a blood test report prepared by Dr. Marianne. This could have been avoided had adequate semi-structured data management been undertaken by GoodLife.



**Figure 2.18** Integration of data from heterogeneous sources. Each source (RDBMS, Object Oriented DBMS, Structured file, Legacy system) represents data differently. They may conform to different data models/different schemas.

## 2.5.2 How to Manage Semi-Structured Data?

Listed below are few ways in which semi-structured data is managed and stored.

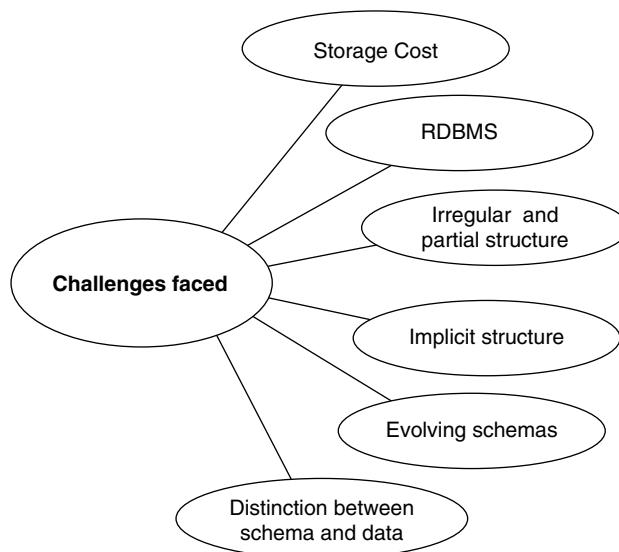
- **Schemas:** These can be used to describe the structure of data. Schemas define the constraints on the structure, content of the document, etc. The problem with schemas is that requirements are ever changing and the changes required in data also lead to changes in schema.
- **Graph-based data models:** These can be used to describe data. This is “schema-less” approach and is also known as “self-describing” as data is presented in such a way that it explains itself. The relationships and hierarchies are represented in the form of a tree-like structure where the vertices contain the object or entity and the leaves contain data.
- **XML:** This is widely used to store and exchange semi-structured data. It allows the user to define tags to store data in hierarchical or nested forms. Schemas in XML are not tightly coupled to data.

This brings us to the next topic – How to store semi-structured data?

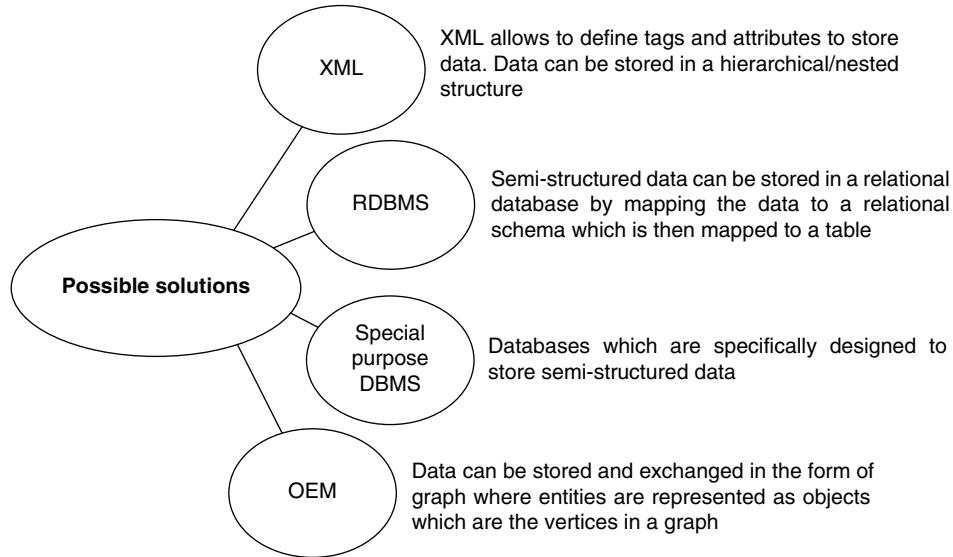
## 2.5.3 How to Store Semi-Structured Data?

Semi-structured data usually has an irregular and partial structure. Data from a few sources may have partial structure while some may have none at all. The structure of data from some sources is implicit which makes it very difficult to interpret relationships between data.

In the case of semi-structured data, schema and data are usually tightly coupled. Same queries may update both schema and data with the schema being updated very frequently. Sometimes the distinction between schema and data is very vague. For example, in some cases the data from source may contain the “status”, i.e. married or single as true or false and consider it as a separate attribute. But in some sources it may be an attribute of a larger set or class. These problems complicate the designing of structure for the data. Figure 2.19 illustrates the challenges faced in storing semi-structured data.



**Figure 2.19** Challenges faced in storing semi-structured data.



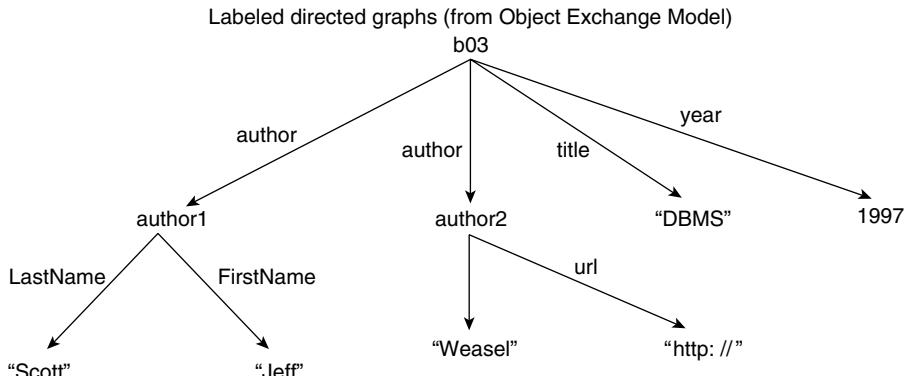
**Figure 2.20** Possible solutions for storing semi-structured data.

The possible solutions to the challenges faced in storing semi-structured data are indicated in Figure 2.20.

#### 2.5.4 Modeling Semi-Structured Data (The OEM Way)

OEM (Object Exchange Model) is a model for storing and exchanging semi-structured data. It structures data in the form of graphs. In OEM, depicted in Figure 2.21, the objects are the entities. The labels are the attributes and the leaf contains the data. It models the hierarchies, nested structures, etc. Indexing and searching a graph-based data model is easier and quicker as it is easy to traverse to the data.

This brings us to the next question – How do we extract information from this data?

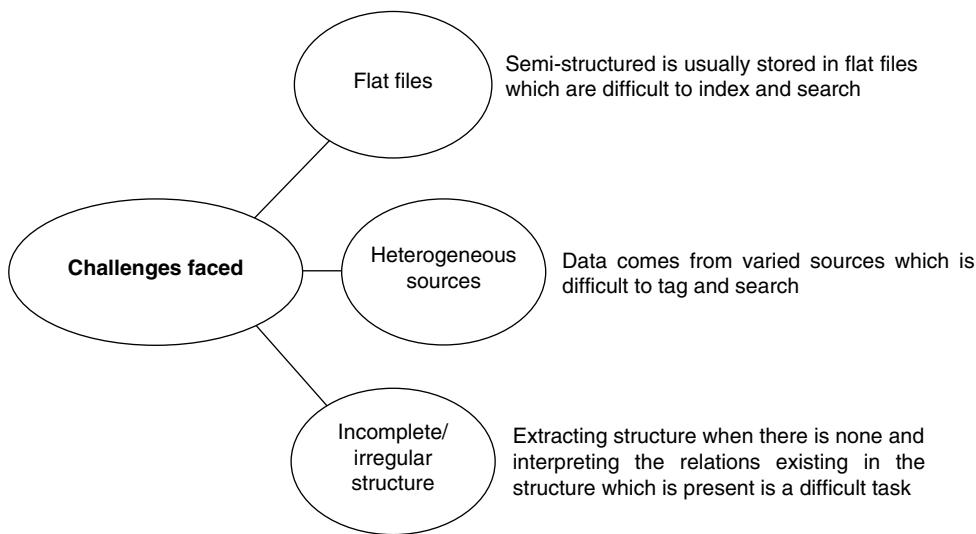


**Figure 2.21** Object Exchange Modeling. Nodes are objects; labels on the arcs are attribute names.

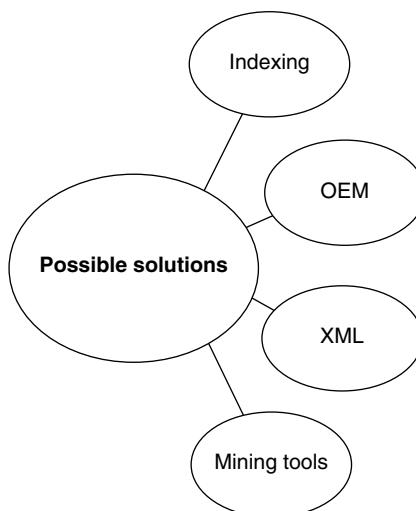
### 2.5.5 How to Extract Information from Semi-Structured Data?

Data coming from heterogeneous sources contain different structures (in some cases none at all!). And, it is difficult to tag and index them. The various challenges faced while extracting information from semi-structured data has been summarized in Figure 2.22.

Now that we have the list of challenges before us, how do we overcome these concerns with semi-structured data? The possible solutions to the challenges are depicted in Figure 2.23 and listed below.



**Figure 2.22** Challenges faced while extracting information from semi-structured data.



**Figure 2.23** Possible solutions for extracting information from semi-structured data.

- **Indexing:** Indexing data in a graph-based model enables quick search.
- **OEM:** This data modeling technique allows for the data to be stored in a graph-based data model which is easier to index and search.
- **XML:** It allows data to be arranged in a hierarchical or tree-like structure which enables indexing and searching.
- **Mining tools:** Various mining tools are available which search data based on graphs, schemas, structures, etc.

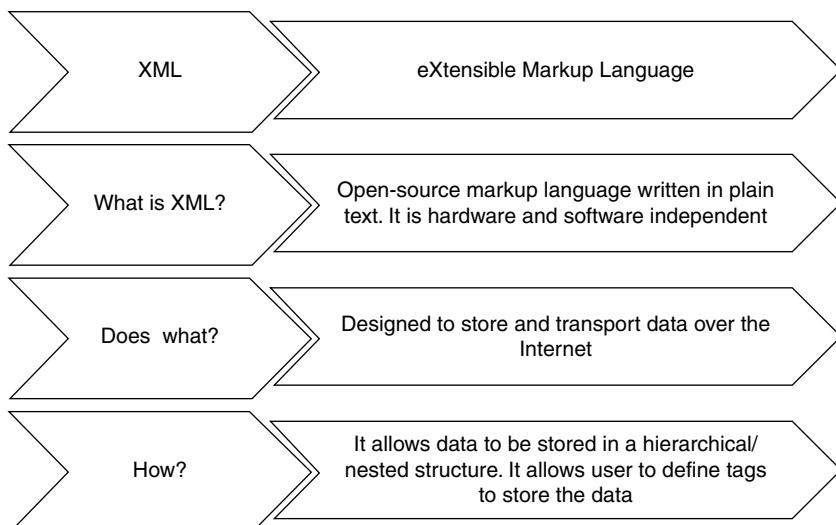
### 2.5.6 XML: A Solution for Semi-Structured Data Management

XML (eXtensible Markup Language) is an open source markup language written in plain text. It is independent of hardware and software. It is designed to store and transport data over the Internet. It allows data to be stored in a hierarchical/nested fashion. In XML, the user can define tags to store data. It also enables separation of content (eXtensible Markup Language) and presentation (eXtensible Stylesheet Language). XML is slowly emerging as a solution for semi-structured data management. Figure 2.24 summarizes the role of XML in semi-structured data management.

XML has no pre-defined tags. The words in the < > (angular brackets) are user-defined tags (Figure 2.25). XML is known as self-describing as data can exist without a schema and schema can be added later. Schema can be described in the XSLT or XML schema.

In brief, the characteristics of XML language are as follows:

- XML (eXtensible Markup Language) is slowly emerging as a standard for exchanging data over the Web.
- It enables separation of content (eXtensible Markup Language) and presentation (eXtensible Stylesheet Language).
- DTD's (Document Type Descriptors) provide partial schemas for XML documents.



**Figure 2.24** XML – A solution for semi-structured data management.

```

<library>
  <book year="2005">
    <title> Database Systems </title>
    <author> <lastname> Date </lastname> </author>
    <publisher> Addison-Wesley </publisher>
  </book>
  <book year="2008">
    <title> Foundation for Object/Relational Databases </title>
    <author> <lastname> Date </lastname> </author>
    <author> <lastname> Darson </lastname> </author>
    <ISBN> <number> 01-23-456 </number> </ISBN>
  </book>
</library>

```

**Figure 2.25** An example of XML.**Table 2.1** Semi-structured data vs. XML

| <i>Semi-Structured Data</i>        | <i>XML</i>                   |
|------------------------------------|------------------------------|
| Consists of attributes             | Consists of tags             |
| Consists of objects                | Consists of elements         |
| Atomic values are the constituents | CDATA (characters) are used. |

The differences between semi-structured data and XML are highlighted in Table 2.1.

## 2.6 DIFFERENCE BETWEEN SEMI-STRUCTURED AND STRUCTURED DATA

Semi-structured data is the same as structured data with one minor exception: semi-structured data requires looking at the data itself to determine structure as opposed to structured data that only requires examining the data element name. Figure 2.26 illustrates the difference between semi-structured and structured data. Semi-structured data is one processing step away from structured data. From a data modeler's point of view, there is no difference between structured and semi-structured data. However, from an analyst's point of view, there is a huge difference because the analyst needs to create the data element source/target mapping, which is traditionally much more complex with semi-structured data.

Consider the following example taken for semi-structured data:

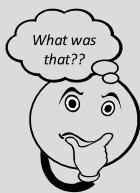
*name:* Patrick Wood  
*email:* ptw@dcs.abc.ac.uk, p.wood@ymail.uk  
*name:*  
*first name:* Mark  
*last name:* Taylor  
*email:* MarkT@dcs.ymail.ac.uk

| Semi-structured                       |                                       | Structured |           |                                      |                    |
|---------------------------------------|---------------------------------------|------------|-----------|--------------------------------------|--------------------|
| Name                                  | Email                                 | First Name | Last Name | Email Id                             | Alternate Email Id |
| Patrick Wood                          | ptw@dcs.abc.ac.uk,<br>p.wood@ymail.uk | Patrick    | Wood      | ptw@dcs.ab<br>c.ac.uk                | p.wood@ymail.uk    |
| first name: Mark<br>last name: Taylor | MarkT@dcs.ymail.ac.uk                 | Mark       | Taylor    | MarkT@dcs.<br>ymail.ac.uk            |                    |
| Alex Bourdoo                          | AlexBourdoo@dcs.ymail<br>.ac.uk       | Alex       | Bourdoo   | AlexBourdo<br>o@dcs.ymail<br>l.ac.uk |                    |

**Figure 2.26** Difference between semi-structured and structured data.

*name:* Alex Bourdoo  
*email:* AlexBourdoo@dcs.ymail.ac.uk

This semi-structured data when stored in the structured format will be in the form of rows and columns each having a defined format as shown in Figure 2.26.



### Remind Me

- *Unstructured Data:* Data which does not conform to a data model or is not in a form which can be used easily by a computer program.
- *Semi-structured Data:* For this data, metadata is available but is not sufficient.
- *Structured Data:* Data which is in an organized form, e.g. in rows and columns or data stored in a database.
- Anything that is in a non-database form is unstructured data.
- Unstructured data can be classified into Bitmap Objects or Textual Objects.
- Web pages are said to be unstructured data even though they are defined by the HTML markup language, which has a rich structure.
- *CAS:* Content Addressable Storage.
- *UIMA:* Unstructured Information Management Architecture.
- UIMA uses analysis engines to analyze the unstructured content to extract implicit information. This information is stored in a structured format.
- *XML (eXtensible Markup Language)* is an open source markup language and is independent of hardware and software.

- XML is known as self-describing because data in XML can exist without a schema.
- Semi-structured data is the same as structured data with one minor exception: semi-

structured data requires looking at the data itself to determine structure as opposed to structured data that only requires examining the data element name.



### *Connect Me (Internet Resources)*

#### **Unstructured Data**

- <http://www.information-management.com/issues/20030201/6287-1.html>
- [http://www.enterpriseitplanet.com/storage/features/article.php/11318\\_3407161\\_2](http://www.enterpriseitplanet.com/storage/features/article.php/11318_3407161_2)
- [http://domino.research.ibm.com/comm/research\\_projects.nsf/pages/uima.index.html](http://domino.research.ibm.com/comm/research_projects.nsf/pages/uima.index.html)
- <http://www.research.ibm.com/UIMA/UIMA%20Architecture%20Highlights.html>

#### **Semi-Structured Data**

- <http://queue.acm.org/detail.cfm?id=1103832>
- [http://www.computerworld.com/s/article/93968/Taming\\_Text](http://www.computerworld.com/s/article/93968/Taming_Text)
- [http://searchstorage.techtarget.com/generic/0,295582,sid5\\_gci1334684,00.html](http://searchstorage.techtarget.com/generic/0,295582,sid5_gci1334684,00.html)
- [http://searchdatamanagement.techtarget.com/generic/0,295582,sid91\\_gci1264550,00.html](http://searchdatamanagement.techtarget.com/generic/0,295582,sid91_gci1264550,00.html)
- [http://searchdatamanagement.techtarget.com/news/article/0,289142,sid91\\_gci1252122,00.html](http://searchdatamanagement.techtarget.com/news/article/0,289142,sid91_gci1252122,00.html)

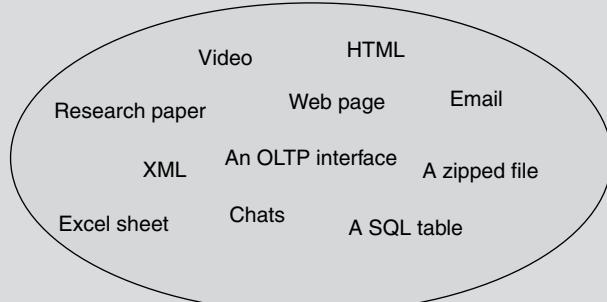
#### **Structured Data**

- <http://www.govtrack.us/articles/20061209data.xpd>
- [http://www.sapdesignguild.org/editions/edition2/sui\\_content.asp](http://www.sapdesignguild.org/editions/edition2/sui_content.asp)



### *Test Me Exercises*

**Classify the given data into three categories: Structured, Semi-Structured, and Unstructured**



**Solution:**

| <i>Unstructured</i> | <i>Semi-Structured</i> | <i>Structured</i> |
|---------------------|------------------------|-------------------|
| Video               | HTML                   | An OLTP interface |
| Chats               | XML                    | A SQL table       |
| Research paper      | Zipped files           |                   |
| Web page            | Email                  |                   |

*Challenge Me*

1. What are the biggest sources of unstructured data in an enterprise?
2. Describe some major challenges associated with unstructured data?
3. What can an enterprise do to tackle the problem of unstructured data?
4. What is one key challenge with semi-structured data?
5. Why semi-structured data?
6. Should structured data stores be confined to contain/hold only structured data?

**Solution:**

1. Email and file services; both generate a lot of data. With email, whenever we “Reply to All” and forward messages, the Exchange Server duplicates and proliferate a message many times over – often with attachments.
2. Volume of data and the continuing tremendous growth of data figure among major challenges of unstructured data. Another challenge is to identify the unstructured data in order to manage it. For example, the only way to identify bitmap images, seismic data, audio or video, is by their filename and extension – there is no way to “look” at the data and know that a given piece of data comprises an image

or other data type. This makes essential management tasks, like data identification, classification, legal discovery and even basic searches, very challenging for an enterprise.

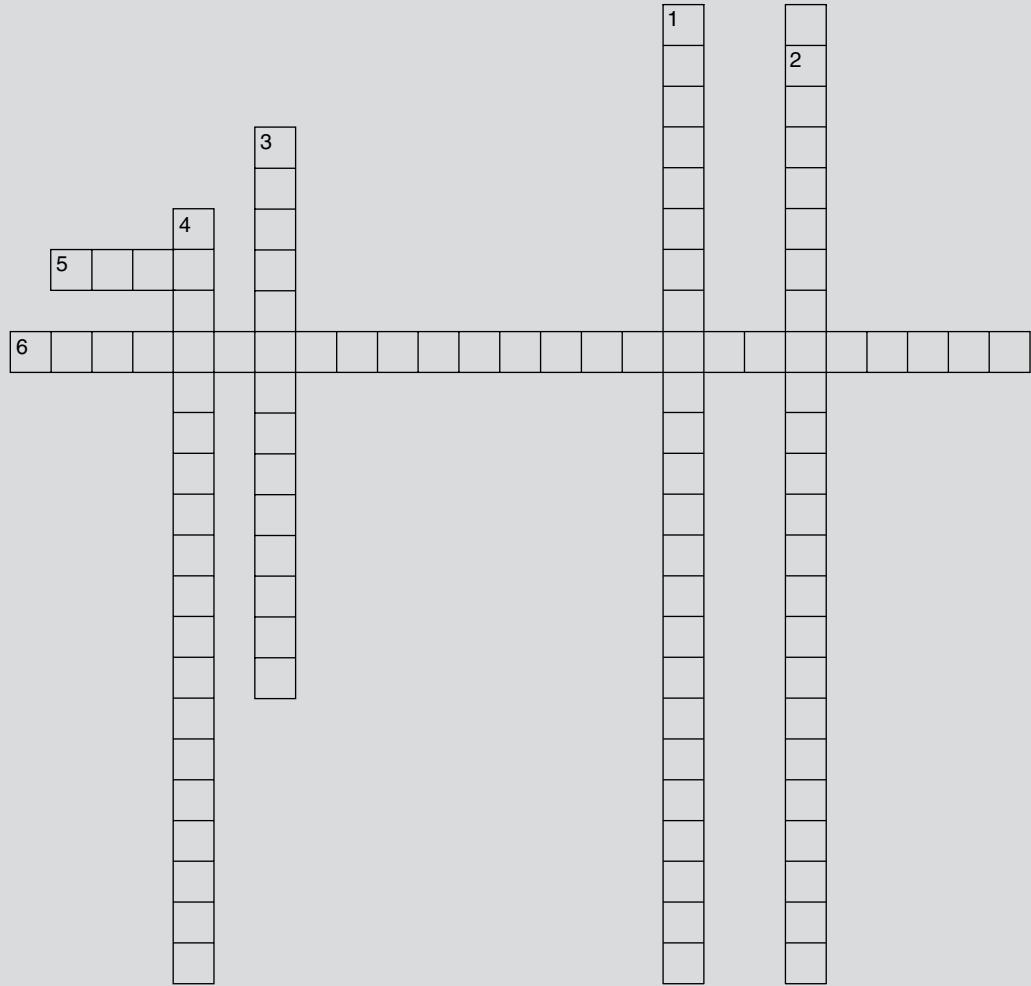
3. Enterprises need to realize the sheer enormity of unstructured data and the knowledge that lies buried in it. They need to come up with policies like naming conventions and standards to be followed across the enterprise. This will bring about some amount of consistency in data. Enterprises need to classify and categorize data according to archival, production based, daily queried data types, etc. and then use technology to manage unstructured data.
4. A schema is one key challenge in the case of semi-structured data. It is not given in advance (often implicit in the data). It is descriptive not prescriptive, partial, rapidly evolving and may be large (compared to the size of the data). Also, types are not what they used to be: objects and attributes are not strongly typed as well as objects in the same collection have different representations.
5. Semi-structured data because:
  - a. raw data is often semi-structured.
  - b. convenient for data integration.
  - c. websites are ultimately graphs.

- d. rapidly evolving schema of the website.
  - e. schema of website does not enforce typing.
  - f. iterative nature of website construction.
6. Enterprises today have 80% of their data in unstructured form and only about 10–20% data is either semi-structured or structured.

Information from all the data formats is required to acquire knowledge. A lot of work is in progress to make storage and retrieval of unstructured, semi-structured, and structured data in a single repository possible.



### *BI Crossword*



**ACROSS**

5. In a relational database, unstructured data may be stored in \_\_\_\_\_. (4)
6. It organizes files based on their metadata and assigns a unique name to every object stored in it. (25)

**DOWN**

1. Enables separation between content and presentation. (24)
2. Provides partial schema for XML document. (23)

3. This data format uses atomic value for its characters. (14)
4. A model for storing semi-structured data in the form of graphs. (19)

**Solution:**

1. eXtensible Markup Language
2. Document Type Descriptors
3. Semi Structured
4. Object Exchange Model
5. BLOB
6. Content Addressable Storage

**UNSOLVED EXERCISES**

1. Compare and contrast structured, semi-structured, and unstructured data.
2. Can semi-structured data be stored as structured data? Explain with the help of an example.
3. Trace the growth of data in any sector (e.g. manufacturing, retail, etc.) or in an area of your interest.
4. Give an example of data, from your life, that has increased to unmanageable levels and that takes considerable amount of your time to manage and store.
5. State an instance where inconsistent data formats caused problems in the integration of data.
6. Can you think of an instance where you came across data that was stored or presented to you in an unstructured, semi-structured, and structured data format?
7. Give an example of data being collected and stored in an unstructured format in your college/university.
8. Suggest a way to add some structure to data that is being collected and stored in an unstructured format in your college/university.
9. Say “Yes” or “No”:
  - a. Is the receipt given to you at the petrol pump in an unstructured form?
  - b. Is the billing of the items purchased at the supermarket done in a structured form?
  - c. Is the form required to be filled for online reservation of railway tickets semi-structured?
  - d. Did your flight ticket have details in a structured format?
  - e. Is the blog entry you made in a structured form?
  - f. Is the discussion you read in the newsroom in a structured form?
  - g. Is the article you read in the e-newspaper in a semi-structured form?
10. Is it important to analyze unstructured data? Explain your answer with an example.
11. How can unstructured data be stored in a relational database?
12. State two ways in which we can extract information from unstructured data.
13. State three ways to store unstructured data.

14. Web pages are said to be unstructured data even though they are defined by the HTML markup language which has a rich structure. Why?
15. What is semi-structured data? List three sources of semi-structured data.
16. How can one manage semi-structured data?
17. Explain OEM (Object Exchange Model) with the help of an example.
18. What is XML? Explain with the help of an example.
19. What are the sources of structured data?
20. How can structured data retrieval be facilitated?
21. What can you say about data in an email? Is it structured, semi-structured, or unstructured? Give reasons in support of your answer.
22. What can you say about data generated in chat conversations? Is it structured, semi-structured, or unstructured? Give reasons in support of your answer.
23. Can XML data be converted into a structured format? Explain with an example.
24. You call up a customer care representative to place a complaint about one of their product offerings. The customer care representative takes down your complaint. What can you say about the format of data in the complaint?
25. Under which category (structured, semi-structured, or unstructured) does a PowerPoint presentation fall?
26. Under which category (structured, semi-structured, or unstructured) does a text file fall?
27. Discuss two best practices for managing the growth of unstructured data.
28. What according to you is the impact of unstructured data on backup and recovery?
29. Under which category (structured, semi-structured, or unstructured) does a census survey form fall?
30. Picture this... A newly opened restaurant wants to collect feedback from its customers on the ambience of the restaurant, the quality and quantity of food served, the hospitality of the restaurant staff, etc. Design an appropriate feedback form for the restaurant and comment on the type (structured, semi-structured, unstructured) of data that will be collected therein.



# 3



## Introduction to OLTP and OLAP

---

### BRIEF CONTENTS

|   |   |
|---|---|
| What's in Store                           | Should OLAP be Performed Directly on Operational Databases? |
| OLTP (On-Line Transaction Processing)     | A Peek into the OLAP Operations on Multidimensional Data    |
| OLAP (On-Line Analytical Processing)      | Leveraging ERP Data Using Analytics                         |
| Different OLAP Architectures              | Solved Exercises  |
| OLTP and OLAP                             | Unsolved Exercises  |
| Data Models for OLTP and OLAP             |   |
| Role of OLAP Tools in the BI Architecture |   |

---

### WHAT'S IN STORE

We assume that you are already familiar with commercial database systems. In this chapter, we will try to build on that knowledge to help you understand and differentiate between OLTP (**O**n-**L**ine **T**ransaction **P**rocessing) systems and OLAP (**O**n-**L**ine **A**nalys**T**ical **P**rocessing) systems. In addition, we will discuss the role of OLAP tools in the BI (data warehousing) architecture and explain why it is not advisable to perform OLAP on operational databases. We will also enhance your knowledge by a brief explanation of ERP and how it is different from OLTP.

We suggest you refer to some of the learning resources suggested at the end of this chapter and also complete the “Test Me” exercises.

---

### 3.1 OLTP (ON-LINE TRANSACTION PROCESSING)

Picture yourself at a point-of-sale (POS) system in a supermarket store. You have picked a bar of chocolate and await your chance in the queue for getting it billed. The cashier scans the chocolate bar’s bar

**Table 3.1** Table structure or schema of the ProductMaster table

| <i>Column Name</i> | <i>Data Type and Length</i> | <i>Constraint</i> | <i>Description</i>                    |
|--------------------|-----------------------------|-------------------|---------------------------------------|
| ProductID          | Character, 7                | Primary Key       | It is not null and unique             |
| ProductName        | Character, 35               | Not Null          | Name of the product must be specified |
| ProductDescription | Character, 50               | Not Null          | A brief description of the product    |
| UnitPrice          | Numeric 8,2                 |                   | The price per unit of the product     |
| QtyInStock         | Numeric 5                   |                   | The units of the product in stock     |

code. Consequent to the scanning of the bar code, some activities take place in the background – the database is accessed; the price and product information is retrieved and displayed on the computer screen; the cashier feeds in the quantity purchased; the application then computes the total, generates the bill, and prints it. You pay the cash and leave. The application has just added a record of your purchase in its database. This was an **On-Line Transaction Processing (OLTP)** system designed to support on-line transactions and query processing. In other words, the POS of the supermarket store was an OLTP system.

OLTP systems refer to a class of systems that manage transaction-oriented applications. These applications are mainly concerned with the entry, storage, and retrieval of data. They are designed to cover most of the day-to-day operations of an organization such as purchasing, inventory, manufacturing, payroll, accounting, etc. OLTP systems are characterized by a large number of short on-line transactions such as INSERT (the above example was a case of insertion wherein a record of final purchase by a customer was added to the database), UPDATE (the price of a product has been raised from \$10 to \$10.5), and DELETE (a product has gone out of demand and therefore the store removes it from the shelf as well as from its database).

Almost all industries today (including airlines, mail-order, supermarkets, banking, insurance, etc.) use OLTP systems to record transactional data. The data captured by OLTP systems is usually stored in commercial relational databases. For example, the database of a supermarket store consists of the following tables to store the data about its transactions, products, employees, inventory supplies, etc.:

- Transactions.
- ProductMaster.
- EmployeeDetails.
- InventorySupplies.
- Suppliers, etc.

Let us look at how data is stored in the ProductMaster table. First, a look at the table structure (also called the schema) depicted in Table 3.1. Then we will have a quick look at a few sample records of the ProductMaster table listed in Table 3.2.

### 3.1.1 Queries that an OLTP System can Process

Let us reconsider our example of the supermarket store POS system, which is an OLTP system. Given below are a set of queries that a typical OLTP system is capable of responding to:

**Table 3.2** A few sample records of the ProductMaster table

| <i>ProductID</i> | <i>ProductName</i> | <i>ProductDescription</i> | <i>UnitPrice</i> | <i>QtyInStock</i> |
|------------------|--------------------|---------------------------|------------------|-------------------|
| P101             | Glucon D           | Energy Drink              | 120.50           | 250               |
| P102             | Boost              | Energy Drink              | 135.00           | 300               |
| P103             | Maxwell DVD        | DVD                       | 45.00            | 500               |
| P104             | Poison             | Perfume                   | 425.00           | 100               |
| P105             | Reynolds           | Pen                       | 15.00            | 125               |
| P106             | Maggie Sauce       | Tomato Sauce              | 54.00            | 250               |

- Search for a particular customer's record.
- Retrieve the product description and unit price of a particular product.
- Filter all products with a unit price equal to or above \$25.
- Filter all products supplied by a particular supplier.
- Search and display the record of a particular supplier.

### 3.1.2 Advantages of an OLTP System

- **Simplicity:** It is designed typically for use by clerks, cashiers, clients, etc.
- **Efficiency:** It allows its users to read, write, and delete data quickly.
- **Fast query processing:** It responds to user actions immediately and also supports transaction processing on demand.

### 3.1.3 Challenges of an OLTP System

- **Security:** An OLTP system requires concurrency control (locking) and recovery mechanisms (logging).
- **OLTP system data content not suitable for decision making:** A typical OLTP system manages the current data within an enterprise/organization. This current data is far too detailed to be easily used for decision making.

### 3.1.4 The Queries that OLTP cannot Answer

Yet again, we go back to our point-of-sale system example. That system helps us perform transactions (such as INSERT, UPDATE, and DELETE) and do simple query processing (locate the record of a particular customer/product/supplier, etc.). How about using this system to handle some complex queries like the ones listed below:

- The supermarket store is deciding on introducing a new product. The key questions they are debating are: "Which product should they introduce?" and "Should it be specific to a few customer segments?"
- The supermarket store is looking at offering some discount on their year-end sale. The questions here are: "How much discount should they offer?" and "Should different discounts be given to different customer segments?"

- The supermarket is looking at rewarding its most consistent salesperson. The question here is: “How to zero in on its most consistent salesperson (consistent on several parameters)?”

All the queries stated above have more to do with analysis than simple reporting. Ideally these queries are not meant to be solved by an OLTP system. Let us look at our next topic – OLAP.

## 3.2 OLAP (ON-LINE ANALYTICAL PROCESSING)

OLAP differs from traditional databases in the way data is conceptualized and stored. In OLAP data is held in the dimensional form rather than the relational form. OLAP's life blood is multi-dimensional data. OLAP tools are based on the multi-dimensional data model. The multi-dimensional data model views data in the form of a data cube. Let us get started with dimensional data.

We will use the data of a supermarket store, “AllGoods” store, for the year “2001” as given in Table 3.3. This data as captured by the OLTP system is under the following column headings: Section, ProductCategoryName, YearQuarter, and SalesAmount. We have a total of 32 records/rows. The Section column can have one value from amongst “Men”, “Women”, “Kid”, and “Infant”. The ProductCategoryName column can have either the value “Accessories” or the value “Clothing”. The YearQuarter column can have one value from amongst “Q1”, “Q2”, “Q3”, and “Q4”. The SalesAmount column records the sales figures for each Section, ProductCategoryName, and YearQuarter.

**Table 3.3** Sample data of “AllGoods” store for the year 2001

| Section | ProductCategoryName | YearQuarter | SalesAmount |
|---------|---------------------|-------------|-------------|
| Men     | Accessories         | Q1          | 3000.50     |
| Men     | Accessories         | Q2          | 1000.50     |
| Men     | Accessories         | Q3          | 3500.50     |
| Men     | Accessories         | Q4          | 2556.50     |
| Women   | Accessories         | Q1          | 1250.50     |
| Women   | Accessories         | Q2          | 1000.50     |
| Women   | Accessories         | Q3          | 1500.50     |
| Women   | Accessories         | Q4          | 1556.50     |
| Kid     | Accessories         | Q1          | 1234.50     |
| Kid     | Accessories         | Q2          | 5678.50     |
| Kid     | Accessories         | Q3          | 1233.50     |
| Kid     | Accessories         | Q4          | 1567.50     |
| Infant  | Accessories         | Q1          | 1555.50     |
| Infant  | Accessories         | Q2          | 2000.50     |

(Continued)

**Table 3.3** (Continued)

| <i>Section</i> | <i>ProductCategoryName</i> | <i>YearQuarter</i> | <i>SalesAmount</i> |
|----------------|----------------------------|--------------------|--------------------|
| Infant         | Accessories                | Q3                 | 3425.50            |
| Infant         | Accessories                | Q4                 | 1775.50            |
| Men            | Clothing                   | Q1                 | 2000.50            |
| Men            | Clothing                   | Q2                 | 1230.50            |
| Men            | Clothing                   | Q3                 | 1456.50            |
| Men            | Clothing                   | Q4                 | 3567.50            |
| Women          | Clothing                   | Q1                 | 4536.50            |
| Women          | Clothing                   | Q2                 | 2345.50            |
| Women          | Clothing                   | Q3                 | 3200.50            |
| Women          | Clothing                   | Q4                 | 1550.50            |
| Kid            | Clothing                   | Q1                 | 1000.50            |
| Kid            | Clothing                   | Q2                 | 6789.50            |
| Kid            | Clothing                   | Q3                 | 8889.50            |
| Kid            | Clothing                   | Q4                 | 7676.50            |
| Infant         | Clothing                   | Q1                 | 2345.50            |
| Infant         | Clothing                   | Q2                 | 2000.50            |
| Infant         | Clothing                   | Q3                 | 3456.50            |
| Infant         | Clothing                   | Q4                 | 5564.50            |

### 3.2.1 One-Dimensional Data

Look at Table 3.4. It displays “AllGoods” store’s sales data by Section, which is one-dimensional data. Although Table 3.4 shows data in two dimensions (horizontal and vertical), in OLAP it is considered to be one dimension as we are looking at the SalesAmount from one particular perspective, i.e. by

**Table 3.4** One-dimensional data by Section

| <i>Section</i> | <i>SalesAmount</i> |
|----------------|--------------------|
| Infant         | 22124.00           |
| Kid            | 34070.00           |
| Men            | 18313.00           |
| Women          | 16941.00           |

Section. We may choose to look at the data from a different perspective, say, by ProductCategoryName.

Table 3.5 presents the sales data of the “AllGoods” stores by ProductCategoryName. This data is again in one dimension as we are looking at the SalesAmount from one particular perspective, i.e. by ProductCategoryName.

**Table 3.5** One-dimensional data by ProductCategoryName

| <i>ProductCategoryName</i> | <i>SalesAmount</i> |
|----------------------------|--------------------|
| Accessories                | 33837.00           |
| Clothing                   | 57611.00           |

One of the most important factors while performing OLAP analysis is time. Table 3.6 presents the “AllGoods” sales data by yet another dimension, i.e. YearQuarter. However, this data is yet another example of one-dimensional data as we are looking at the SalesAmount from one particular perspective, i.e. by YearQuarter.

**Table 3.6** One-dimensional data by YearQuarter

| <i>ProductCategoryName</i> | <i>SalesAmount</i> |
|----------------------------|--------------------|
| Q1                         | 16924.00           |
| Q2                         | 22046.00           |
| Q3                         | 26663.00           |
| Q4                         | 25815.00           |

### 3.2.2 Two-Dimensional Data

So far it has been good. One-dimensional data was easy. What if, the requirement was to view the company’s data by calendar quarters and product categories? Here, two-dimensional data comes into play. Table 3.7 gives you a clear idea of the two-dimensional data. In this table, two dimensions (YearQuarter and ProductCategoryName) have been combined.

**Table 3.7** Two-dimensional data by YearQuarter and ProductCategoryName

| <i>YearQuarter</i> | <i>Accessories</i> | <i>Clothing</i> | <i>SalesAmount</i> |
|--------------------|--------------------|-----------------|--------------------|
| Q1                 | 7041               | 9883            | <b>16924</b>       |
| Q2                 | 9680               | 12366           | <b>22046</b>       |
| Q3                 | 9660               | 17003           | <b>26663</b>       |
| Q4                 | 7456               | 18359           | <b>25815</b>       |
| <b>Total</b>       | <b>33837</b>       | <b>57611</b>    | <b>91448</b>       |

The two-dimensional depiction of data allows one the liberty to think about dimensions as a kind of coordinate system. In Table 3.7, data has been plotted along two dimensions as we can now look at the SalesAmount from two perspectives, i.e. by YearQuarter and ProductCategoryName. The calendar quarters have been listed along the vertical axis and the product categories have been listed across the horizontal axis. Each unique pair of values of these two dimensions corresponds to a single point of SalesAmount data. For example, the Accessories sales for Q2 add up to \$9680.00 whereas the Clothing sales for the same quarter total up to \$12366.00. Their sales figures correspond to a single point of SalesAmount data, i.e. \$22046.

### 3.2.3 Three-Dimensional Data

What if the company's analyst wishes to view the data – all of it – along all the three dimensions (YearQuarter, ProductCategoryName, and Section) and all on the same table at the same time? For this the analyst needs a three-dimensional view of data as arranged in Table 3.8. In this table, one can now look at the data by all the three dimensions/perspectives, i.e. Section, ProductCategoryName, YearQuarter. If the analyst wants to look for the section which recorded maximum Accessories sales in Q2, then by giving a quick glance to Table 3.8, he can conclude that it is the Kid section.

**Table 3.8** Three-dimensional data by Section, ProductCategoryName, and YearQuarter

| ProductCategoryName | YearQuarter | Men          | Women        | Kid          | Infant       | Total        |
|---------------------|-------------|--------------|--------------|--------------|--------------|--------------|
| <b>Accessories</b>  | Q1          | 3000.5       | 1250.5       | 1234.5       | 1555.5       | <b>7041</b>  |
|                     | Q2          | 1000.5       | 1000.5       | 5678.5       | 2000.5       | <b>9680</b>  |
|                     | Q3          | 3500.5       | 1500.5       | 1233.5       | 3425.5       | <b>9660</b>  |
|                     | Q4          | 2556.5       | 1556.5       | 1567.5       | 1775.5       | <b>7456</b>  |
| <b>Clothing</b>     | Q1          | 2000.5       | 4536.5       | 1000.5       | 2345.5       | <b>9883</b>  |
|                     | Q2          | 1230.5       | 2345.5       | 6789.5       | 2000.5       | <b>12366</b> |
|                     | Q3          | 1456.5       | 3200.5       | 8889.5       | 3456.5       | <b>17003</b> |
|                     | Q4          | 3567.5       | 1550.5       | 7676.5       | 5564.5       | <b>18359</b> |
| <b>Total</b>        |             | <b>18313</b> | <b>16941</b> | <b>34070</b> | <b>22124</b> | <b>91448</b> |

### 3.2.4 Should We Go Beyond the Third Dimension?

Well, if the question is “Can you go beyond the third dimension?” the answer is YES! If at all there is any constraint, it is because of the limits of your software. But if the question is “Should you go beyond the third dimension?” we will say it is entirely on what data has been captured by your operational/transactional systems and what kind of queries you wish your OLAP system to respond to.

Now that we understand multi-dimensional data, it is time to look at the functionalities and characteristics of an OLAP system. OLAP systems are characterized by a low volume of transactions that involve very complex queries. Some typical applications of OLAP are: budgeting, sales forecasting, sales reporting, business process management, etc.

**Example:** Assume a financial analyst reports that the sales by the company have gone up. The next question is “Which Section is most responsible for this increase?” The answer to this question is usually followed by a barrage of questions such as “Which store in this Section is most responsible for the increase?” or “Which particular product category or categories registered the maximum increase?” The answers to these are provided by multidimensional analysis or OLAP.

### 3.2.5 Queries that an OLAP System can Process

Let us go back to our example of a company’s (“AllGoods”) sales data viewed along three dimensions: Section, ProductCategoryName, and YearQuarter. Given below are a set of queries, related to our example, that a typical OLAP system is capable of responding to:

- What will be the future sales trend for “Accessories” in the “Kid’s” Section?
- Given the customers buying pattern, will it be profitable to launch product “XYZ” in the “Kid’s” Section?
- What impact will a 5% increase in the price of products have on the customers?

### 3.2.6 Advantages of an OLAP System

- Multidimensional data representation.
- Consistency of information.
- “What if” analysis.
- Provides a single platform for all information and business needs – planning, budgeting, forecasting, reporting, and analysis.
- Fast and interactive ad hoc exploration.

## 3.3 DIFFERENT OLAP ARCHITECTURES

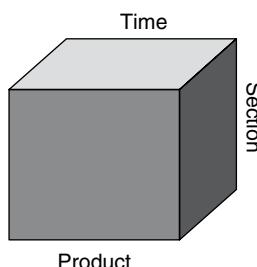
---

Different types of OLAP architecture are:

- Multidimensional OLAP (MOLAP).
- Relational OLAP (ROLAP).
- Hybrid OLAP (HOLAP).

### 3.3.1 MOLAP (Multidimensional On-Line Analytical Processing)

In MOLAP, data is stored in a multidimensional cube (Figure 3.1). The storage is in proprietary formats and not in the relational database.



**Figure 3.1** OLAP cube with Time, Product, and Section dimensions.

### ***Advantages***

- Fast data retrieval.
- Optimal for slicing and dicing.
- Can perform complex calculations. All calculations are pre-generated when the cube is created.

### ***Disadvantages***

- Limited in the amount of data that it can handle. The reason being as all calculations are pre-generated when the cube is created, it is not possible to include a large amount of data in the cube itself. The cube, however, can be derived from a large amount of data.
- Additional investment in human and capital resources may be required as the cube technology is proprietary and might not exist in the enterprise.

### **3.3.2 ROLAP (Relational On-Line Analytical Processing)**

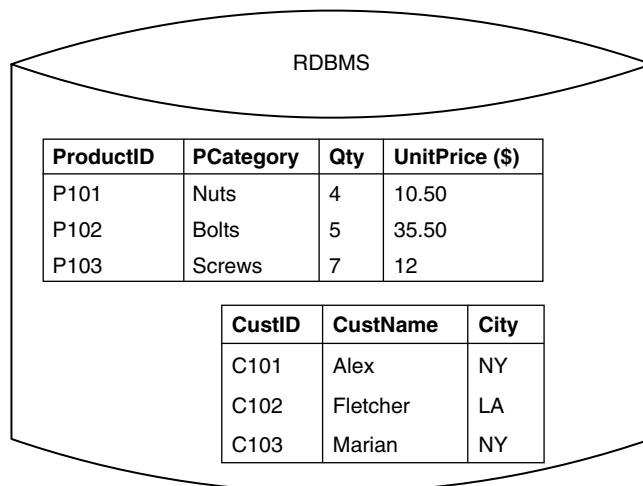
In ROLAP, data is stored in a relational database (Figure 3.2). In essence, each action of slicing and dicing is equivalent to adding a “WHERE” clause in the SQL statement.

### ***Advantages***

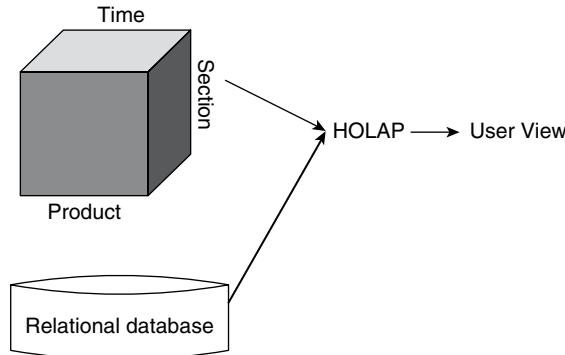
- Can handle large amount of data (limited only by the data size of the underlying database).
- Can leverage functionalities inherent in the relational database.

### ***Disadvantages***

- Difficult to perform complex calculations using SQL.
- Performance can be slow. As each ROLAP report is essentially an SQL query (or multiple SQL queries) in the relational database, the query time can be long if the underlying data size is large.



**Figure 3.2** Data stored in relational database (ROLAP).



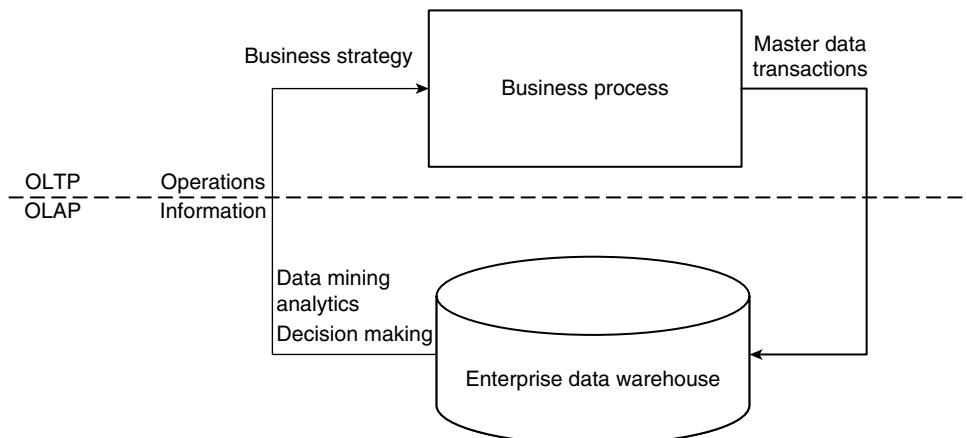
**Figure 3.3** HOLAP.

### 3.3.3 HOLAP (Hybrid On-Line Analytical Processing)

HOLAP technologies attempt to combine the advantages of MOLAP and ROLAP (Figure 3.3). On the one hand, HOLAP leverages the greater scalability of ROLAP. On the other, HOLAP leverages the cube technology for faster performance and for summary-type information. However, HOLAP can also “drill through” into the underlying relational data from the cube.

## 3.4 OLTP AND OLAP

As depicted in Figure 3.4, OLTP helps in the execution of day-to-day operations (in alignment with the business strategy) of an organization/enterprise. It helps keep record of each and every transaction taking place on day-to-day basis. The transaction records are stored in commercial relational database systems. Data from multiple disparate transactional systems is then brought together in an enterprise data warehouse after successful extraction, cleansing (error detection and rectification), and transformation (data converted from legacy or host format to warehouse format). This data is then used for analysis, for



**Figure 3.4** OLTP vs. OLAP.

unravelling hidden patterns and trends, and for decision-making. The decisions so taken further help bring efficiency in the operations of the organization/enterprise. A comparison of the features of OLTP and OLAP has been given in Table 3.9.

**Table 3.9** Comparison of features of OLTP and OLAP

| Feature                    | OLTP<br>(On-Line Transaction Processing)  | OLAP<br>(On-Line Analytical Processing)  |
|----------------------------|---|--|
| <b>Focus</b>               | Data in   | Data out   |
| <b>Source of data</b>      | Operational/Transactional Data  | Data extracted from various operational data sources, transformed and loaded into the data warehouse   |
| <b>Purpose of data</b>     | Manages (controls and executes) basic business tasks  | Assists in planning, budgeting, forecasting, and decision making   |
| <b>Data Contents</b>       | Current data. Far too detailed – not suitable for decision making   | Historical data. Has support for summarization and aggregation. Stores and manages data at various levels of granularity, thereby suitable for decision making |
| <b>Inserts and updates</b> | Very frequent updates and inserts   | Periodic updates to refresh the data warehouse   |
| <b>Queries</b>             | Simple queries, often returning fewer records   | Often complex queries involving aggregations   |
| <b>Processing speed</b>    | Usually returns fast  | Queries usually take a long time (several hours) to execute and return   |
| <b>Space requirements</b>  | Relatively small, particularly when historical data is either purged or archived  | Comparatively huge because of the existence of aggregation structures and historical data  |
| <b>Database design</b>     | Typically normalized tables. OLTP system adopts ER (Entity Relationship) model  | Typically de-normalized tables; uses Star or Snowflake schema  |
| <b>Access</b>              | Field level access  | Typically aggregated access to data of business interest   |
| <b>Operations</b>          | Read/write  | Mostly read  |
| <b>Backup and recovery</b> | Regular backups of operational data are mandatory. Requires concurrency control (locking) and recovery mechanisms (logging) | Instead of regular backups, data warehouse is refreshed periodically using data from operational data sources  |
| <b>Indexes</b>             | Few   | Many   |
| <b>Joins</b>               | Many  | Few  |

(Continued)

**Table 3.9** (Continued)

| <i>Feature</i>                     | <i>OLTP<br/>(On-Line Transaction Processing)</i>   | <i>OLAP<br/>(On-Line Analytical Processing)</i>  |
|------------------------------------|--|--|
| <b>Derived data and aggregates</b> | Rare   | Common   |
| <b>Data structures</b>             | Complex  | Multidimensional   |
| <b>Few sample queries</b>          | <ul style="list-style-type: none"> <li>• Search &amp; locate student(s) record(s)</li> <li>• Print students scores</li> <li>• Filter records where student(s) have scored above 90% marks</li> </ul> | <ul style="list-style-type: none"> <li>• Which courses have productivity impact on-the-job?</li> <li>• How much training is needed on future technologies for non-linear growth in BI?</li> <li>• Why consider investing in DSS experience lab?</li> </ul> |

## 3.5 DATA MODELS FOR OLTP AND OLAP

An OLTP system usually adopts an Entity Relationship (ER) model whereas an OLAP system adopts either a Star or a Snowflake model. We assume that you are familiar with the ER model. The Star and Snowflake models will be covered subsequently in Chapter 7, “Multidimensional Data Modeling”. For now, we will leave you with the ER design and brief introduction of the Star and Snowflake models.

### 3.5.1 Data Model for OLTP

Figure 3.5 depicts an Entity Relationship (ER) data model for OLTP. In this model, we have considered the following three entities:

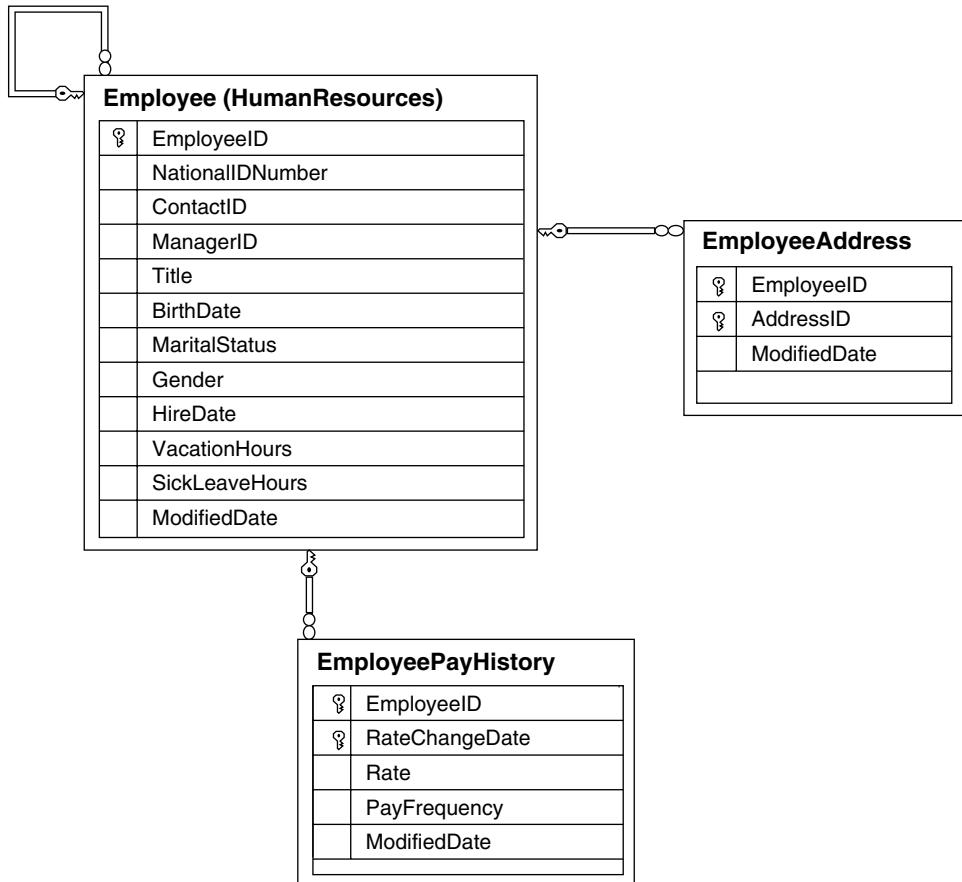
1. Employee (EmployeeID is the primary key).
2. EmployeeAddress (EmployeeID is a foreign key referencing to the EmployeeID attribute of Employee entity).
3. EmployeePayHistory (EmployeeID is a foreign key referencing to the EmployeeID attribute of Employee entity).

We see the following two relationships:

- There is a (**1: M cardinality**) between Employee and EmployeeAddress entities. This means that an instance of Employee entity can be related with multiple instances of EmployeeAddress entity.
- There is also a (**1: M cardinality**) between Employee and EmployeePayHistory entities. This means that an instance of Employee entity can be related with multiple instances of EmployeePayHistory entity.

### 3.5.2 Data Model for OLAP

A multidimensional model can exist in the form of a Star schema, a Snowflake schema, or a Fact Constellation/Galaxy schema. We consider here two models – Star and Snowflake. As already stated, detailed explanation about these models will be given in Chapter 7. Here a brief explanation is provided for easy comprehension.



**Figure 3.5** ER data model for OLTP.

Just before we explain the Star and the Snowflake models, we need to understand two terms – fact and dimensions. In general, a dimension is a perspective or entity with respect to which an organization wants to keep records. For example, “AllGoods” store wants to keep records of the store’s sale with respect to “time”, “product”, “customer”, “employee”. These dimensions allow the store to keep track of things such as the quarterly sales of products, the customers to whom the products were sold, and the employees of the store who were involved in materializing the sales. Each of these dimensions may have a table associated with it, called the dimension table. A dimension table such as “Product” may have attributes like “ProductName”, “ProductCategory”, and “UnitPrice”. Now let us look at what are facts. Facts are numerical measures/quantities by which we want to analyze relationships between dimensions. Examples of facts are “Total (sales amount in dollars)”, “Quantity (number of units sold)”, “Discount (amount in dollars of discount offered)”, etc.

### **Star Model**

Figure 3.6 depicts the Star data model for OLAP. This model has a central fact table which is connected to four dimensions. Each dimension is represented by only one table and each table has a set of attributes.

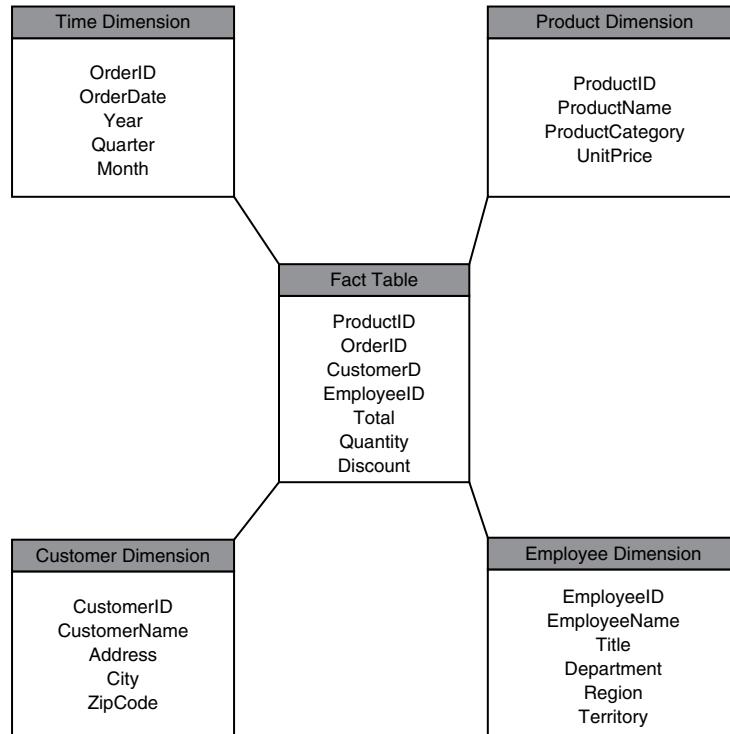


Figure 3.6 Star data model for OLAP.

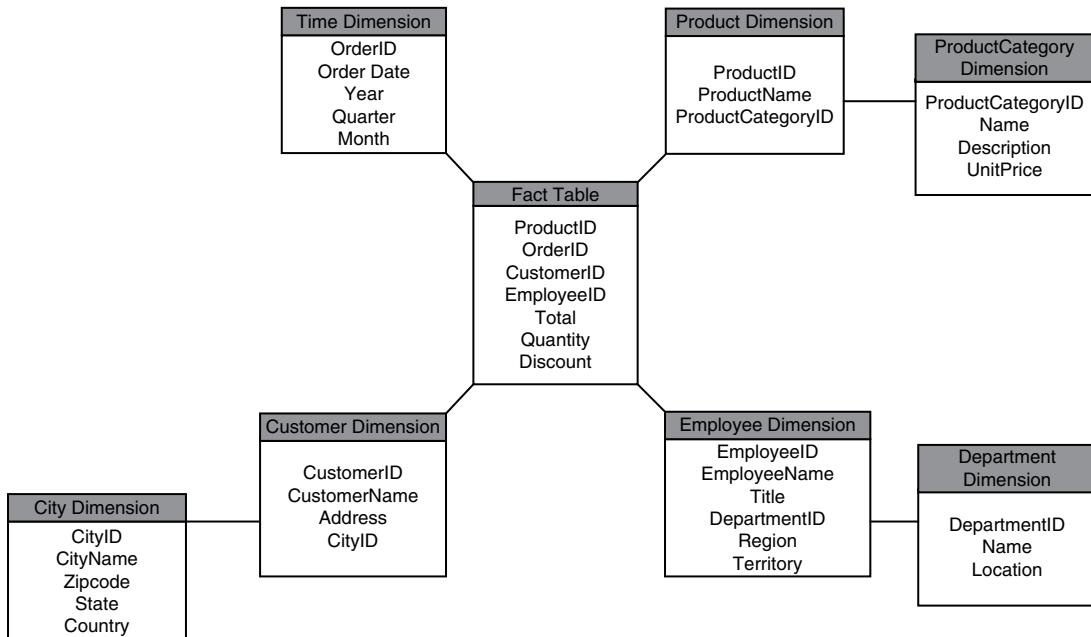


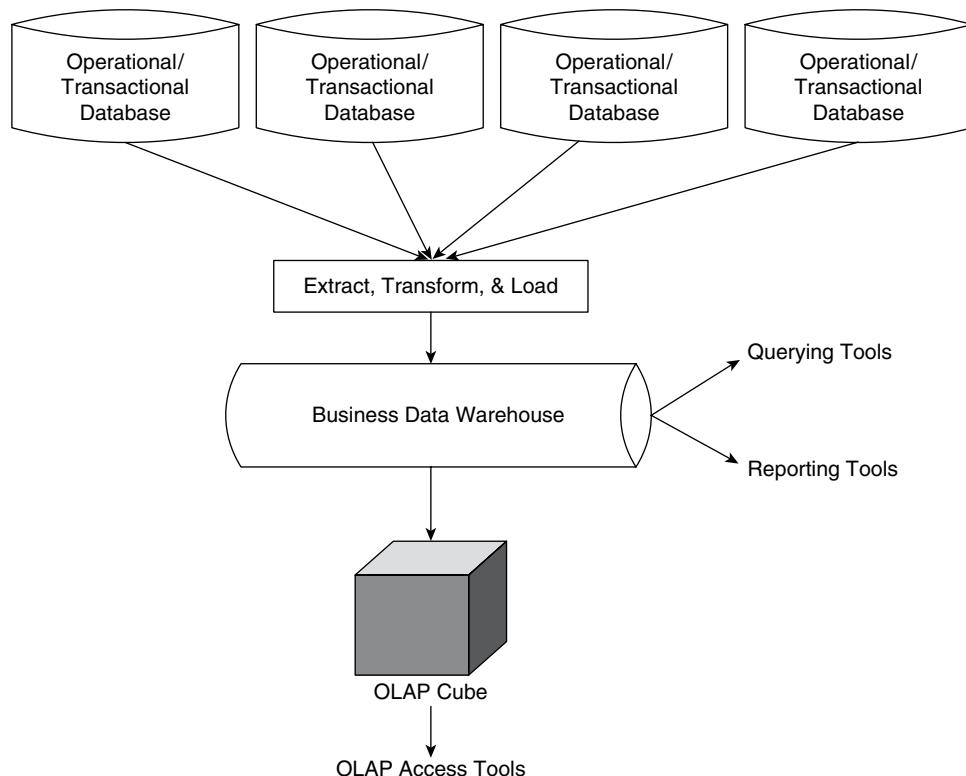
Figure 3.7 Snowflake data model for OLAP.

### **Snowflake Model**

In the Snowflake schema of Figure 3.7, there is a central fact table connected to four dimensions. The “Product” dimension is further normalized to “ProductCategory” dimension. Similarly, the “Employee” dimension is further normalized to the “Department” dimension. By now, you would have guessed that normalization of the dimension tables definitely helps in reducing redundancy; however, it adversely impacts the performance as more joins will be needed to execute a query.

## **3.6 ROLE OF OLAP TOOLS IN THE BI ARCHITECTURE**

The Business Intelligence (BI) architecture of a typical enterprise has a multitude of applications. Most of these applications execute in silos. More often, these applications will have their own backend databases. Refer to Figure 3.8. Data has to be brought/extracted from these multifarious database systems scattered around the enterprise, cleansed (error detection and rectification), transformed (conversion of data from a legacy or host format to the data warehouse format), and loaded into a common business data warehouse. The OLAP system then extracts information from the data warehouse and stores it in a multidimensional hierarchical database using either a relational OLAP (ROLAP) or multidimensional OLAP (MOLAP). Once the data is safely housed in the cube, users can use OLAP query and reporting



**Figure 3.8** OLAP in BI.

tools, analysis tools, and/or data mining tools (e.g. trend analysis, unearthing hidden patterns, alerts, predictions, etc.).

An OLAP system with its adequate tools can help produce Roll-up reports (e.g. if you are viewing the quarterly sales data of your company, you may want to go one level up in the hierarchy and view the annual sales data), Drill-down reports (e.g. if you are viewing the quarterly sales data of your company, you may want to go one level down in the hierarchy and view the sales data for months in a particular quarter), Drill-through reports (e.g. sometimes it may be required to trace back to the data in the operational relational database system), aggregations, summaries, pivot tables, etc. – all focused on varied views/perspectives on the data.

### **3.7 SHOULD OLAP BE PERFORMED DIRECTLY ON OPERATIONAL DATABASES?**

---

Once again, if the question is “Can you perform OLAP directly on operational databases?” The answer is “Yes!” But if the question is “Should you perform OLAP directly on operational databases?” The answer is “No!” Why? This is to ensure the high performance of both the systems. Both the systems (OLTP and OLAP) are designed for different functionalities. We recommend separate databases for OLTP and OLAP systems. OLTP queries can run on operational databases while the OLAP systems will need a separate data warehouse to be built for them. (Building a data warehouse will be explained in detail in subsequent chapters. However, for now, you should know that a data warehouse houses the data extracted from multiple data sources after cleansing and transformation.)

OLTP is to help with the running of the day-to-day operations of an enterprise/organization. It is designed to search for particular records using indexing, etc. On the other hand, OLAP systems deal with the computations of large groups of data at the summarized level. If you perform OLAP queries on operational databases, it will severely degrade the performance of operational tasks.

OLTP systems support multiple concurrent transactions. Therefore OLTP systems have support for concurrency control (locking) and recovery mechanisms (logging). An OLAP system, on the other hand, requires mostly a read-only access to data records for summarization and aggregation. If concurrency control and recovery mechanisms are applied for such OLAP operations, it will severely impact the throughput of an OLAP system.

### **3.8 A PEEK INTO THE OLAP OPERATIONS ON MULTIDIMENSIONAL DATA**

---

There are a number of OLAP data cube operations available that allow interactive querying and analysis of data. Some common OLAP operations on multidimensional data are:

- Slice.
- Dice.
- Roll-up or Drill-up.
- Drill-down.

- Pivot.
- Drill-across.
- Drill-through.

Here is a brief explanation of the above-stated operations.

### 3.8.1 Slice

Slicing is filtering/selecting the data using one dimension of the cube. In Figure 3.9, data is sliced/filtered along the Section dimension using the criterion Section = “Infant” or Section = “Kid”.

| Sum of SalesAmount | ProductCategoryName ▾ |              |              |
|--------------------|-----------------------|--------------|--------------|
| Section ▾          | Accessories           | Clothing     | Grand Total  |
| Infant             | 8757                  | 13367        | 22124        |
| Kid                | 9714                  | 24356        | 34070        |
| <b>Grand Total</b> | <b>18471</b>          | <b>37723</b> | <b>56194</b> |

**Figure 3.9** Slicing the data on the Section dimension.

### 3.8.2 Dice

Dicing is also about filtering the data but using two or more dimensions. In Figure 3.10, the data is sliced/filtered along three dimensions – Section, ProductCategoryName, and YearQuarter. The selection criteria used for these dimensions are: (YearQuarter = “Q3” or YearQuarter = “Q4”), (ProductCategoryName = “Clothing”), and (Section = “Infant” or Section = “Kid”).

| Sum of SalesAmount | Section ▾             |               |               |              |
|--------------------|-----------------------|---------------|---------------|--------------|
| YearQuarter ▾      | ProductCategoryName ▾ | Infant        | Kid           | Grand Total  |
| Q3                 | Clothing              | 3456.5        | 8889.5        | 12346        |
| <b>Q3 Total</b>    |                       | <b>3456.5</b> | <b>8889.5</b> | <b>12346</b> |
| Q4                 | Clothing              | 5564.5        | 7676.5        | 13241        |
| <b>Q4 Total</b>    |                       | <b>5564.5</b> | <b>7676.5</b> | <b>13241</b> |
| <b>Grand Total</b> |                       | <b>9021</b>   | <b>16566</b>  | <b>25587</b> |

**Figure 3.10** Dicing the data on the YearQuarter, ProductCategoryName, and Section dimensions.

### 3.8.3 Roll-Up

The Roll-up operation is also called as Drill-up operation. In roll-up the data is viewed at a higher level of hierarchy of a dimension. For example, let us consider the YearQuarter dimension. The hierarchy in the “AllGoods” stores data (Table 3.3) is Year > YearQuarter, i.e. Year is at a higher level than

YearQuarters. A year has four quarters, i.e. quarter 1 (Q1), quarter 2 (Q2), quarter 3 (Q3), and quarter 4 (Q4). We can view the data by quarters (at a lower level of hierarchy). If we so desire we can also drill-up or roll-up and view the data by year (at a higher level of hierarchy).

Table 3.10 shows the “AllGoods” stores data along the YearQuarter (Q1, Q2, Q3, and Q4) and ProductCategoryName dimensions. The data of Table 3.10 is shown or rolled-up in Table 3.11 at a higher level of hierarchy of the YearQuarter dimension.

**Table 3.10** Data of “AllGoods” stores viewed along the YearQuarter and ProductCategoryName dimensions

| YearQuarter  | Accessories  | Clothing     | SalesAmount  |
|--------------|--------------|--------------|--------------|
| Q1           | 7041         | 9883         | <b>16924</b> |
| Q2           | 9680         | 12366        | <b>22046</b> |
| Q3           | 9660         | 17003        | <b>26663</b> |
| Q4           | 7456         | 18359        | <b>25815</b> |
| <b>Total</b> | <b>33837</b> | <b>57611</b> | <b>91448</b> |

**Table 3.11** Sales data of “AllGoods” stores viewed along the ProductCategoryName dimension for the Year 2001 (inclusive of all quarters: Q1, Q2, Q3, and Q4)

| ProductCategoryName | SalesAmount |
|---------------------|-------------|
| Accessories         | 33837.00    |
| Clothing            | 57611.00    |

### 3.8.4 Drill-Down

In this case, the data is viewed at a lower level of hierarchy of a dimension. In other words, drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. In Figure 3.11, the data is seen at a lower level of hierarchy of the YearQuarter dimension. Here, the total is also viewed by YearQuarter.

### 3.8.5 Pivot

Pivot is also called Rotate. In order to provide an alternative representation of the data, the pivot operation rotates the data axes in view. Let us again observe Figure 3.11. Here the data is displayed along the YearQuarter and ProductCategoryName axes. In the pivot table (Figure 3.12), the ProductCategoryName and YearQuarter axes are rotated.

| Sum of SalesAmount |         | ProductCategoryName |              |              |  |
|--------------------|---------|---------------------|--------------|--------------|--|
| YearQuarter        | Section | Accessories         | Clothing     | Grand Total  |  |
| Q1                 | Infant  | 1555.5              | 2345.5       | 3901         |  |
|                    | Kid     | 1234.5              | 1000.5       | 2235         |  |
|                    | Men     | 3000.5              | 2000.5       | 5001         |  |
|                    | Women   | 1250.5              | 4536.5       | 5787         |  |
| <b>Q1 Total</b>    |         | <b>7041</b>         | <b>9883</b>  | <b>16924</b> |  |
| Q2                 | Infant  | 2000.5              | 2000.5       | 4001         |  |
|                    | Kid     | 5678.5              | 6789.5       | 12468        |  |
|                    | Men     | 1000.5              | 1230.5       | 2231         |  |
|                    | Women   | 1000.5              | 2345.5       | 3346         |  |
| <b>Q2 Total</b>    |         | <b>9680</b>         | <b>12366</b> | <b>22046</b> |  |
| Q3                 |         | 9660                | 17003        | 26663        |  |
| Q4                 |         | 7456                | 18359        | 25815        |  |
| <b>Grand Total</b> |         | <b>33837</b>        | <b>57611</b> | <b>91448</b> |  |

**Figure 3.11** Drilling down the data on the YearQuarter dimension.

| Sum of SalesAmount  | YearQuarter | Section     |             |             | Q1 Total     | Q2           | Q3           | Q4           | Grand Total  |
|---------------------|-------------|-------------|-------------|-------------|--------------|--------------|--------------|--------------|--------------|
| ProductCategoryName | Infant      | Kid         | Men         | Women       |              |              |              |              |              |
| Accessories         | 1555.5      | 1234.5      | 3000.5      | 1250.5      | 7041         | 9680         | 9660         | 7456         | 33837        |
| Clothing            | 2345.5      | 1000.5      | 2000.5      | 4536.5      | 9883         | 12366        | 17003        | 18359        | 57611        |
| <b>Grand Total</b>  | <b>3901</b> | <b>2235</b> | <b>5001</b> | <b>5787</b> | <b>16924</b> | <b>22046</b> | <b>26663</b> | <b>25815</b> | <b>91448</b> |

**Figure 3.12** Pivot table.

### 3.8.6 Drill-Across

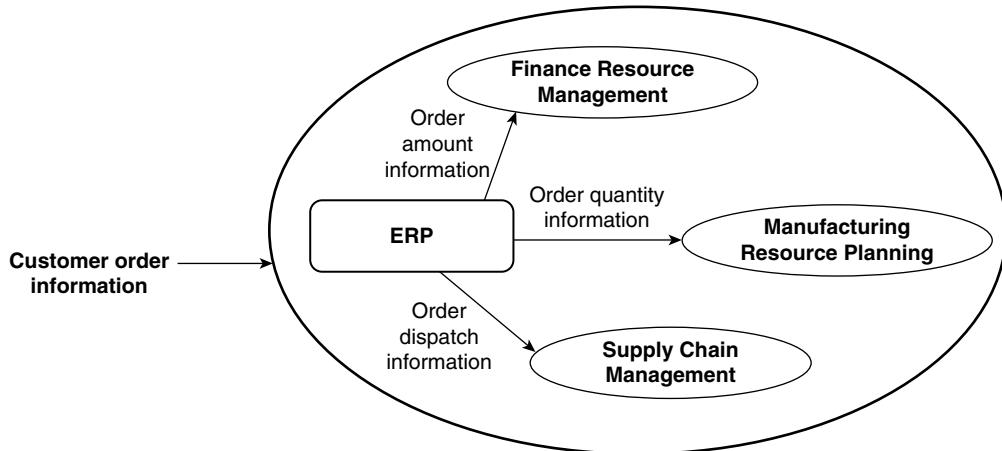
In drill-across, the data is viewed across two or more fact tables.

### 3.8.7 Drill-Through

In drill-through, the data is traced back to the operational data source usually commercial databases.

## 3.9 LEVERAGING ERP DATA USING ANALYTICS

A typical enterprise usually has a chaotic mix of business applications; for example, an application for managing the inventory, another for managing the purchases, yet another for managing the manufacturing, and so on. While a few of them exist in silos, a few others are tied together with the help of complex interface programs. Owing to some applications' existence in silos, there is a huge probability that the same data may exist at more than one place, raising the question on the accuracy and consistency of data. Here is a need for a system that could integrate and automate the business processes from end to end, for example from planning to manufacturing to sales. Enter the ERP software.



**Figure 3.13** Customer order processing in an ERP system.

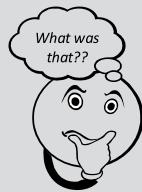
Let us start by understanding what is an ERP system. ERP (Enterprise Resource Planning) systems typically automate transaction processes. Here is an example how an ERP system processes a customer's order through various transaction processes. As depicted in Figure 3.13, a customer places an order. The finance department checks the credit line for the customer to either accept the customer's order or reject it depending on whether the customer has outstanding dues against him. The order, if accepted, is passed to the sales department. The sales department checks to find out whether the ordered goods are in stock or, whether the production/manufacturing needs to be informed to fulfil the order. If the goods are not in stock, the production/manufacturing comes back with a date on which they will be able to ship the order. This way the transaction processes occur to fulfil a customer's order in an ERP system.

Though ERP provides several business benefits, here we enumerate the top three:

- Consistency and reliability of data across the various units of the organization.
- Streamlining the transactional process.
- A few basic reports to serve the operational (day-to-day) needs.

An ERP system is able to solve some, if not all, information needs of the organization. It is also able to successfully integrate several business processes across the organization's supply chain. It is adept at capturing, storing, and moving the data across the various units smoothly. However, an ERP system is inept at serving the analytical and reporting needs of the organization. Why? Because it has limitations in terms of integrating data from existing applications and data from external sources. This integration of the organization's internal transactional data with data from existing applications and external data is imperative if it is to serve the analytical and reporting needs of the organization. So, as a solution to the ERP system shortcomings emerged BI (Business Intelligence) that can effect the desired data integration and do much more.

Our next chapter is on **Getting Started with Business Intelligence**.



## Remind Me

- OLTP – Simple queries, often returning fewer records.
- OLAP – Often complex queries involving aggregations.
- OLTP – Operational/ Transactional Data.
- OLAP – Data extracted from various operational data sources, transformed and loaded into the data warehouse.
- Examples of OLTP – CRM, SCM, ERP, etc.
- Examples of OLAP – Data mining, text mining, Web mining.
- ER (Entity Relationship) is the data model for OLTP systems whereas Star and Snowflake are the data models for OLAP systems.

- Some OLAP operations on multidimensional data are: Slice, Dice, Roll-up, Drill-down, Drill-through, Drill-across, Pivot/rotate, etc.
- OLAP queries on operational databases severely degrade the performance of operational tasks.
- ERP (Enterprise Resource Planning) typically automates the transaction processes of the organization.
- ERP systems are adept at capturing, storing and moving the data across the various units smoothly. They are, however, inept at serving the analytical and reporting needs of the organization.



## Point Me (Books)

- *Business Intelligence for Dummies* by Swain Scheps.
- *Data Mining – Concepts and Techniques* by Jiawei Han and Micheline Kamber.



## Connect Me (Internet Resources)

### **OLAP**

[http://www.dwreview.com/OLAP/Introduction\\_OLAP.html](http://www.dwreview.com/OLAP/Introduction_OLAP.html)

### **Data warehousing**

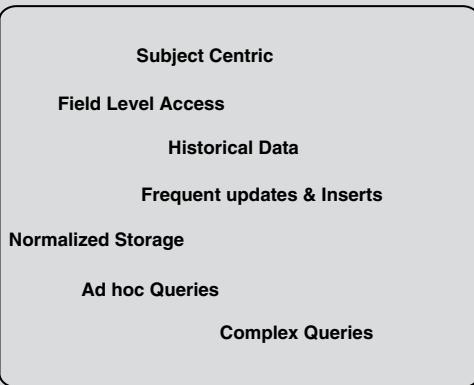
[http://www.dwreview.com/DW\\_Overview.html](http://www.dwreview.com/DW_Overview.html)

<http://www.dwreview.com/Articles/index.html>



## *Test Me Exercises*

Arrange the following features into their respective categories: OLTP or OLAP



### **Solution:**

| OLAP               | OLTP                            |
|--------------------|---------------------------------|
| 1. Complex Queries | 1. Field Level Access           |
| 2. Ad Hoc Queries  | 2. Frequent Updates and Inserts |
| 3. Historical Data | 3. Normalized Storage           |
| 4. Subject Centric |                                 |



## *Solve Me*

1. List a few examples of OLTP transactions taking place in your neighborhood.
2. Give an example where you think that data being collected from an OLTP system might/should be warehoused.
3. List what you think an OLTP system might be recording for every transaction that you make at a supermarket store.
4. An XYZ airline maintains a data warehouse storing all its OLTP transactions with its customers along with their preferences based on their feedback. Think of ways in which this data can help the airline.
5. Which data is being warehoused in your workplace/organization and why?
6. Think of how performing data mining on the data stored in your workplace might be of help to your organization.
7. Suppose X keeps all his bills and payment slips of the various transactions he makes in a month. Can this data be of any use to X? Why and how can it be used?
8. List the various types of data that you have archived or stored in the past.

9. Can you derive some meaning and information from the archived data which will help you in any way in the future?
10. List the services or sectors where you think data should be warehoused and data mining could be used to make predictions.

**Who am I?**

**The following are a set of six clues to help you guess me.**

1. Mostly read-only operations ✓
2. Current data X
3. Focuses on information out ✓
4. Database design – ER-based X
5. Summarizations and aggregations ✓
6. Complex query ✓

**Solution:**

OLAP systems



*Challenge Me*

### Scenario-Based Question

1. Picture an enterprise “XYZ” which is witnessing a mammoth growth in the volume of data. The enterprise has recognized that the data it possesses is its most valuable resource. The enterprise had a few OLTP applications running in silos. Just last year, the enterprise purchased an ERP system. The ERP system has solved some, if not all, information needs of the enterprise. It is able to successfully integrate several business processes across enterprise’s supply chain. It is adept at capturing, storing, and moving the data across various units smoothly. It is, however, inept at serving the analytical and reporting needs of the enterprise.

Why is ERP unable to serve the analytical and reporting needs of the enterprise “XYZ”?

**Solution:** ERP is unable to solve the analytical and reporting needs of the enterprise because it fails to integrate the data from some of the existing applications and also the data from some external sources.

### Individual Activity

2. Given the pivot table in Figure 3.12, rotate the data such that you have the ProductCategoryName and Section along the vertical axes and YearQuarter along the horizontal axes.

**Solution:** See Table 3.12.

**Table 3.12** Individual Activity

| Sum of SalesAmount       | ProductCategoryName | Section | YearQuarter | Q1          | Q2          | Q3          | Q4          | Grand Total  |
|--------------------------|---------------------|---------|-------------|-------------|-------------|-------------|-------------|--------------|
| Accessories              | Infant              |         |             | 1555.5      | 2000.5      | 3425.5      | 1775.5      | 8757         |
|                          | Kid                 |         |             | 1234.5      | 5678.5      | 1233.5      | 1567.5      | 9714         |
|                          | Men                 |         |             | 3000.5      | 1000.5      | 3500.5      | 2556.5      | 10058        |
|                          | Women               |         |             | 1250.5      | 1000.5      | 1500.5      | 1556.5      | 5308         |
| <b>Accessories Total</b> |                     |         |             | <b>7041</b> | <b>9680</b> | <b>9660</b> | <b>7456</b> | <b>33837</b> |
| Clothing                 | Infant              |         |             | 2345.5      | 2000.5      | 3456.5      | 5564.5      | 13367        |

*(Continued)*

**Table 3.12** (Continued)

|                       |       |              |              |              |              |              |
|-----------------------|-------|--------------|--------------|--------------|--------------|--------------|
|                       | Kid   | 1000.5       | 6789.5       | 8889.5       | 7676.5       | 24356        |
|                       | Men   | 2000.5       | 1230.5       | 1456.5       | 3567.5       | 8255         |
|                       | Women | 4536.5       | 2345.5       | 3200.5       | 1550.5       | 11633        |
| <b>Clothing Total</b> |       | <b>9883</b>  | <b>12366</b> | <b>17003</b> | <b>18359</b> | <b>57611</b> |
| <b>Grand Total</b>    |       | <b>16924</b> | <b>22046</b> | <b>26663</b> | <b>25815</b> | <b>91448</b> |

**Team Activity**

3. Pair up with a colleague and analyze the sample data in Table 3.13.

**Table 3.13** A sample data set

| OrderID | VehicleID | SourceCity | DestinationCity | YearQuarter | AmountCharged |
|---------|-----------|------------|-----------------|-------------|---------------|
| 101     | V111      | Mysore     | Bangalore       | Q1          | 1500          |
| 102     | V211      | Bangalore  | Mysore          | Q1          | 1500          |
| 103     | V311      | Bangalore  | Mangalore       | Q1          | 2500          |
| 104     | V411      | Mangalore  | Bangalore       | Q1          | 2575          |
| 105     | V511      | Mysore     | Mangalore       | Q1          | 3200          |
| 106     | V611      | Mysore     | Bangalore       | Q2          | 1500          |
| 107     | V711      | Mysore     | Bangalore       | Q2          | 1500          |
| 108     | V811      | Mysore     | Bangalore       | Q2          | 1500          |
| 109     | V911      | Mysore     | Bangalore       | Q3          | 1500          |
| 110     | V111      | Mysore     | Bangalore       | Q3          | 1500          |
| 111     | V211      | Bangalore  | Mysore          | Q4          | 1500          |
| 112     | V311      | Bangalore  | Mysore          | Q4          | 1500          |
| 113     | V411      | Bangalore  | Mysore          | Q4          | 1500          |
| 114     | V511      | Bangalore  | Mysore          | Q4          | 1500          |
| 115     | V611      | Bangalore  | Mysore          | Q4          | 1500          |
| 116     | V711      | Mangalore  | Bangalore       | Q4          | 2800          |
| 117     | V811      | Mangalore  | Bangalore       | Q1          | 2800          |
| 118     | V911      | Mangalore  | Bangalore       | Q1          | 2800          |
| 119     | V111      | Mangalore  | Bangalore       | Q1          | 2800          |
| 120     | V211      | Mangalore  | Bangalore       | Q1          | 2800          |
| 121     | V311      | Mysore     | Bangalore       | Q1          | 1600          |
| 122     | V411      | Mysore     | Bangalore       | Q2          | 1500          |
| 123     | V511      | Mysore     | Bangalore       | Q2          | 1500          |
| 124     | V611      | Mysore     | Bangalore       | Q2          | 1600          |
| 125     | V711      | Mumbai     | Pune            | Q3          | 2000          |

(Continued)

**Table 3.13** (Continued)

| <i>OrderID</i> | <i>VehicleID</i> | <i>SourceCity</i> | <i>DestinationCity</i> | <i>YearQuarter</i> | <i>AmountCharged</i> |
|----------------|------------------|-------------------|------------------------|--------------------|----------------------|
| 126            | V811             | Pune              | Mumbai                 | Q3                 | 2000                 |
| 127            | V911             | Pune              | Mumbai                 | Q4                 | 2000                 |
| 128            | V111             | Pune              | Mumbai                 | Q4                 | 2000                 |
| 129            | V211             | Pune              | Mumbai                 | Q4                 | 2000                 |
| 130            | V311             | Mumbai            | Pune                   | Q4                 | 2000                 |
| 131            | V411             | Mumbai            | Pune                   | Q4                 | 2000                 |
| 132            | V511             | Mumbai            | Pune                   | Q4                 | 2000                 |

Perform the following OLAP operations on the data in Table 3.13:

- (a) Slice
- (b) Dice
- (c) Pivot

**Solution:** We provide below the output of a Dice operation. Here the data has been diced/filtered along three dimensions: SourceCity, DestinationCity, and VehicleID.

The criteria are where (*SourceCity* = “Mysore” or *SourceCity* = “Pune”), (*DestinationCity* = “Bangalore” or *DestinationCity* = “Mumbai”) and (*VehicleID* = “V111” or *VehicleID* = “V411” or *VehicleID* = “V711”).

Perform the Slice and Pivot operations on the data in Table 3.13.

| <b>Sum of Amountcharged</b>            |                                 |                        | <b>Vehicle ID</b> |             |             |                    |
|--|---------------------------------|------------------------|-------------------|-------------|-------------|--------------------|
| <b>YearQuarter</b>                     | <b>SourceCity</b>               | <b>DestinationCity</b> | <b>V111</b>       | <b>V411</b> | <b>V711</b> | <b>Grand Total</b> |
| <input type="checkbox"/> Q1            | <input type="checkbox"/> Mysore | Bangalore              | 1500              |             |             | 1500               |
|  | <b>Mysore Total</b>             |                        | <b>1500</b>       |             |             | <b>1500</b>        |
| <b>Q1 Total</b>                        |                                 |                        | <b>1500</b>       |             |             | <b>1500</b>        |
| <input type="checkbox"/> Q2            | <input type="checkbox"/> Mysore | Bangalore              |                   | 1500        | 1500        | 3000               |
|  | <b>Mysore Total</b>             |                        |                   | <b>1500</b> | <b>1500</b> | <b>3000</b>        |
| <b>Q2 Total</b>                        |                                 |                        |                   | <b>1500</b> | <b>1500</b> | <b>3000</b>        |
| <input checked="" type="checkbox"/> Q3 |                                 |                        | 1500              |             |             | 1500               |
| <input type="checkbox"/> Q4            | <input type="checkbox"/> Pune   | Mumbai                 | 2000              |             |             | 2000               |
|  | <b>Pune Total</b>               |                        | <b>2000</b>       |             |             | <b>2000</b>        |
| <b>Q4 Total</b>                        |                                 |                        | <b>2000</b>       |             |             | <b>2000</b>        |
| <b>Grand Total</b>                     |                                 |                        | <b>5000</b>       | <b>1500</b> | <b>1500</b> | <b>8000</b>        |

## SOLVED EXERCISES

**Problem:** Construct an ER diagram and dimensional model for the following scenario: There is a publishing house which operates from its offices spread across the country. For their convenience, the publishing house classifies the country into four regions (North, South, East, and West). Each of their offices belongs to one of the regions. Several authors are in contract with the publishing house for the

publication of their work. An author can be associated with more than one publishing house. An author can be in contract for publication of one or more of their works (books), either with the same publishing house or multiple publishing houses. But a particular book by an author can be published by only one publishing house. It can also happen that the book has been co-authored.

**Solution:** The steps to create an ER diagram as follows:

### Step 1: Identify all possible entities

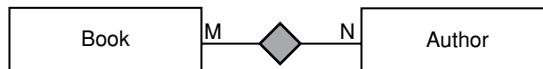
The entities are:

- Book.
- Author.
- Publisher.
- Region.

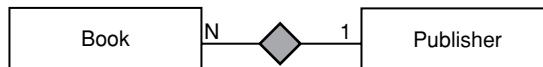
### Step 2: Identify all the relationships that exist between all the entities

Relationships between entities are as follows:

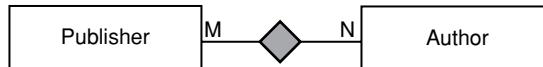
- An author could have written several books. A book could also have been co-authored (more than one author for the book).



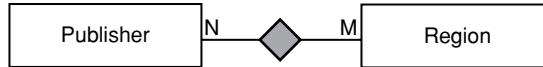
- A publishing house publishes several books. However, a particular book can be published by only one publishing house.



- A publishing house has several authors in contract with it. An author can have contract with several publishing houses.



- A publishing house has its presence in several regions. Likewise, a region has several publishing houses.



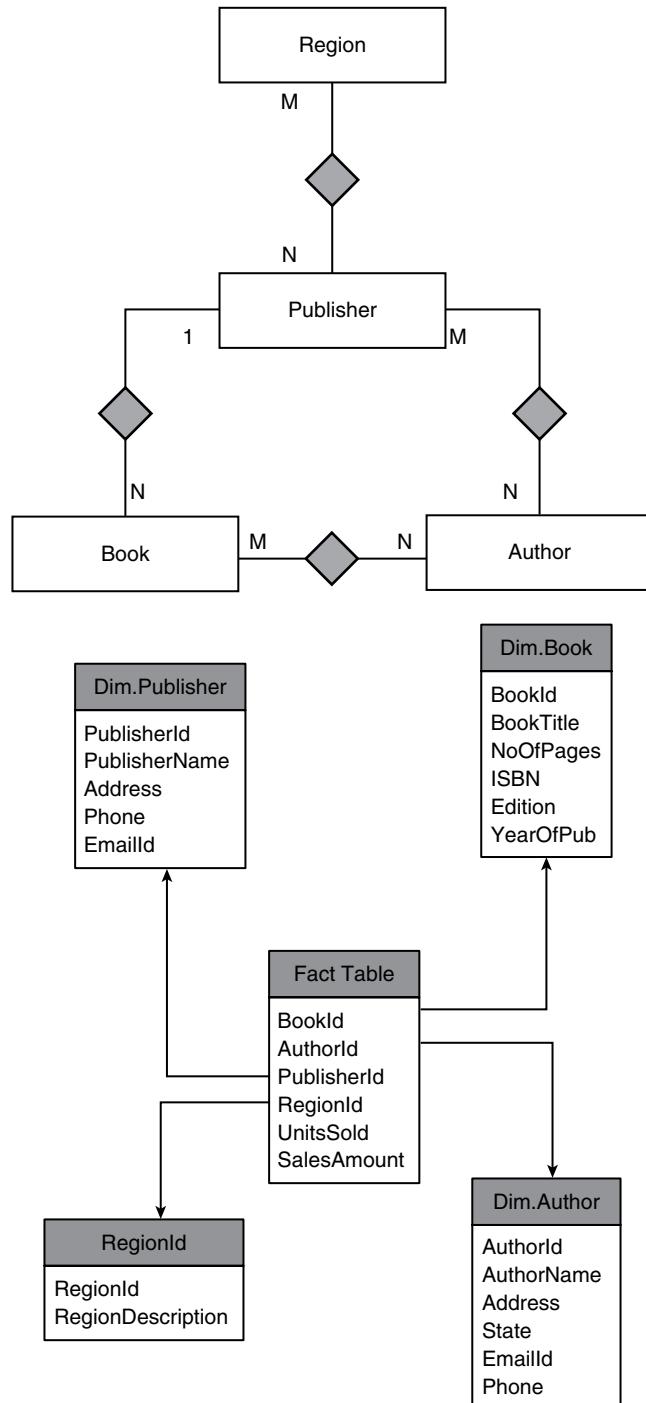
### Step 3: Identify the cardinality of the relationships (such as 1:1, 1:M, M:N)

The cardinalities of the relationships are as follows:

- M:N from Book to Author.
- N:1 from Book to Publisher.
- M:N from Publisher to Author.
- N:M from Publisher to Region.

### Step 4: Draw the ER diagram

*The dimensional model:* Let us look at the Star schema for the above scenario.



## UNSOLVED EXERCISES

---

1. How is OLTP different from ERP?
2. Should OLAP be performed directly on operational databases?
3. How is OLAP different from OLTP?
4. Explain the multidimensional data using an example.
5. How is MOLAP different from ROLAP?
6. State two advantages and two disadvantages of ROLAP.
7. Explain with an example the terms “Aggregations” and “Summarizations”.
8. What are fact and dimension tables?
9. Explain the slice and dice operations on data using an example.
10. Explain the drill-through and drill-across operations using an example.
11. How is a Snowflake schema different from a Star schema?
12. State the difference in data models for OLTP and OLAP.
13. Construct an ER diagram for a car insurance company that has a set of customers, each of whom owns one or more cars. Each car has associated with it zero to any number of recorded accidents.
14. Construct an ER diagram to illustrate the working of a bank. Make all the necessary assumptions.
15. Picture a retail company that sells a variety of products. The company would like to analyze their sales data by region, by product, by time, and by salesperson. Draw Star and Snowflake schema to depict it.
16. Explain the role of OLAP tools in the BI architecture.
17. What is a fact constellation?
18. What are the challenges faced by an OLTP system?
19. Comment on the type of data (structured, semi-structured, or unstructured) that is generated by OLTP system.
20. Which system (OLTP/OLAP) has support for concurrency control and recovery and why?
21. State a scenario where there is a requirement for an OLAP system.
22. State an instance from your daily life where an OLTP system is being used.
23. Why OLTP database designs are not generally a good idea for a data warehouse?
24. Is it important to have your data warehouse on a different system than your OLTP system?  
Explain your answer.
25. Compare data warehouse database and OLTP database.

# 4



## Getting Started with Business Intelligence

### BRIEF CONTENTS

|   |  |
|---|--|
| What's in Store?  | Where is BI being used?; When should you use BI?; What can BI deliver? |
| Using Analytical Information for Decision Support   | Evolution of BI and Role of DSS, EIS, MIS, and Digital Dashboards      |
| Information Sources before Dawn of BI?  | Need for BI at Virtually all Levels                                    |
| Definitions and Examples in Business Intelligence, Data Mining, Analytics, Machine Learning, Data Science | BI for Past, Present, and Future                                       |
| Looking at "Data" from Many Perspectives  | The BI Value Chain   |
| Business Intelligence (BI) Defined  | Introduction to Business Analytics                                     |
| Why BI?   | Unsolved Exercises   |

### WHAT'S IN STORE

You are now familiar with the big picture of a business enterprise and the role of IT in an enterprise. You can also now distinguish between OLTP and OLAP systems. Let's get to the heart of the subject matter of this book, viz., analytics that support business decisions. This is also referred to as Business Intelligence.

In this chapter we will familiarize you with the definition of Business Intelligence and the associated terminologies and the evolution of business intelligence technology domain to its current state. We will explain why businesses choose to leverage analytics for decision making. We will share several examples of industries like retail and hotel and about situations that are very familiar to us. You will be able to compare and contrast business analytics with ERP and data warehousing.

We suggest you conduct self-research in your area of interest leveraging the information available on the Internet.

## 4.1 USING ANALYTICAL INFORMATION FOR DECISION SUPPORT

---

In the past, leading market research firms noticed that often senior executives in businesses leveraged “numerical information” to support their decisions. They started using the term “Business Intelligence” (BI) for the set of concepts and processes that allows a business executive to make informed decisions. The IT applications providing such “numerical information” were commonly called “analytical applications” to distinguish them from transaction-oriented applications. The decision making became informed decision making with the use of BI. What is an “informed decision”? It is a decision based on fact and fact alone. Why is it required to make informed decisions? The simple reason is informed decisions based on fact, not on gut feeling, more often than not are the correct decisions. It’s easy to communicate “facts” to the large number of stakeholders. A large dispersed set of decision makers can arrive at the same conclusion when facts are presented and interpreted the same way. This type of decision making will lead to business benefits. It will provide insight into the operational efficiencies; it will help explore untapped opportunities; and above all it will serve as a window to the business dynamics and performance. It will help provide answers to questions, like *“Who are my most profitable customers?”*, *“Which are our most profitable products?”*, *“Which is the most profitable marketing channel?”*, *“What are the various up-sell and cross-sell opportunities?”*, *“Who are my best performing employees?”*, *“How is my company performing in terms of the customer expectations?”*, etc.

It’s is true that business executives, operations staff, and planning analysts all made decisions even when “Business Intelligence” or Business Analytics were not there. The evolution of BI made decision making faster, reliable, consistent, and highly team oriented.

## 4.2 INFORMATION SOURCES BEFORE DAWN OF BI?

---

Decision makers invest in obtaining the market facts and internal functions such as finance and marketing sales to evolve business plans. Some of the frequently used information sources are as follows:

- **Marketing research:** This analysis helps understand better the marketplace in which the enterprise in question is operating. It is about understanding the customers, the competitors, the products, the changing market dynamics, etc. It is to answer questions such as “Whether the launch of product X in region A will be successful or not?”, “Will the customers be receptive to the launch of product X?”, “Should we discontinue item Z?”, “Where should items A and B be placed on the shop shelves?”, etc.
- **Statistical data:** This is essentially about unravelling hidden patterns, spotting trends, etc. through proven mathematical techniques for understanding raw data. For example, variance in production rate, correlation of sales with campaigns, cluster analysis of shopping patterns, etc. help decision makers see new opportunities or innovate products or services.
- **Management reporting:** Most enterprises have their own IT teams dedicated to churn out ad hoc reports for the management. Often times they invest in specialized tools to prepare reports in graphical format.
- **Market survey:** Enterprises also employ third-party agencies to conduct consumer surveys and competitive analysis. They also use benchmark data to understand their strengths, weaknesses, and specific market opportunities they could exploit as well as risks that might reduce their revenue or market share.

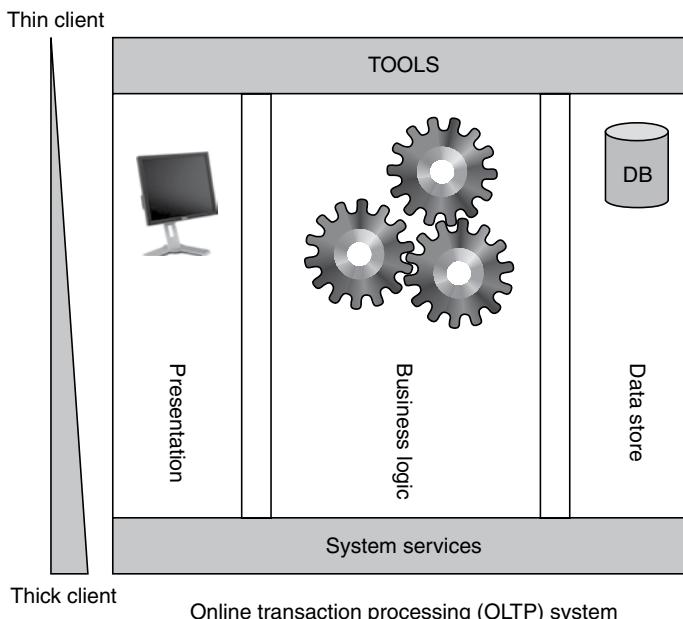
In fact, enterprises use one or more of the above methods but in more systematic ways, thanks to the availability of data in digital form and emergence of new delivery channels like the Internet and mobile devices.

### 4.3 DEFINITIONS AND EXAMPLES IN BUSINESS INTELLIGENCE, DATA MINING, ANALYTICS, MACHINE LEARNING, DATA SCIENCE

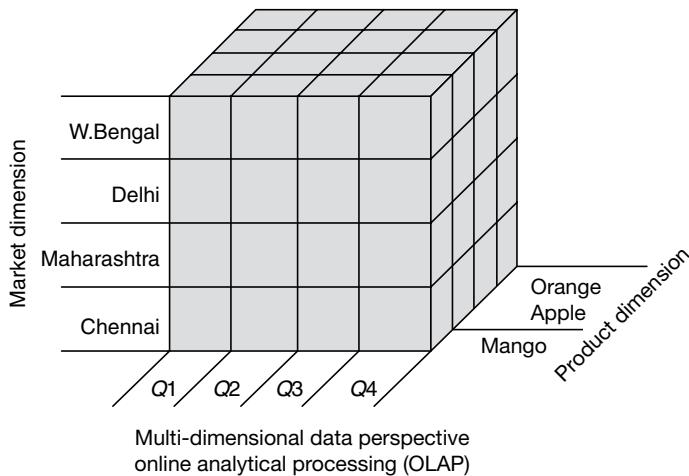
1. **Online Transaction Processing (OLTP):** OLTP systems refer to a class of IT applications that process and manage transaction-oriented digital data coming as inputs to the system and produce updated databases and management reports for routine decision making.

*Example:* An invoicing IT application will take inputs such as customer name, shipping address, products purchased, product price, tax rates and other parameters like date to generate a database (RDBMS Table) of invoices and also provide management a summary report of daily or weekly invoices raised by the company (Figure 4.1).

2. **Online Analytics Processing (OLAP):** OLAP is the system software used to combine data from several IT applications and provide business users with analyzed information to make informed business decisions. While relational databases are considered to be two-dimensional, OLAP data is multidimensional, meaning the information can be compared in many different ways. Thus, OLAP allows business users to easily and selectively extract and view data from different points of view.



**Figure 4.1** A typical OLTP application.



**Figure 4.2** Multidimensional data structure.

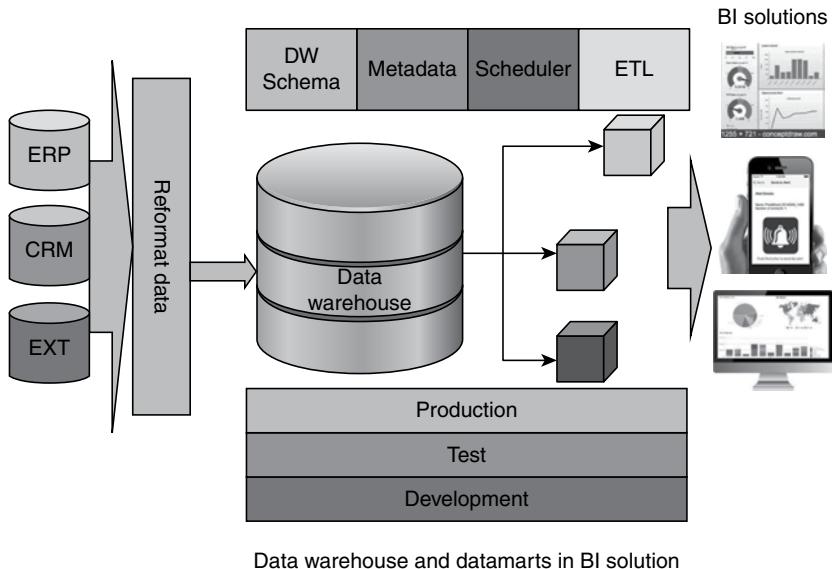
*Example:* A sales OLAP application will analyze product sales data from different perspectives like sales by product, sales by region, sales by sales person, sales by partners and so on and compute on-the-fly totals, averages, etc. to support decision making (Figure 4.2).

3. **Business Intelligence (BI):** BI can be defined as a set of *concepts* and *methodologies* to improve the decision making in businesses through the use of facts and fact-based IT systems. Thus the goal of BI is improved business decisions. It is more than technologies. It encompasses core *concepts* such as *Extract-Transform-Load (ETL)*, *Data Warehousing*, *Datamarts*, *Metadata*, *Metrics and KPIs*, *Scorecards*, *Dashboards* and *OLAP Reporting* as well as *methodologies* specific to ETL, Data Warehousing, Master Data Management and Data Governance. BI uses technology tools for ETL, Data Warehousing, Reporting, OLAP and Dashboards to achieve the decision support goals.

*Example:* A Human Capital Management data mart will get data from many IT applications holding employee-specific data such as employee profile, payroll, training, compensation, project performance and analyze employee productivity by department, management level, years of experience, sex and qualification. Such data marts will need ETL and OLAP reporting tools to function. Larger data warehouses will store centralized multi-dimensional data for several data marts.

Data Warehouse is a federated repository for all the data that an enterprise's various business systems collect. Typically the historical data of the enterprise is organized by subject areas such as employee, marketing, customer, product and so on. The purpose of creating data warehouse is to guide management in decision making with integrated and summarized facts. Such data warehouse creation will also need ETL tools, multi-dimensional data storage and OLAP reporting or data mining tools.

*Example:* An enterprise data warehouse in a bank will hold extracted summarized data taken from several OLTP IT applications in subject areas of loans, forex, corporate banking, personal banking, treasury, etc. Knowledge workers analyze the data in the warehouse from multiple perspectives like time (Day/Week/Month/Year), Bank location, Customer category to make business decisions and stay ahead of competition (Figure 4.3).



**Figure 4.3** An enterprise data warehouse.

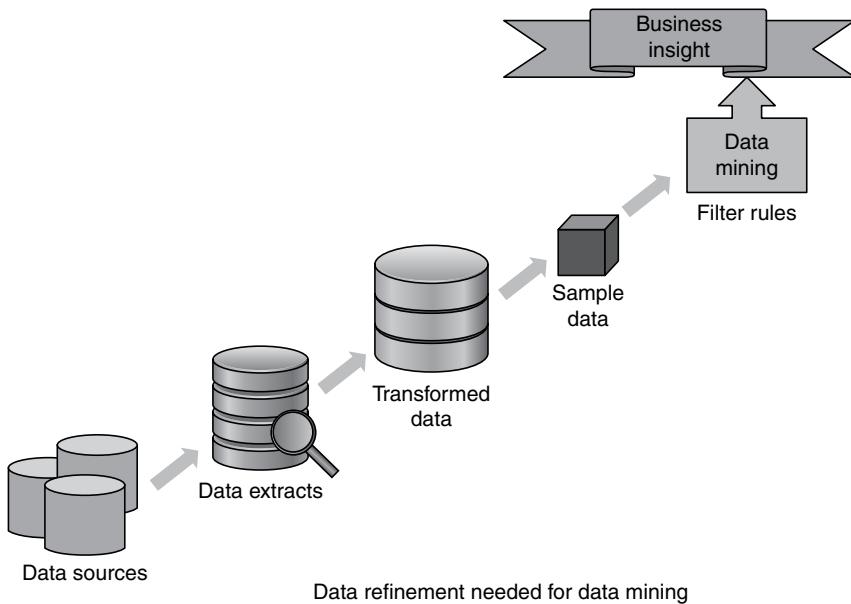
- 4. Data Mining:** Data mining is the computational process of sifting through existing business data to identify new patterns and establish relationships that will help in strategic decision making. This process of data mining will require pre-processing of data by ETL or getting data from data warehouse as well as post-processing for displaying the new insights found. Data mining uses the following techniques to discover new patterns in data:

- *Association*: looking for patterns where one event is connected to another event.
- *Sequence analysis*: looking for patterns where one event leads to another later event.
- *Classification*: looking for new patterns and groups that could be of business value.
- *Clustering*: finding and visually documenting groups of facts not previously known.
- *Forecasting*: discovering patterns in data that can lead to reasonable predictions about the future business events, conditions, etc.
- *Anomaly detection*: Finding unusual patterns in data sets.

*Example:* A data-mining specialist looking at online retail store purchases with buyer demographic and psychographic data may find new segments of buyers like individuals in old age homes, patients needing home healthcare services, children in residential schools and customize product/service offerings for them (Figure 4.4).

- 5. Big Data:** Big data is a high volume, high velocity, high variety information asset that demands cost-effective and innovative forms of information processing for enhanced business insight and decision making. Hence, big data involves homogeneous voluminous data that could be structured (as in RDBMS) or unstructured (as in blogs, tweets, Facebook comments, emails) and the content is in different varieties (audio, pictures, large text). Handling this type of data will need newer and innovative technologies for capturing, storing, searching, integrating, analyzing and presenting newly found insights.

*Example:* A big data application in telecom industry could be to analyze millions of calls data, billing data, marketing data, competitive data, carrier usage data, data usage and customer



**Figure 4.4** Data mining process: From data to insights.

profiles to accurately recommend a service that will meet the needs of the customer. In this situation, the volume and split second response by the big data analytics application is critical to engage the customer on call and close deals (Figure 4.5).

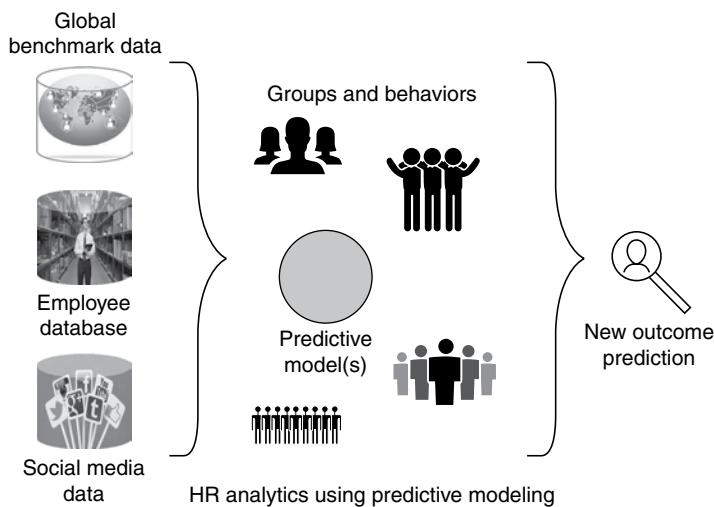


**Figure 4.5** Big data: Data from multiple disparate sources.

6. **Analytics:** Analytics is the computational field of examining raw data with the purpose of finding, drawing conclusions and communicating inferences. Data analytics focuses on inference - the process of deriving a conclusion based solely on what is already known by the researcher. Analytics relies on the simultaneous application of statistics, operations research, programming

to quantify observations. Analytics often uses data visualization techniques to communicate insight. Enterprises apply analytics to business data to describe (Exploratory analytics), predict (Predictive analytics) and automate to improve business operations (Prescriptive analytics). Increasingly, “Business analytics” is used to describe statistical and mathematical data analysis that clusters, segments, scores and predicts what scenarios are most likely to happen in businesses.

*Example:* HR analytics can be seen as the application of statistical techniques (like factor analysis, regression and correlation) and the synthesis of multiple sources to create meaningful insights – say, employee retention in office X of the company is driven by factors Y and Z (Figure 4.6).



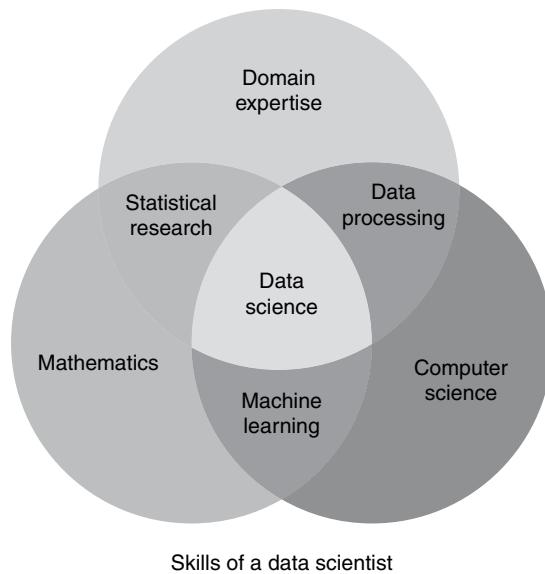
**Figure 4.6** Applying analytics to business data to describe, predict and prescribe.

**7. Data Science:** This is the science of extracting knowledge from data. The aim of data science is again to bring out hidden patterns amongst datasets using statistical and mathematical modeling techniques. Data science is an interdisciplinary field (Artificial Intelligence, Algorithms, Pattern recognition, NLP, Machine learning, Analytics, Computer science, Programming, Data mining, Business domain knowledge, etc.) about processes and systems to extract knowledge or insights from data. Data scientists use their data and analytical ability to find and interpret rich data sources, manage large amounts of data, combine data sources, ensure datasets quality, create visualizations to communicate understanding of analysis build mathematical models using the data, and enhance specific industry vertical business performance. The data scientist processes more heterogeneous, incomplete, messy and unstructured data than the highly curated data of the past. Digitized text, audio, and visual content, like sensor and blog data, etc. are the sources. As individuals may not easily acquire all these competencies, it is generally approached as a team competence.

*Example:* Data science solution for electricity company could encompass areas such as:

- Improve their monitoring and forecasts of energy consumption.
- Predict potential power outages and equipment failures.
- Drastically reduce their operational costs.
- Pinpoint inefficiencies.
- Manage energy procurement with greater precision.

This will need expertise to leverage smart grid data and model capacity, consumption, losses, and outages and suggest measures for optimum and profitable operations (Figure 4.7).

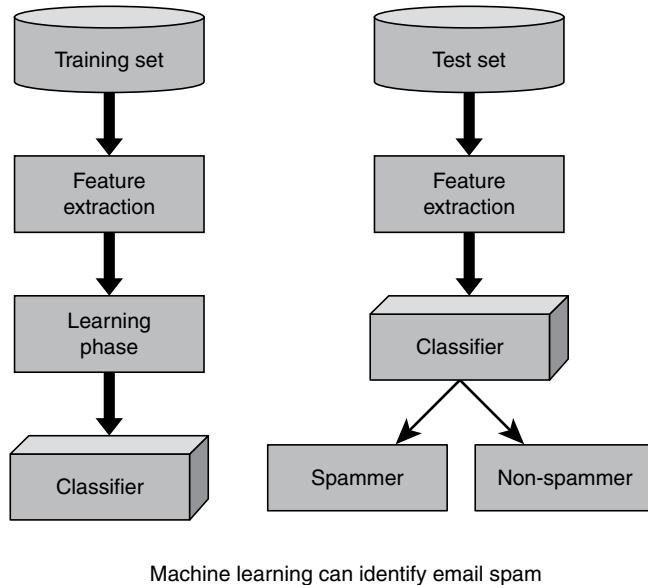


**Figure 4.7** Data science: Coming together of various skills.

8. **Machine Learning:** Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data. Machine learning is about building algorithms that build a model out of the input samples and make data driven decisions. Machine learning leverages concepts like decision trees, regression, clustering, collaborative filtering, artificial neural networks, inductive logic programming, support vector machines, etc. to develop algorithms to automate analytical model building to solve a particular type of problem. Here, as models are exposed to new data, they are able to independently adapt.

*Example:* Online recommendations that we see when you buy a book from online book store use machine learning. Here the algorithms look for people with similar profile and interests and their buying habits to build a model. As and when users buy new books from the store, it applies the filtering and ranking techniques to select high probability buyers. Most often it will be a very successful recommendation. This is today used for suggesting restaurants, hotel rooms to stay, electronics as well as gifts. Thus, machine learning can promote quick and smart actions without human intervention in real-time (Figure 4.8).

9. **Data Lake:** A data lake is a storage repository that holds a vast amount of collected raw data in its native format until it is needed for processing. The concept of a data lake is closely tied to Apache Hadoop and its ecosystem of open source projects extensively used in big data applications. Data lakes reduce data integration challenges. The data lake supports the following capabilities to:
  - Capture and store raw data **at scale for a low cost**.
  - Store many **types** of data (Structured to unstructured) in the same repository.



**Figure 4.8** Machine learning: AI in action.

- Perform transformations on the data.
- **Schema on read design** rather than schema at write time.
- Allow different communities of users like data scientists to gain access to raw data for their processing needs without much latency.

In many enterprises, the EDW (enterprise data warehouse) was created to consolidate information from many different sources so that reporting, data mining and analytics could serve decision makers. The enterprise data warehouse was designed to create a single version of the truth that could be used over and over again.

## 4.4 LOOKING AT “DATA” FROM MANY PERSPECTIVES

Globally, organizations have harnessed the power of “business, science or commerce facts or data” to make strategic decisions by using information technology. Over the years the data has posed several challenges such as volume, growth rate, variety of data, storage of data, retrieval of information, processing of data including complex computation, searching, ordering, merging, protecting, sharing, archival, backup and restoration. Many tools have been developed to solve these data-related challenges. In this section let us look at data from several perspectives before we start learning about different technologies and tools associated with data handling. Here are some perspectives to consider:

1. Data lifecycle perspective.
2. Data storage for processing.
3. Data processing and analysis perspective.
4. Data from business decision support perspective.

5. Data quality management aspects.
6. Related technology influences of data.

Each of these perspectives will help you understand the current state-of-the-industry and appreciate the different data technologies and management trends.

#### 4.4.1 Data Lifecycle Perspective

The data lifecycle stretches through multiple phases as data is created, used, shared, updated, stored and eventually either archived or disposed. Every business whether it is a bank, utility company, airline, telecom carrier, hotel or hospital will be generating business data each second. Picture these typical transactions we do that the businesses have to remember:

1. Withdrawal of cash in an ATM
2. Pay our electricity bill online
3. Cancel an airline ticket
4. Change the postpaid mobile phone plan
5. Check-in into a hotel
6. Download music from online store
7. Request for a lab test in a hospital
8. Shopping check-out at a retail store and payment
9. Post a comment in Facebook

The data could be viewed from a lifecycle perspective going through phases such as creation, storage, backup, processing like computation, share, update, archive and dispose (rarely).

One thing that will occur to us immediately is about the type of the data we deal with. The data in the current day context can be of different types or variety like numeric data, alphanumeric text, long text, date, currency, picture, audio segment, video clip, news feed, survey response, machine generated data (location) and so on. Modern data management is about handling the entire data lifecycle. That is everything that happens to a piece of data from the moment it is created to the moment it is destroyed.

Data management involves several activities like planning, collecting, assuring, describing, preserving, searching, integrating, securing, analyzing, formatting, sharing and archival.

We have witnessed six distinct phases in data management.

*Phase I:* Initially, data was processed manually including recording computing, reporting.

*Phase II:* This phase used punched-card equipment and electro-mechanical machines to sort and tabulate millions of records.

*Phase III:* The third phase was the age of stored data on magnetic tapes or disks and used computers to perform batch processing on sequential files.

*Phase IV:* The fourth phase introduced the concept of a Relational Database Management Systems (RDBMS) with a database schema and online access to the stored data.

*Phase V:* This phase automated access to relational databases and added distributed and client-server processing.

*Phase VI:* We are now in the early stages of sixth generation systems that store richer data types, such as complex eBooks, images, voice and video data.

#### 4.4.2 Data Storage (Raw) for Processing

We all agree that we need to store data for processing or back-up or archiving for later use. We need to understand the TYPES of data that need to be stored as well as VOLUME. Let us look at the data explosion we are witnessing in the era of Internet.

*According to the 2014 EMC/IDC Digital Universe report, data is doubling in size every two years. In 2013, more than 4.4 zeta bytes of data had been created; by 2020, the report predicts that number will explode by a factor of 10 to 44 zeta bytes - 44 trillion gigabytes. The report also notes that people - consumers and workers - created some two-thirds of 2013's data; in the next decade, more data will be created by things - sensors and embedded devices. In the report, IDC estimates that the IoT had nearly 200 billion connected devices in 2013 and predicts that number will grow 50% by 2020 as more devices are connected to the Internet - smartphones, cars, sensor networks, sports tracking monitors and more.*

We could look at the data storage from the user angle as well. We as individuals need to store data - documents we author, slides we prepare, spreadsheets we model, audio files we record/collect, videos we use, images we like and so on as well as keep a backup. We use hard disks, USB, cloud storage as personal data stores.

When it comes to a team working in a business office, they rely on network servers, public cloud, private cloud for storage at a function or department level. When it comes to enterprise level, there will be a huge storage facility on the network as well as cloud along with sophisticated backup, restore and archival facilities running to several terabytes or petabytes. Typically enterprises store critical data across multiple external storage systems. Many are using virtualized storage systems with automatic provisioning. Enterprises implement storage groups in order to plan, deploy, operate and administer storage systems. Storage technology segments today include Storage Area Networks (SAN), Backup-recovery-archival systems, remote replication, storage virtualization, cloud storage and local replication.

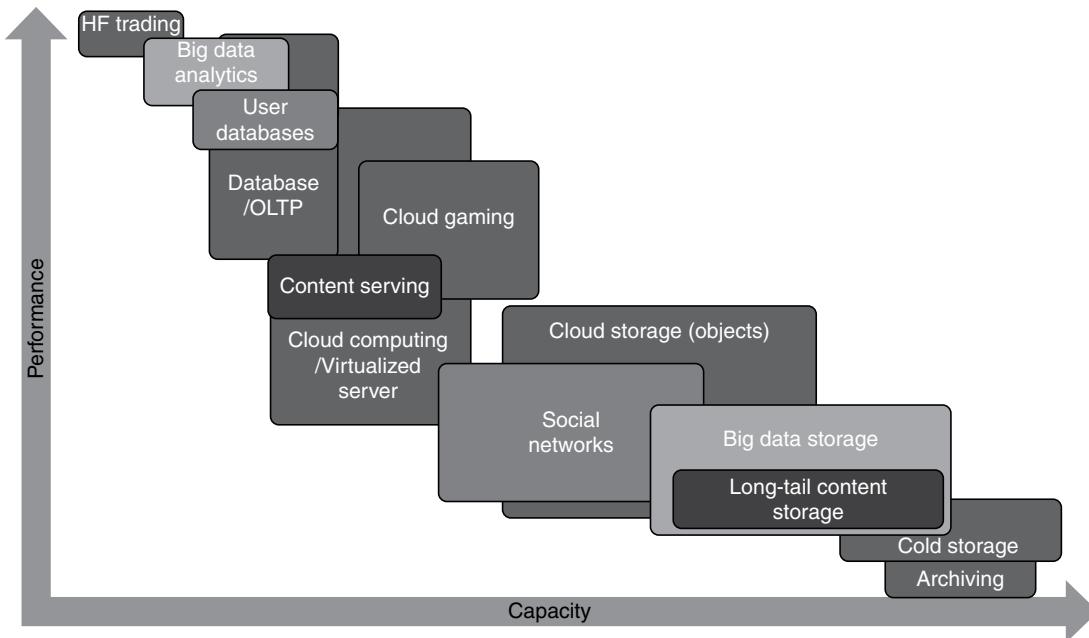
Some of the critical activities enterprises take-up to reliably store data include:

1. Managing data storage growth by increasing capacity.
2. Designing and implementing disaster recovery processes.
3. Deploying virtual storage environment.
4. Consolidating servers.
5. Implementing data archival solutions.
6. Moving data to cloud.

Enterprises consider availability, capacity, reliability, price-performance, power consumption (eco-friendly) and noise while deploying enterprise storage solutions. The applications used in the enterprise determine the user access performance needs. Figure 4.9 gives a snapshot of the applications demanding storage performance in a modern-day enterprise.

#### 4.4.3 Data Processing and Analysis Perspective

Let us now look at the computing platforms that the businesses have used to extract meaningful information from data. Here also we can see several generations of computer hardware used for data processing starting with the mainframe computers, mini computers (IBM AS400), then client-server based open computing systems using Unix operating system, personal computers, laptops and at present



**Figure 4.9** Various Data Storage: Their capacity and performance.

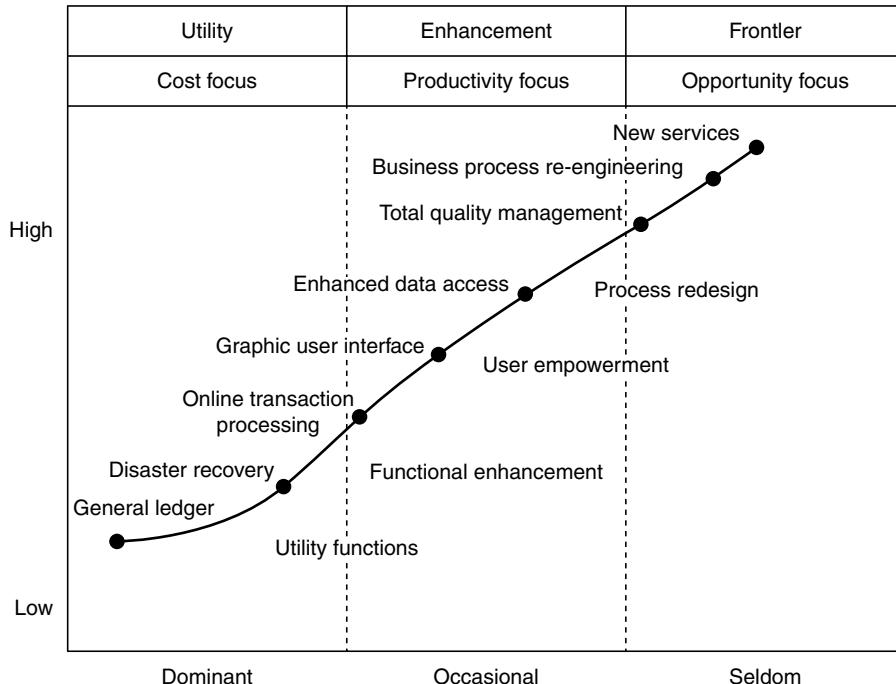
smartphones and tablets. While laptops, PCs, smartphones and tablets are used for personal computing, networked PCs, Internet connected smart devices can access even enterprise applications. Enterprises that serve thousands to millions of users tend to run large-scale applications on data center computers. The next extension of such computing is the cloud computing. The IT applications that run in enterprises could be classified as in Figure 4.10.

OLTP systems refer to a class of IT applications that process and manage transaction-oriented digital data coming as inputs to the system and produce updated databases and management reports for routine decision making. OLTP applications use RDBMS (such as Oracle, SQL Server, and DB2) as the main system for application data management. RDBMS provides SQL language interface for extracting, updating and combining data from different sources.

### ***Data Analysis Technologies***

Enterprises have discovered the value of “Data” as one of the key corporate assets. IT product companies have continuously invented solutions to address data handling problems in areas of volume, structure, computation capabilities, data integration, data analysis using statistical principles, ad-hoc querying, data replication, data security, discovering new patterns of data, using data to create self-learning models and so on. Let us look at these technological developments that have influenced the use of historical data for smart decision making across all levels in enterprises.

- 1. Spreadsheets:** Spreadsheets represent the use of software from PCs to tablets for manipulation of data using the structure of rows and columns. This digital document helps in numerical data entry, computation, sorting, searching, aggregating, perform statistical analysis, graph datasets and even create basic data mining tasks. MS Excel is an example of spreadsheet.



**Figure 4.10** Classification of IT applications of an enterprise.

2. **Ad-hoc Query Tools:** In any enterprise not all decision makers are satisfied with standard reports. They do not want to wait too long for the IT function to extract data from multiple sources and prepare a new report to their specification. Ad-hoc query tools allow extracting data from several RDBMS and other types of files and extract needed information as and when required. Crystal reports, BIRT, etc. are some examples of ad-hoc query tool.
3. **ETL Tools:** Extract-Transform-Load (ETL) is the process of combining selected data from multiple forces, re-format them and create a new database or a data warehouse. They support a library of standard transformations needed to handle different data types and even automate data transformation. Informatica is an example of such ETL tool.
4. **Business Intelligence (BI) and Reporting Tools:** When the enterprise creates data marts and data warehouses, the users will need sophisticated reporting tools that can show insights using scorecards, dashboards, pivot charts, etc. BI reporting tools allow end users to leverage the power of data warehouse by building role-specific visualizations. It is important to note that many of these tools support mobile devices like smartphones and tablets and alert users whenever there is a threshold breach.
5. **Data Mining Tools:** Data mining is the computational process of sifting through existing business data to identify new patterns and establish relationships that will help in strategic decision making. Such tools use principles of association, classification and clustering of data to discover new patterns. SAS miner is an example of data mining tool.
6. **Big Data Analytics Tools:** According to the Gartner IT Glossary, Big Data is high-volume, high-velocity and high-variety information asset that demands cost-effective, innovative forms

of information processing for enhanced insight and decision making. *Volume* refers to the amount of data. Many factors contribute to high volume: sensor and machine-generated data, networks, social media and much more. Enterprises are awash with terabytes and, increasingly, petabytes of big data. *Variety* refers to the number of types of data. Big data extends beyond structured data such as numbers, dates and strings to include unstructured data such as text, video, audio, click streams, 3D data and log files. *Velocity* refers to the speed of data processing. The pace at which data streams in from sources such as mobile devices, clickstreams, high-frequency stock trading, and machine-to-machine processes is massive and continuously fast moving. Big data mining and analytics help uncover hidden data patterns, unknown correlations and other useful business information. However, big data tools can analyze high-volume, high-velocity and high-variety information assets far better than conventional tools than RDBMS within a reasonable response time. R, Weka, Talend, Pentaho, BIRT/Actuate, Rapid miner, Hive, Mahout are some examples of big data analytics tools.

#### 4.4.4 Data from Business Decision Support Perspective

In most businesses, data moves from function to function to support business decision making at all levels of the enterprise. The age of the data and its value is highly related; data has the highest value as soon as it is created. But considering the challenges like big data, distributed storage, parallel processing, the trend is to find business value while the data is still in computer memory. This concept is evolving in the form of in-memory databases and stream processing.

We also need to see adding context and aggregation refines data progressively. Here is the value of refining data to move up in the value chain.

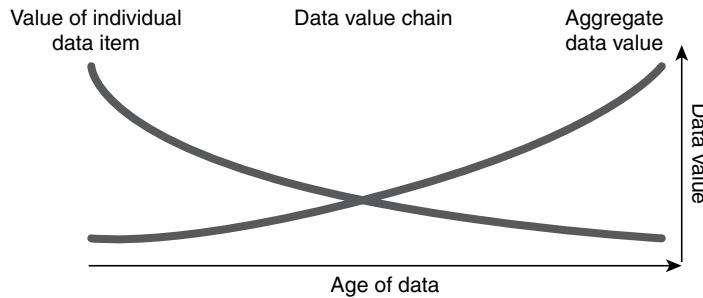
The true potential of data is realized through business results. While computer systems are capable of delivering both data and refined data or information to end users, they need to apply that information to take actions. Data mining can produce new knowledge about employee attrition, new market segments that can be targeted expanding the enterprise knowledge. For example, increasing cross-channel sales which is a business value, requires data about your current customers and the products they own.

We can take the value process to the next level using applied analytics. Here we use data mining as well as concepts of analytics like forecasting, predicting and even prescribing best action in a given business context. In this situation decision makers are influencing their future business performance by leveraging analytics. Analytics will also help in building enterprise smart decision framework based on very large datasets and high speed processing like Hadoop and MapReduce in big data processing.

The fast data segment of the enterprise addresses a number of critical requirements, which include the ability to ingest and interact with the data feed, make decisions on each event in the feed, and apply real-time analytics to provide visibility into fast streams of incoming data.

Another way to look at the value of information is shown in Figure 4.11.

First, let us look at the left-hand side of the figure. We can infer that all enterprises care about some entities or objects about which they would like to capture data to support business decisions. This is information/knowledge they have. While running business transactions, data gets generated and this “data” is stored in computer systems. Now we move to the right-hand side of the figure. This digital data could be shared with decision makers directly or aggregated, analyzed and interpreted by different roles in the enterprise. This processed data in general can be termed information useful for decision making. There can be various subject areas (as used in data warehousing terminology) of this “Information” and different consumers of such information.



**Figure 4.11** Age of data vs. its value.

Whatever be the “Processing” or “Analysis” we perform on the stored data (real-time to historical data), decision makers will find use for this. Hence analytics is termed as “The oil of the 21st century”.

#### 4.4.5 Data Quality Management Aspects

The quality of the data has total impact on the “Findings” from the data set and has a huge potential to misguide decision makers if the quality of the data is poor. Enterprises spend significant money to track and correct the quality of the data generated at different sources, departments, purchased, imported from other external sources.

Let us look at some of the common data problems or causes of poor data in companies.

1. **Typographical errors:** Often times the names, descriptions, etc. manually entered into computers may have typographical errors or be misspelled.
2. **Duplicates:** There could be duplicate records occurring that have the same information. Such duplicates will pose problems when columns with numeric information such as invoice amount, salary, etc. are totaled generating erroneous totals. RDBMS can automatically check such errors if set-up properly.
3. **Wrong codes:** Sometimes company may have given codes to entities like Employees, Items, etc. and that could be entered wrongly into computer systems. Such records will be hard to trace.
4. **Missing fields or records:** Data for some of the columns may be missing or not entered at all. RDBMS avoids such problems by making data fields mandatory. Sometimes an entire entry could have been missed in the file system.
5. **Computed using incorrect formulae:** Sometimes incorrect formulae may have been entered into worksheets that are used for data capture.
6. **Non-conformance entries:** It is always possible to enter numeric data into an alphanumeric column type. If such errors have occurred during data merge or bulk transfer, they may go undetected.
7. **Data accuracy:** It is possible that during data capture one may enter numeric data, say, with 2 decimal places but while storing the system may just store as integer. We lose the accuracy/precision during such errors.
8. **Data correctness:** When you have several numeric fields or alphanumeric fields defined in a sequence and data is shifted by one column, you can see the problems it can create.
9. **Un-updated data:** Some of the data sources could have become obsolete and data may not be current. Taking data from obsolete sources will provide wrong business picture to users.

**10. Swapped fields:** Data like first name, last name could have got interchanged while processing and can result in mismatches.

**11. Varying default values:** System generated defaults for some columns pose data analysis challenges.

In order to control the quality of your analysis and output results, you will need to make thorough checks on the data set. While it may not be possible to check for errors in billions of records, you will need robust data quality management processes. Some of the attributes of data quality are:

1. **Relevance:** Data meet the needs of users at different levels.
2. **Accuracy:** Data values obtained are close to the true values.
3. **Credibility:** Users have confidence in the statistics and trust the objectivity of the data.
4. **Accessibility:** Data can be readily located and accessed by authorized users using tools.
5. **Validity:** It refers to the correctness and reasonableness of data.
6. **Integrity:** What data is missing important relationship linkages? The inability to link related records together may actually introduce duplication across your systems.
7. **Interpretability:** Users can readily understand, use and analyze the data as well as describe the limitations of the data set.
8. **Coherence:** Statistical definitions and methods are consistent and any variations in methodology that might affect data values are made clear.
9. **Timeliness:** Delays between data collection and availability are minimized, without compromising accuracy and reliability.
10. **Periodicity:** Vital statistics are shared regularly so that they serve the ongoing needs of policy-makers for up-to-date information.
11. **Confidentiality:** Data-management practices are aligned with established confidentiality standards for data storage, backup, transfer and retrieval.

It is very important to assess the quality of the data before taking up data analysis work as it may be following in GIGO paradigm (Garbage-In-Garbage-Out).

The quality, reliability, security and usability of enterprise data dictate the business performance goals of the organization in driving new business opportunities, enhancing customer base, managing risk and compliance, generating smart decision support models, and reducing operating costs. The practice used to achieve these results is termed as “Data Governance”.

The scope of data governance covers areas like:

1. Defining data policies related to creation, storage, security, transmission, sharing, etc.
2. Communication of data policies with IT and business functions.
3. Be the central agency for managing data requirements for all key functions.
4. Documentation structural data standards.
5. Development of unified standard data models for persistence and sharing.
6. Active monitoring of data policy applications to data expectations.
7. Assessing the requirements from across the line of business landscape.
8. Driving information security at all levels of the enterprise.
9. Single agency for selection of data management tools in the enterprise.
10. Data architecture formulation, regulatory compliance and audits.
11. Promote the value of enterprise data assets.
12. Data risks management.
13. Metadata and business glossary management.

#### 4.4.6 Related Technology Influences of Data

Let us focus our attention on some of the technologies that are closely associated with data technologies and the synergy is creating a larger business impact. These include social media, cloud and mobility. Some of the key technology trends of 2015 for these technologies are indicated here.

Trends shaping data landscape include:

1. **Data Democratization:** Knowledge workers at all levels in the enterprise will have access to the information they need to perform their roles.
2. **Internet of Things (IoT):** More and more sensors will be capturing data at its source and huge amounts of data will be available for the enterprise to make use of.
3. **Algorithmic Decision Making:** Enterprises will code the “Learning algorithms” with patterns of decision making and automate decision making in several areas.
4. **Unstructured and Real-time Data:** Enterprises tend to focus on near real-time data with technologies for in-memory processing for reducing the latency of information availability for decision-making. Globally, enterprises are looking at social media analytics and real-time analytics for gaining competitive advantage in the marketplace.
5. **NOSQL Adoption Will Increase:** Enterprises do not want to focus on the structure of the data to be predefined as schema but would like to interpret data at runtime and store data without predefined structure.
6. **Increasing Cloud Storage:** Enterprises will be using cloud for data integration and sharing to facilitate real-time analytics.
7. **Data Architecture and Data Lakes:** Enterprises will dedicate teams to build reliable data platforms and data lakes will be built to address analytics needs.

Some of the social media trends include:

1. **In-the-moment-updates:** Enterprises will support near real-time updates about the areas of interest for the user groups or live streaming.
2. **Visual Search:** Facilities to search content visually and with context will increase. Rich media and cross social media platforms through integration will also increase.
3. **Enhanced Mobile Commerce:** Social media apps will allow more In-app interaction features including purchase.
4. **Data privacy Challenges:** Users will witness more challenges with regard to their personal data and security threats.
5. **Growing Importance of Content:** Social connections will leverage the interaction platform for exchanging content of common interest.

Some of the cloud computing trends include:

1. **Hybrid Cloud Environments:** Enterprises will use combination of public and private clouds to reduce costs, support mobile workforce securely.
2. **Cloud-based Decision Frameworks:** With data integration and real-time data coming to cloud, many decision-making frameworks for connected users are likely to happen in cloud platforms.
3. **Cloud Service Brokers:** It is likely that agencies will be the intermediaries who will negotiate the services quality delivered by cloud service providers.
4. **Applications Design Alignment:** Enterprises will align many applications to cloud environments to realize benefits of cloud.

Some of the mobility trends include:

1. **Mobile-first Applications:** Enterprises will develop mobile apps that will exploit the core capabilities of the mobile devices and sensors.
2. **Powerful Work-life Integrations:** Applications will provide smooth handoff to multiple devices to enable users working on multiple devices to complete tasks.
3. **Mobile Security:** Enterprises continue to focus on security considerations of mobile centric applications to ensure no data assets are lost to hackers.

## 4.5 BUSINESS INTELLIGENCE (BI) DEFINED

---

Howard Dresner, of the Gartner Group, in 1989 coined the term BI. He defined BI as “**a set of concepts and methodologies to improve decision making in business through use of facts and fact-based systems.**” Let us take a while to understand the terms used in the definition:

- The goal of BI is improved decision making. Yes, decisions were made earlier too (without BI). The use of BI should lead to improved decision making.
- BI is more than just technologies. It is a group of concepts and methodologies.
- It is fact-based. Decisions are no longer made on gut feeling or purely on hunch. They have to be backed by facts.

BI uses a set of processes, technologies, and tools (such as Informatica/IBM Datastage/Ab initio for extracting the data, SAS/IBM SPSS for analyzing the data, and IBM Cognos/Business Object for reporting the data) to transform raw data into meaningful information. BI mines the information to provide knowledge (KDD – Knowledge Discovery from Data) and uses the knowledge gained to provide beneficial insights; the insights then lead to impactful decision making which in turn provides business benefits such as increased profitability, increased productivity, reduced costs, improved operations, etc. The transformation of raw data to business benefits through BI may be depicted as

**Raw Data → Meaningful Information → Knowledge Discovery → Beneficial Insights → Impactful Decisions → Business Benefits**

In short, Business Intelligence is about providing the right information in the right format to the right decision makers at the right time. Some important features of Business Intelligence have been described below:

- **Fact-based decision making:** Decisions made through Business Intelligence are based purely on fact and history. It is about staying in tune with the data flowing through your business systems.  
Refer to the case study brief on “GoodFood Restaurants Inc.”. Let us try to understand fact-based decision making using the example of our “GoodFood” restaurant. Every restaurant will report the quantity of food wasted across the globe within six hours from the closing hour of restaurant. This data is aggregated and shared among all chefs, back office staff, operations manager, and marketing campaign teams. A team analyzes reasons and spot drivers of variance and set target to reduce wastage week-by-week. The same team tracks data and initiates actions to correct the process to reduce waste and achieve set target.
- **Single version of truth:** Put simply, a single version of truth means that if the **same piece of data** is available at more than one place, all the copies of the data should agree wholly and in every respect. BI helps provide single version of truth.

In our above example of the restaurant, picture a customer walking into the restaurant a little late for lunch. He asks at the reception about the availability of a particular cuisine (say, Thai cuisine) at the buffet lunch. The receptionist confirms the availability after checking on the networked computer system, and the customer proceeds to the dining area. On the way, the customer comes across the head waiter and asks the same question. The head waiter too confirms the availability, checking on his PDA (Personal Digital Assistant). This is the “single version of truth” wherein the same piece of information shared by multiple persons (the receptionist and the head waiter in our case) agrees wholly and in every respect.

- **360 degree perspective on your business:** BI allows looking at the business from various perspectives. Each person in the project/program team will look at the data from his/her role and will look for attributes that add value for decision making in his/her role.

In the GoodFood example, a “reservation table number” helps the steward escort guests to the right place in the dining area, helps the chef visit the guests to describe the “day’s speciality”, and helps the service staff reach for cleaning and rearrange table whenever needed. Similarly, the food wastage will be viewed by different departments with different perspectives – the finance by cost of wastage, the housekeeping by disposal methods, chefs by reason for rejection by guests, the quality team for finding innovative approaches to reduction, and the information systems team for devising measures that indicate improvement in processes.

- **Virtual team members on the same page:** In today’s business context, not all stakeholders or decision makers will be in the same building/geographic location. Businesses are highly distributed in nature and executives travel extensively. The team of people who work on a common purpose/project/business goal but are spread across geographic locations is termed as a virtual team. Technologies like BI bring them together and provide them the same facts at the speed of light in personalized forms.

## Picture this...

The GoodFood Restaurant chain team members are now spread across 10 countries but they are highly wired in the digital world. The chefs discuss new recipes, procurement teams find new suppliers, marketing team members share campaign innovations, and all operations managers and executives hold conference calls and video-based meetings anytime in planned ways. Every member is trained to use the restaurant portal, critical applications, and the email system. The performance dashboards are used to provide key business metrics based on their role for all executives. Even customers interact over mail, portal, Facebook, and Twitter.

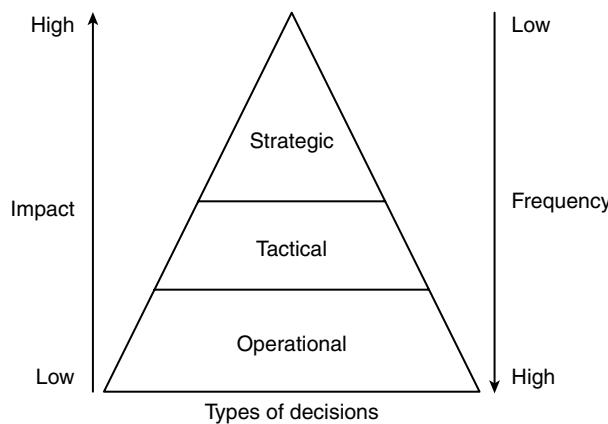
### 4.5.1 Visibility into Enterprise Performance

Business Intelligence provides a clear insight into the enterprise’s performance. This is by way of an operational dashboard. An operational dashboard makes use of visual forms such as charts, gauges, tables, matrix, indicators, etc.

In our example of the GoodFood Restaurant, the restaurant owner views the progress of the chain of restaurants on a single dashboard. This quickly tells him, with relevant data, which branch is doing excellent business, where the business is average with data, etc. Timely information as provided by the dashboard leads to early remedial procedures.

BI supports decision making at all levels in the enterprise as described below and as depicted in Figure 4.12.

- **Strategic level:** BI helps enterprises see the “big picture”. It supports decision making for the long term. Such decisions affect the entire organization. For the GoodFood Restaurants Inc., strategic decisions may include: “Where could be the next 5 restaurants?”, “Is cruise liner catering more attractive than flight kitchen management opportunity?”, etc.
- **Tactical level:** Tactical decisions in comparison to strategic decisions are made more frequently. In terms of the impact of decisions, tactical decisions affect a single unit(s)/department(s). For the GoodFood Restaurants Inc., tactical decisions may include: “How much discount should we offer for holidaying cruise liner guests?”, “What are the right months in the year for encouraging customers redeem loyalty points?”, etc.
- **Operational level:** Operational decisions are made even more frequently. The impact, however, is restricted to a single unit/department or function. These decisions help in conducting the day-to-day operations of business. For the GoodFood Restaurants Inc., an operational level decision may be: What menu item needs to be dropped this week to handle bad weather?



**Figure 4.12** Types of decisions supported by BI.

## 4.6 WHY BI? HOW CAN YOU ACHIEVE YOUR STATED OBJECTIVES?

1. You want to gain competitive advantage in the marketplace (based on better, faster and fact based decisions).
2. You want to retain your customers by making relevant recommendations of products and services to them.
3. You want to improve employee productivity by identifying and removing bottleneck processes.
4. You want to optimize your service offerings.
5. You want to bring NEW value to your business.

The above stated objectives can be achieved by harnessing the power of Business Intelligence (BI).

## 4.7 SOME IMPORTANT QUESTIONS ABOUT BI - WHERE, WHEN AND WHAT

---

### 4.7.1 Where is BI being used?

1. Netflix
2. Google
3. Yahoo
4. LinkedIn
5. Walmart
6. Facebook, etc.

### 4.7.2 When should you use BI?

If the answer to one or several of the below questions is “Yes”, you should consider BI:

1. Do you find yourself in situations where either you lack information to make the right decisions or you experience too much information but not enough insight?
2. Do you often end up discussing and debating the validity and accuracy of your reporting?
3. Are you unsure whether the key indicators on which you base your decisions are correct and whether they are up to date?
4. Do you spend a fair amount of time finding the needed information, classifying and structuring it and then distributing it to the right people at the right time?
5. Do you find yourself “drowning in spreadsheets” and do you find it tedious and cumbersome to integrate and consolidate data flowing in from multiple disparate sources?
6. Do you find it difficult and time-consuming to conduct ad-hoc analyses yourself?
7. Do you face problems of “information inaccessibility”; or is access dependent on one or two key persons?

### 4.7.3 What can BI deliver?

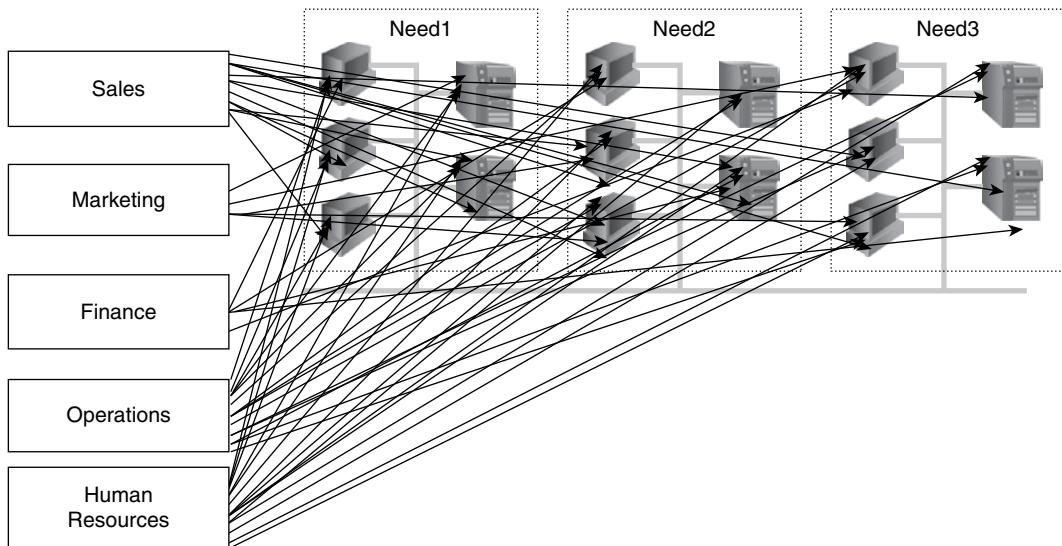
1. BI can deliver “single version of the truth”.
2. BI can deliver information that is actionable.
3. BI can glean “insights” from corporate data.

## 4.8 EVOLUTION OF BI AND ROLE OF DSS, EIS, MIS, AND DIGITAL DASHBOARDS

---

You may be interested to know how large businesses that have several IT systems (OLTP systems) provided “management information” for decision makers. Figure 4.13 illustrates the process typically followed for provision of “management information” to decision makers.

The IT function team members typically used to take responsibility for MIS (Management Information System). The new report preparation would involve all the phases of software lifecycle, viz., requirements gathering, analysis, design of new schema to combine data from several IT applications, programming to read data from existing IT applications, populating new schema, and then generating the required report. The major challenges associated with this approach include:



**Figure 4.13** Passing on “management information” to decision makers.

- Long delay between the request for and delivery of reports.
- Inaccurate figures as the data would have been copied to a new schema while IT applications could be updating their own databases resulting in multiple versions of truth.
- The copies of data (called extracts) could not serve any new requirement and had to be discarded. Too many versions can be totally confusing!
- Executives’ requirements would have changed by the time the report is taken to them, resulting in dissatisfaction of the service delivered. This increased the divide between those manning IT and management staff in many situations.

Several external forces started taking interest in tackling these challenges as the “solution” would be of high commercial value. First, IT function within enterprises started looking at new tools to connect to heterogeneous databases and pull data from them quickly. RDBMS vendors started offering adapters to connect to different vendor RDBMS products/versions but were expensive. Second, the new software companies started developing and marketing multidimensional databases, hardware solutions for handling queries faster (Netezza...), and new reporting tools that could seamlessly combine data from multiple sources. Third, in academia, research demonstrated newer approaches to de-normalization, OLAP, and data integration concepts with middleware. Business Intelligence solutions are a product of all these three developments. Some vendors integrated all the BI functionality into specific problem domain such as sales analysis and started delivering domain-customized BI tools. Today, we are in the third generation of BI tools. Some BI solutions are described below:

- **Ad hoc reporting:** Ad hoc reporting systems are used to meet the requirements of individual decision makers in the form, data set collection, and frequency. They are different from the regular management information systems as they are used to analyze the other information systems that are employed in the operational activities of the organization. The terms “MIS” and “information

systems” are often used interchangeably, rather incorrectly though. Information systems are not intended to support decision making, but MIS is there to support management functions. The reporting tools typically have the ability to combine data from multiple sources, store metadata, store report specifications for faster re-runs, and deliver reports in multiple forms such as PDF, Document, or worksheet formats.

- **DSS (Decision Support System):** In the 1970s, DSS became an area of research. It is an information system which supports business decision making activities. Also called knowledge-based system, DSS is known to support decision making that is required to run day-to-day operations. DSS supports applications such as inventory, point of sales systems, etc. It essentially supports operational decision making. The emphasis is on use of business graphics to present information from multiple sources.
- **EIS (Executive Information System):** EIS comes with powerful reporting and analytical abilities. It supports decision making at the senior management level, i.e. strategic decisions. It provides easy access to not just the internal data but also external data which are relevant in realizing the strategic goals of the enterprise. EIS typically focuses on metrics or KPI (discussed later in the book) that indicates the health of the functions/projects or business performance. It enables organizations to integrate and coordinate business process and has support for metric-based performance. Often times it is considered as a specialized form of DSS.

#### 4.8.1 Difference Between ERP (Enterprise Resource Planning) and BI

In Chapter 3, we briefly discussed ERP (Enterprise Resource Planning). Now, that we understand BI, it is worthwhile to look at a few points of difference, outlined in Table 4.1, between ERP and BI as an enterprise application.

#### 4.8.2 Is Data Warehouse Synonymous with BI?

Sometimes we see the terms data warehouse and BI being used interchangeably. Although used interchangeably, the term data warehouse is not a synonym for BI. Well, an organization might have a data

**Table 4.1** Differences between ERP and BI

| ERP  | <i>BI as an Enterprise Application</i>   |
|--|--|
| ERP is for data gathering, aggregation, search, update, etc.   | BI is for data retrieval.  |
| Essentially an operational/transactional/OLTP system.  | Essentially OLAP.  |
| Supports the capture, storage, and flow of data across multiple units of an organization.                    | Supports the integration of data from varied data sources, transforms the data as per business requirements, and stores it in the business data warehouse. |
| Has support for a few pre-built reports which usually help meet the transactional needs of the organization. | Supports advanced form of reporting (boardroom quality) and visualization. Has support for dynamic reports, drill down reports, drill across reports, etc. |
| Has little or no support for analytical needs of the organization.   | Supports the analytical needs of the organization.   |

warehouse but if adequate front-end tools are not made available to the people, what use is the data warehouse? In other words, *BI is the front-end while DW (data warehouse) is the back-end*. Yes, DW stores data but if the stored data is not converted to meaningful information and action is not based on the information, what use is the data? BI is more than just data warehouse. BI is also about analytics and reporting. BI is an umbrella term which encompasses marketing research, analytics, reporting, dashboards, data warehouse, data mining, etc.

## 4.9 NEED FOR BI AT VIRTUALLY ALL LEVELS

---

- **There is too much data, but too little insight!**

We have humungous amount of data and the volume and velocity of it continues to grow by leaps and bounds. There is a greater need than ever before to manage this data. There is a realization by enterprises/organizations that they do not have the information required to run their businesses efficiently. Data and information exists but more often in silos and its accuracy cannot always be trusted. More often, how the information is entered differs markedly from how it needs to be used to make business decisions. The definitions of data (the data schema) might vary from silos to silos. There is a need to integrate this data, and convert it into meaningful information that can be utilized effectively to drive business.

- **Business Intelligence has been there in the boardroom for long. There is a need to expand business intelligence from the boardroom to the front lines!**

Companies have to react faster to changing market conditions. This entails integrating business intelligence into operational processes. For example, in view of the prevailing market conditions, there may be an urgent need of alerting a call center worker to offer a particular promotion to a targeted customer segment. This is made possible by the availability of business intelligence tools at almost all levels of the corporations.

- **Structured and unstructured data need to converge!**

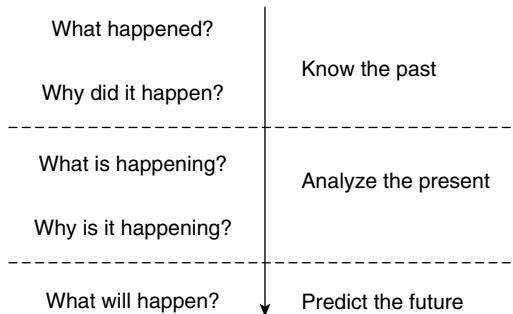
Unstructured data such as emails, voicemail messages, memos, etc. are rich sources of information. Along with the data that is available in the conventional rows and columns format (structured data), there is a need to blend the unstructured data to support better decision making. For example, adding suggestions/complaints/comments and other inputs from customers into a BI application can allow retailers to better their market segmentation analysis.

## 4.10 BI FOR PAST, PRESENT, AND FUTURE

---

BI is not only for the present context and scenario but also takes into consideration the past and the future. As indicated in Figure 4.14, BI with its set of standard reports helps answer questions such as “What happened?”, “When did it happen?”, “Where did it happen?”, etc. These are the reports that are generated on a regular basis, e.g. monthly or quarterly reports. These reports are useful in the short term and not necessarily seen as a support for long-term decisions.

BI with its statistical analysis capabilities allows us to dig deep into the current and past data to determine “Why is this happening?”, “Are we missing out on opportunities?”, etc. For example, retail stores can use statistical analysis to determine: Why the customers prefer a particular brand over another?



**Figure 4.14** BI for past, present, and future.

BI also helps in forecasting and predictive modelling. It helps get answers to questions like “What if the trend continues?”, “What is likely to happen next?”, “How much inventory will be needed and by when?”, “What will be in demand?”, etc.

Let us again get back to our now very familiar restaurant example. Every quarter, the restaurant owner along with his managers from the various restaurant branches looks through the spending on inventory supplies (food raw materials and soft drinks) by every restaurant branch. This way they are able to get a clear picture on which branch/branches spends the maximum on inventory supplies and the possible reason behind it. (The reason could be: There was a greater influx of customers on certain days, or there was a huge wastage owing to the food being spoilt.) This is a simple case of knowing the past.

### Picture yet another scenario...

One of the branch managers has raised a concern with the restaurant owner saying that of late (the last couple of weeks) he has observed a decline in the number of customers. He has also checked the possible reason behind it and found that a new ice-cream parlour has started service in the vicinity and is drawing a good crowd. The youth who used to frequent the restaurant for a quick bite are now slowly becoming a regular at the parlour. This is a simple case of analyzing the present.

Now let us look at predictive modelling in the light of our restaurant example. It has been observed by almost all restaurant branch managers that business picks up during the festive season. There is a surge in the number of people who prefer to eat out. Going by the trend and with the festive season (“Christmas” and “New Year”) round the corner, it is the right time to think about stockpiling the inventory supplies. Now, the simple question is: “What to stockpile?” and “How much?”

## 4.11 THE BI VALUE CHAIN

Let us try and understand the BI value chain. We would like to depict it as follows:

**Transformation → Storage → Delivery**

Data from different OLTP/transactional systems is brought together into an enterprise data warehouse (it could have been very easily a data mart as well). This is after the data has been cleansed and is free of all errors/defects. The data has also been transformed. One of the reasons behind transformation is to convert the data existing in different formats in the various data sources to a unified format. The data is

then loaded into the data warehouse. The next step in the BI value chain is data/information delivery. This topic will be dealt with in greater depths in the chapters to follow.

## 4.12 INTRODUCTION TO BUSINESS ANALYTICS

Business analytics is heavily dependent on data. For its successful implementation, business analytics requires a high volume of high quality data. The challenges faced by business analytics are: storage, integration, reconciliation of data from multiple disparate sources across several business functions, and the continuous updates to the data warehouse. Let us take a while to understand the difference between business intelligence and business analytics (Table 4.2).

**Table 4.2** Differences between business intelligence and business analytics

|                               | <i>Business Intelligence</i>   | <i>Business Analytics</i>  |
|-------------------------------|--|--|
| <b>Answers the questions:</b> | <ul style="list-style-type: none"> <li>• What happened?</li> <li>• When did it happen?</li> <li>• Who is accountable for what happened?</li> <li>• How many?</li> <li>• How often?</li> <li>• Where did it happen?</li> </ul>                    | <ul style="list-style-type: none"> <li>• Why did it happen?</li> <li>• Will it happen again?</li> <li>• What will happen if we change <math>x</math>?</li> <li>• What else does the data tell us that we never thought to ask?</li> <li>• What is the best that can happen?</li> </ul> |
| <b>Makes use of:</b>          | <ul style="list-style-type: none"> <li>• Reporting (KPIs, metrics)</li> <li>• Automated Monitoring/Alerting (thresholds)</li> <li>• Dashboards/Scorecards</li> <li>• OLAP (Cubes, Slice &amp; Dice, Drilling)</li> <li>• Ad hoc query</li> </ul> | <ul style="list-style-type: none"> <li>• Statistical/Quantitative Analysis</li> <li>• Data Mining</li> <li>• Predictive Modeling</li> <li>• Design of experiments to extract learning out of business data</li> <li>• Multivariate Testing</li> </ul>                                  |

What benefits does analysis of business data lead to? It can help businesses optimize existing processes, better understand customer behavior, help recognize opportunities, and also help in spotting problems before they happen.

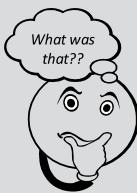
*Gartner defines an “analytic” application as packaged BI capabilities for a particular domain or business problem.* [<http://www.gartner.com/technology/research/it-glossary/>]

Let us look at a few basic domains within business analytics:

- Marketing analytics.
- Customer analytics.
- Retail sales analytics.
- Financial services analytics.
- Supply chain analytics.
- Transportation analytics, etc.

GoodFood Restaurants Inc. has always relied on analytics for driving its decisions. We discuss here the food quality domain of GoodFood Restaurants. The decision to set up “International Master Chefs School” in 2006 was based on the analysis of feedback which was collected from customers. Some of its outlets in the USA and UK had received very good feedback from the customers with a special mention on the awesome quality of food and good dining experience. However, a few outlets in the USA, UK,

and most of its outlets in Australia had received mixed feedback ranging from “Good” to “Average” to “Poor”. An analysis of the feedback data led the management of GoodFood Restaurants Inc. to make the decision on setting up “International Master Chefs School” where all the chefs in service at the various outlets can come together to learn a consistent approach in preparing and serving the various cuisines and share their best practices. The decision paid off and today, GoodFood Restaurants Inc. is a name to reckon with.



### *Remind Me (Forget Me Not!)*

- OLTP systems refer to a class of IT applications that process and manage transaction-oriented digital data coming as inputs to the system and produce updated databases and management reports for routine decision making.
- Data mining is the computational process of sifting through existing business data to identify new patterns and establish relationships that will help in strategic decision making.
- Data Science is the science of extracting knowledge from data.
- Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed.



### *Point Me (Books)*

- *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die* by Eric Siegel and Thomas H. Davenport (2013).
- *Data Science for Business: What you need to know about Data Mining and Data-Analytic Thinking* by Foster Provost and Tom Fawcett (2013).



### *Connect Me (Internet Resources)*

- <http://searchbusinessanalytics.techtarget.com/news/1507222/Text-analytics-search-bolster-business-intelligence-software-stack-says-TDWI>
- <http://searchbusinessanalytics.techtarget.com/news/2240019695/BI-search-platform-eyes-middle-ground-between-BI-and-unstructured-data>
- [http://www.dwreview.com/DW\\_Overview.html](http://www.dwreview.com/DW_Overview.html)



## Test Me Exercises

### Match me

| Column A                    | Column B          |
|-----------------------------|-------------------|
| ERP                         | Data Retrieval    |
| BI                          | Filtering of data |
| Single version of truth     | Dashboards        |
| Data slicing                | BI                |
| Charts, gauges, pivot table | Data input        |

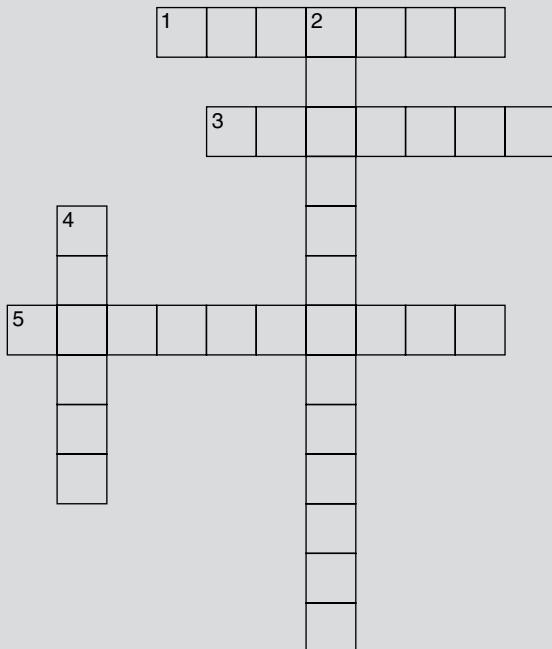
### Solution:

| Column A                    | Column B          |
|-----------------------------|-------------------|
| ERP                         | Data input        |
| BI                          | Data retrieval    |
| Single version of truth     | BI                |
| Data slicing                | Filtering of data |
| Charts, gauges, pivot table | Dashboards        |



## BI Crossword

### Introduction to BI



#### ACROSS

1. An act of piling up
3. An act of pulling out data
5. A company or a corporate

#### DOWN

2. A place where the inventory of data resides
4. Process of getting a metal from its ore

#### Solution:

- |                   |               |
|-------------------|---------------|
| 1. Loading        | 4. Mining     |
| 2. Data warehouse | 5. Enterprise |
| 3. Extract        |               |

## UNSOLVED EXERCISES

---

1. What is Business Intelligence (BI)?
2. Why is BI required by businesses?
3. What is fact-based decision making? Explain with the help of examples.
4. What is the single version of truth? Explain with the help of an example.
5. How are DSS, EIS, and MIS related to Business Intelligence?
6. How is ERP different from BI?
7. Can the terms BI and data warehouse be used interchangeably? Justify your answer.
8. Is BI required for operational decision making? Explain.
9. Can reports be drawn on OLTP system?
10. How do you think BI contributes to the future?
11. Is BI used only to analyze past data? Comment.
12. Explain how BI contributes to providing visibility into the enterprise performance?
13. Give a few examples of strategic, tactical and operational decisions.
14. How can BI be leveraged to provide a 360 degree perspective on the business?
15. Describe the BI value chain.
16. What is business analytics?
17. Assume you are in the hospitality business. What are the key questions that BI will help you answer?
18. Assume you are the owner of an automobile store. What are the key questions that BI will help you answer?
19. Think of an example from your college life where you ran into difficulty because of the single version of truth being compromised.
20. Picture this scenario... An enterprise has an ERP system in place, which has been running successfully. Would you still suggest them to go for Business Intelligence and why?
21. Explain a few techniques of Data Mining to discover new patterns in data.
22. Explain the difference between descriptive, predicative and prescriptive analytics.
23. What is exploratory analysis? Explain with an example.
24. What are the skills required to transform into a Data Scientist?
25. What is Internet of Things (IoT)?

# 5



## BI Definitions and Concepts

---

### BRIEF CONTENTS

|                                    |                               |
|------------------------------------|-------------------------------|
| What's in Store                    | BI Roles and Responsibilities |
| BI Component Framework             | Best Practices in BI/DW       |
| Who is BI for?                     | The Complete BI Professional  |
| BI Users                           | Popular BI Tools              |
| Business Intelligence Applications | Unsolved Exercises            |

---

### WHAT'S IN STORE

Chapter 4, “Getting Started with Business Intelligence”, has familiarized you with the industry definition of Business Intelligence, the key terms used in the BI technology domain, the evolution of BI, and different methods adopted to serve managements with the information requested by them. This chapter leads you to learn about BI in depth.

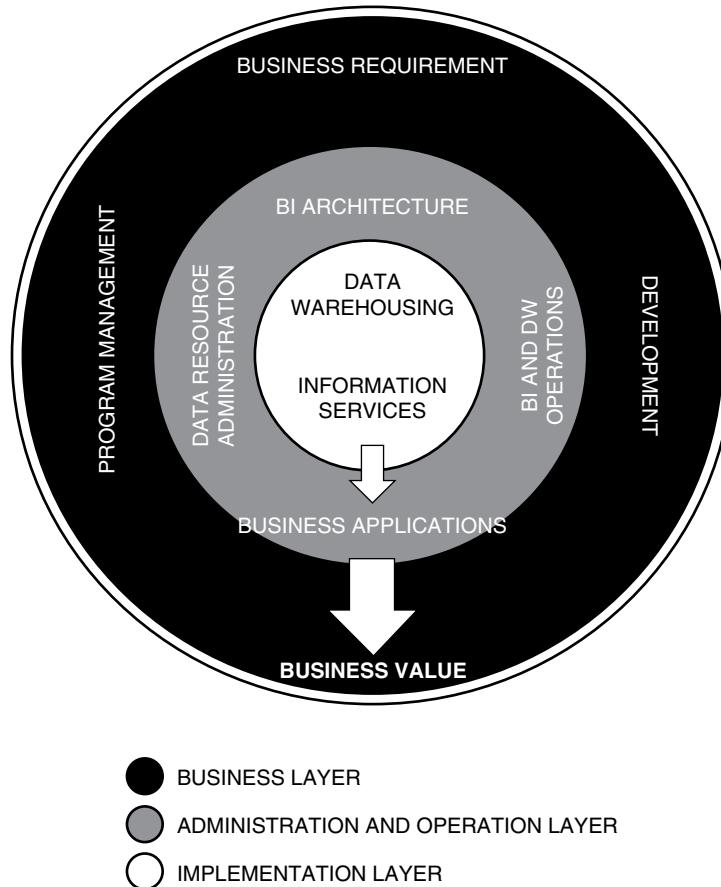
This chapter will introduce you to the BI framework and the various layers that constitute the BI architecture; familiarize you with casual and power users of BI; introduce you to BI applications, etc. We suggest the topics “Best Practices in BI/DW” and “The Making of a Complete BI Professional” as a “Must Read” for those who wish to gain a good understanding of the BI space.

We suggest you refer to the learning resources suggested at the end of almost every topic and also complete the “Test Me” exercises. You will get deeper knowledge by interacting with people who have shared their project experiences in blogs. We suggest you make your own notes/bookmarks while reading through the chapter.

---

### 5.1 BI COMPONENT FRAMEWORK

In today’s warehouse environment, the organizations which are successful are those with sound architectures. Ever wondered, why architectures are important? The answer is simple: They support the



**Figure 5.1** The BI component framework.

functional, technical, and data needs of the enterprise. In other words, they help the organization/enterprise become better equipped to respond to the business questions/queries posed by the users.

As depicted in Figure 5.1 the **BI component framework** can be divided into three major layers:

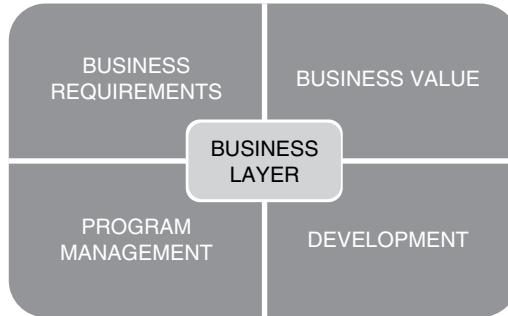
- Business layer.
- Administration and Operation layer.
- Implementation layer.

### 5.1.1 Business Layer

Figure 5.2 depicts the business layer of the BI component framework. This layer consists of four components – business requirements, business value, program management, and development.

#### ***Business Requirements***

The requirements are a product of three steps of a business process, namely, Business drivers, Business goals, and Business strategies:



**Figure 5.2** The business layer.

- **Business drivers:** These are the impulses that initiate the need to act. A few examples of business drivers are: changing workforce, changing labour laws, changing economy, changing technology, etc.
- **Business goals:** These are the targets to be achieved in response to business drivers. A few examples of business goals are: increased productivity, improved market share, improved profit margins, improved customer satisfaction, cost reduction, etc.
- **Business strategies:** These are the planned course of action that will help achieve the set goals. A few examples of business strategies are: outsourcing, global delivery model, partnerships, customer retention programs, employee retention programs, competitive pricing, etc.

### **Business Value**

When a strategy is implemented against certain business goals, then certain costs (monetary, time, effort, information produced by data integration and analysis, application of knowledge from past experience, etc.) are involved. However, the final output of the process should create such value for the business whose ratio to the costs involved should be a feasible ratio. The business value can be measured in terms of ROI (Return on Investment), ROA (Return on Assets), TCO (Total Cost of Ownership), TVO (Total Value of Ownership), etc. Let us understand these terms with the help of a few examples:

- **Return on Investment (ROI):** We take the example of “Digicom”, a digital electronics goods company which has an online community platform that allows their prospective clients to engage with their users. “Digicom” has been using social media (mainly Twitter and Facebook) to help get new clients and to increase the number of prospects/leads. They attribute 10% of their daily revenue to social media. Now, that is an ROI from social media!
- **Return on Asset (ROA):** Suppose a company, “Electronics Today”, has a net income of \$1 million and has total assets of \$5 million. Then, its ROA is 20%. So, ROA is the earning generated from invested capital (assets).
- **Total Cost of Ownership (TCO):** Let us understand TCO in the context of a vehicle. TCO defines the cost of owning a vehicle from the time of purchase by the owner, through its operation and maintenance to the time it leaves the possession of the owner.
- **Total Value of Ownership (TVO):** TVO has replaced the simple concept of Owner’s Equity in some companies. It could include a variety of subcategories such as stock, undistributed dividends, retained earnings or profit, or excess capital contributed. In its simplest form, the basic accounting equation containing TVO as a component is

$$\text{Assets} = \text{Liabilities} + \text{Owner's Equity}, \text{or if you like TVO}$$

### **Program Management**

This component of the business layer ensures that people, projects, and priorities work in a manner in which individual processes are compatible with each other so as to ensure seamless integration and smooth functioning of the entire program. It should attend to each of the following:

- Business priorities.
- Mission and goals.
- Strategies and risks.
- Multiple projects.
- Dependencies.
- Cost and value.
- Business rules.
- Infrastructure.

### **Development**

The process of development consists of *database/data warehouse development* (consisting of ETL, data profiling, data cleansing, and database tools), *data integration system development* (consisting of data integration tools and data quality tools), and *business analytics development* (about processes and various technologies used).

## **5.1.2 Administration and Operation Layer**

Figure 5.3 depicts the administration and operation layer of the BI component framework. This layer consists of four components – BI architecture, BI and DW operations, data resource administration, and business applications.

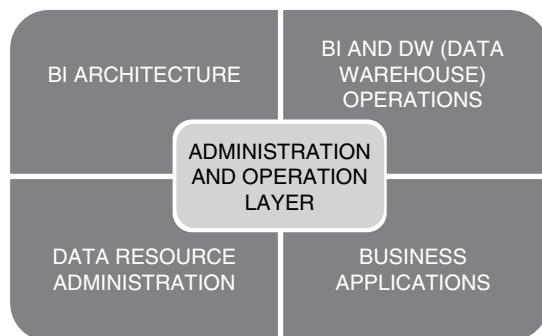
### **BI Architecture**

The various components of BI architecture are depicted in Figure 5.4.

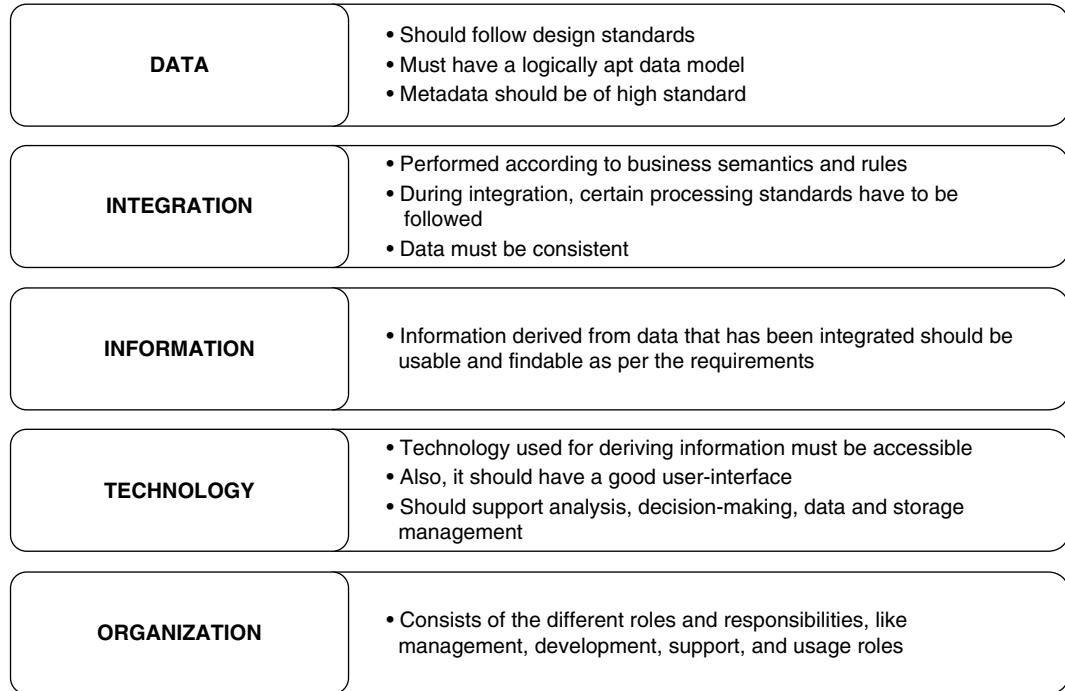
### **BI and DW Operations**

Data warehouse administration requires the usage of various tools to monitor the performance and usage of the warehouse, and perform administrative tasks on it. Some of these tools are:

- Backup and restore.



**Figure 5.3** The administration and operation layer.



**Figure 5.4** The components of BI architecture.

- Security.
- Configuration management.
- Database management.

### **Data Resource Administration**

It involves *data governance* and *metadata management*.

### **Data Governance**

It is a technique for controlling data quality, which is used to assess, improve, manage, and maintain information. It helps define standards that are required to maintain data quality. The distribution of roles for data governance is as follows:

- Data ownership.
- Data stewardship.
- Data custodianship.

### **Metadata Management**

Picture yourself looking at a CD/DVD of music. What do you notice? Well, there is the date of recording, the name of the artist, the genre of music, the songs in the album, copyright information, etc. All this information constitutes the metadata for the CD/DVD of music. In the context of a camera, the data is the photographic image. The metadata then is the date and time when the photograph was

taken. In simple words, metadata is data about data. When used in the context of a data warehouse, it is the data that defines the warehouse objects. Few examples of metadata are timestamp at which the data was extracted, the data sources from where metadata has been extracted, and the missing fields/columns that have been added by data cleaning or integration processes. Metadata management involves tracking, assessment, and maintenance of metadata. Metadata can be divided into four groups:

- Business metadata.
- Process metadata.
- Technical metadata.
- Application metadata.

Let us understand the various types of metadata with the help of an example:

### Picture this...

“ElectronicsForAll” is an India-based company with branches in all the major cities of India. Each branch has its own set of databases. The company has a data warehouse which is sourced from these several databases physically located at different locations (major cities of India such as Delhi, Mumbai, Chennai, etc.). The company also has a dashboard (a web application) that projects the daily and quarterly revenue earnings of the various branches. Let us now look at some of the technical details:

A partial structure of table “RevenueEarnings” existing in the databases at various branches of “ElectronicsForAll” is shown in Table 5.1. Remember, it is only a partial structure.

**Table 5.1** RevenueEarnings

| BranchID | LocationID | Quarter | Year | RevenueEarned | CostIncurred | Profit | ProfitMargins |
|----------|------------|---------|------|---------------|--------------|--------|---------------|
|----------|------------|---------|------|---------------|--------------|--------|---------------|

Table 5.1 has columns such as “BranchID”, “LocationID”, “Quarter”, “Year”, “RevenueEarned”, “CostIncurred”, “Profit”, “ProfitMargins”, etc. Each of these columns has a data type and a data size such as:

| Column Name | Data Type and Data Size | Constraints |
|-------------|-------------------------|-------------|
| BranchID    | Varchar(15)             | Not Null    |
| LocationID  | Varchar(30)             | Not Null    |
| Quarter     | Varchar(2)              | Not Null    |

Further, Table 5.1 is indexed on “BranchID”. The following constitutes the **technical metadata**:

- Data locations.
- Data formats.
- Technical names.
- Data sizes (in the example above, 15 as the maximum length for “BranchID”).
- Data types (in the example above, Varchar(15) for “BranchID”).
- Indexing (in the example above, the index on “BranchID”).
- Data structures, etc.

The data warehouse for the company “ElectronicsForAll” is asynchronously and incrementally updated by using the data updates from the data sources to the data warehouse.

The following constitutes the ***process metadata***:

- Source/target maps.
- Transformation rules.
- Data cleansing rules.
- Extract audit trail.
- Transform audit trail.
- Load audit trail.
- Data quality audit, etc.

The company’s dashboard is built on top of the data warehouse. The dashboard is accessible by all the branch heads and senior executives of the company.

The following constitutes the ***application metadata***:

- **Data access history such as:**
  - Who is accessing? Who has accessed?
  - Frequency of access?
  - When was it accessed?
  - How was it accessed?, etc.

***Business metadata*** captures information such as business definitions, structure, and hierarchy of the data, aggregation rules, ownership characteristics (who are the data stewards/data owners/data custodians), subject areas (such as finance, market share, etc.), and business rule-based descriptions of transformation rules and definitions of business metrics (such as which branch is the top revenue grosser, which branch has incurred the maximum expenses, how many branches are there in each location, etc.).

### ***Business Applications***

The application of technology to produce value for the business refers to the generation of information or *intelligence* from data assets like data warehouses/data marts. Using BI tools, we can generate *strategic*, *financial*, *customer*, or *risk* intelligence. This information can be obtained through various BI applications, such as DSS (decision support system), EIS (executive information system), OLAP (on-line analytical processing), data mining and discovery, etc.

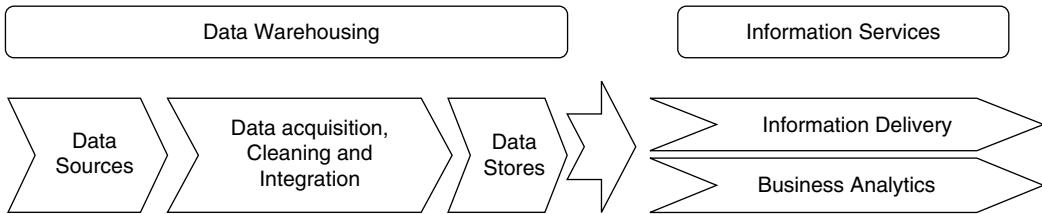
Few of the BI applications are discussed in the section “Business Intelligence Applications” later in this chapter.

#### **5.1.3 Implementation Layer**

The implementation layer of the BI component framework is depicted in Figure 5.5. This layer consists of technical components that are required for data capture, transformation and cleaning, converting data into information, and finally delivering that information to leverage business goals and produce value for the organization.

#### ***Data Warehousing***

- It is the process which prepares the basic repository of data (called *data warehouse*) that becomes the data source where we extract information from.



**Figure 5.5** The implementation layer.

- A *data warehouse* is a data store. It is structured on the dimensional model schema, which is optimized for data retrieval rather than update.
- Data warehousing must play the following five distinct roles:
  - Intake.
  - Integration.
  - Distribution.
  - Delivery.
  - Access.

Let us look at the example of “AllGoods” company. “AllGoods” company is a global company with presence in all the major countries of the world. The company has several branches in each country. Each branch has its own set of databases. The President of “AllGoods” company’s US division has asked his analyst to present the company’s sales per product category (such as “Accessories”, “Clothing”, etc.) per section (such as “Men”, “Women”, “Kids”, and “Infant”) per branch for the first quarter. The challenge here is to collate data spread out over several databases across several cities. In case of “AllGoods” company this task doesn’t pose much of a challenge as the company has a data warehouse depicted in Figure 5.6.

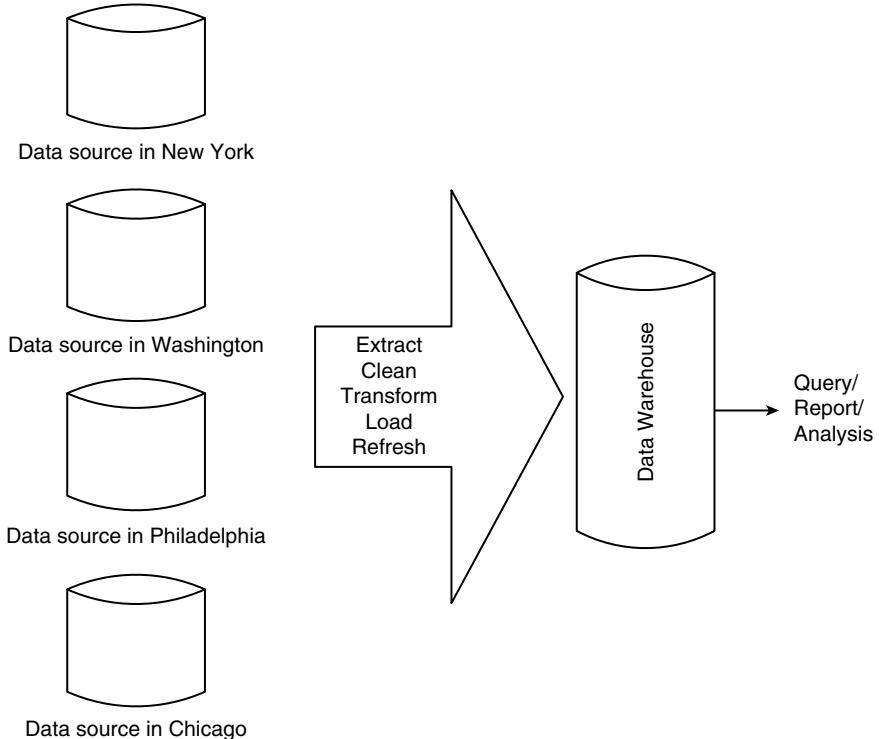
A data warehouse is built by extracting data from multiple heterogeneous and external sources, cleaning to detect errors in the data and rectify them wherever possible, integrating, transforming the data from legacy format to warehouse format, and then loading the data after sorting and summarizing. The data warehouse is also periodically refreshed by propagating the updates from the data sources to the warehouse. Refer to Figure 5.6.

Details on building a data warehouse are covered in Chapter 6.

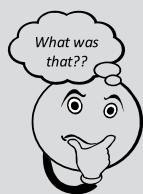
### **Information Services**

The following information related tasks are performed in the information services layer of the BI component framework:

- It is not only the process of producing information; rather, it also involves ensuring that the information produced is aligned with business requirements and can be acted upon to produce value for the company.
- Information is delivered in the form of KPIs, reports, charts, dashboards or scorecards, etc., or in the form of analytics.
- Data mining is a practice used to increase the body of knowledge.
- Applied analytics is generally used to drive action and produce outcomes.



**Figure 5.6** Data warehouse framework of “AllGoods” company.



### Remind Me

- The BI component framework can be divided into 3 layers:
  - The business layer.
  - The administration and operations layer.
  - The implementation layer.
- Business drivers are the impulses that initiate the need to act.
- Business goals are the targets to be achieved in response to the business drivers.
- Business strategies are the planned course of action that will help achieve the set goals.
- Data governance is a technique for controlling data quality.
- Metadata management involves tracking, assessment, and maintenance of metadata.
- A data warehouse is a data store.
- A data warehouse is structured on the dimensional model schema, which is optimized for data retrieval rather than updates.
- Applied analytics is used to drive action and produce outcomes.



### Connect Me (Internet Resources)

- <http://www.tdwi.org/>



### Test Me Exercises

#### Fill me

1. \_\_\_\_\_ is a data store, optimized for data retrieval rather than updates.
2. \_\_\_\_\_ are the impulses/forces that initiate the need to act.
3. \_\_\_\_\_ is a technique for controlling data quality.
4. The \_\_\_\_\_ layer consists of technical components that are required for data capture, transformation and cleaning, converting data into information and finally delivering that information to leverage business goals and produce value for the organization.

5. The expansion for EIS is \_\_\_\_\_.
6. The expansion for DSS is \_\_\_\_\_.
7. Using BI tools, we can generate \_\_\_\_\_, \_\_\_\_\_, \_\_\_\_\_, or \_\_\_\_\_ intelligence.

#### Solution:

1. Data warehouse
2. Business drivers
3. Data governance
4. Implementation
5. Executive Information System
6. Decision Support System
7. Customer, strategic, financial, risk

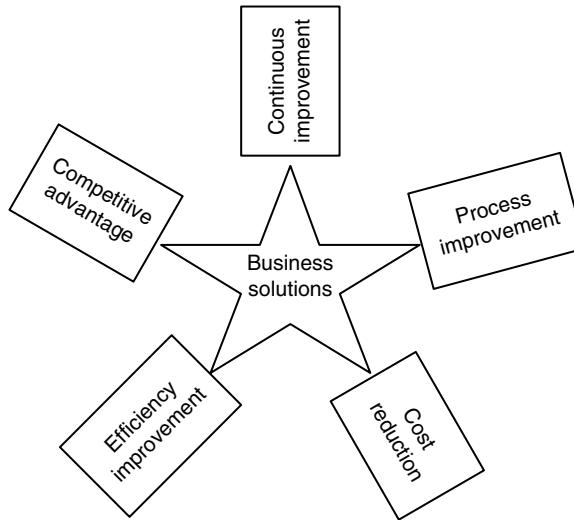
After understanding the BI framework, let us now move towards knowing the users in the BI space and how BI can be leveraged for process and performance improvement.

## 5.2 WHO IS BI FOR?

It is a misnomer to believe that BI is only for managers or the executive class. True, it is used more often by them. But *does that mean that BI can be used only for management and control?* The answer is: NO! Let us try to unravel this myth by looking at a few areas (listed below) where BI has made/is making an impact:

- BI for management.
- Operational BI.
- BI for process improvement.
- BI for performance improvement.
- BI to improve customer experience (QCE – Quality of Customer Experience).

Figure 5.7 depicts business solutions provided by Business Intelligence.



**Figure 5.7** Business solutions provided by Business Intelligence.

### 5.2.1 BI for Management

BI is a very powerful weapon in the hands of managers. With the help of BI front-end tools, they are able to use the information made available to gain business value. Gone are the days when managers had to wait for the end of the quarter to take stock of the situation. And more often than not, it was a tad too late to take action. Now with BI, it is right information at the right time. There is no need to wait till the end of the quarter when the quarter results would indicate whether business is going the profit way or the loss way. BI helps report:

- How sales are in the various regions?
- Whether the project is on budget or is overshooting?
- Whether costs are exceeding budgets?
- Whether customers are dissatisfied with a particular service or product offering?
- What items customers buy the most?
- Whether employees are dissatisfied with the changes in the company's policies?
- Whether imparting technical training to employees before placing them on a project is likely to enhance their productivity?
- What is it that your company is best at?

### 5.2.2 Operational BI

Whoever said BI relied on historical data only, needs to think again. *Does BI help in the daily operations of a company?* Yes it does. *How?* Let us explain with an example. Assume a typical manufacturing-to-shipment scenario. A customer places an order. Before accepting the order, the customer service representative might want to check whether adequate inventory is available. For this he may look at a report generated within an order inventory system or furnished through a BI solution.

*What is the primary difference between BI for management and operational BI?* Operational BI will have to interact either directly with a transaction system or be sourced by a data warehouse that is updated

in near real time several times in a day. This is not to say that BI for management will not use near real time data. It may, but it can also be based on weekly or monthly data.

### 5.2.3 BI for Process Improvement

We have heard often times that BI leads to enhancement in the performance of enterprises. But the question to ask here is: *Does it also contribute to improvement of processes?* The answer again is: “Yes!” Sometimes the process itself can be a bottleneck. Here is an example to explain the same. A retail store was running into the cash flow problem. Goods from the retail store were delivered to the customers on time. So on-time delivery was not a problem. The invoice was sent to the customer after about a week but before 10 days. If the delivery-to-invoice time could be curtailed, chances were that the retail store might not experience the cash flow problem. BI had helped the company to monitor this process and identify the road-block. Using the BI-generated information, the store was able to act to reduce the delivery-to-invoice time to a couple of days and come out victorious.

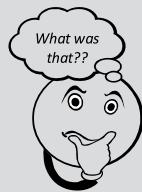
### 5.2.4 BI for Performance Improvement

Let us begin by understanding “*What is the measure for business performance?*” The obvious measures are: revenue, margin, profitability, etc. “*How does BI help to boost revenue?*” Let us try to understand this with an example. BI helps a leather shoe company – XYZ – identify customers who are regular buyers of leather shoes. The company’s representative makes it a point to send across catalogs of other leather accessories from various brands to the identified customers. This is essentially cross-selling. A BI-driven investigation of sales history data can help determine the products that customers are likely to buy together, for example “bread and egg” or “bread and butter” or “butter and jelly”, etc. These are few other examples of cross-selling. Cross-selling is also called “Market Basket Analysis”. “*How does BI help companies maintain a steady cash-flow?*” The answer is: by identifying late-paying customers. “*How does BI help increase profitability?*” A company can use BI tools to understand the customer demographics before deciding to launch a product in a particular region. The launch of the right product will mean better profits as there will be more customers for it.

### 5.2.5 BI to Improve Customer Experience

Businesses are all about providing an enhanced quality of customer experience. Let us take an example. BI can help an airlines identify frequent fliers. The airlines can then come up with schemes such as upgradation from economy to business class based on availability, giving preference to frequent fliers club. A car dealer can use BI to monitor its warranty program and track down the root causes for warranty problems. A telecom company can make use of BI tools to determine the customers who are frequent in sending out text messages and then can target those customers either for cross-sell or up-sell. BI will help all these companies serve the customers better and win the customers satisfaction, loyalty, and advocacy. This will definitely do wonders for the business.

Decision making is not new. It existed even without IT support. But it was more of a gut-feel kind of stuff and less based on facts. Now, IT and businesses have joined hands to make BI a mission-critical endeavor, and one of the essentials for successful businesses.



### *Remind Me*

- BI is not only for senior management/executive class. BI is also for monitoring and managing the day-to-day operations of the organization.
- BI is for process improvement.
- BI is for performance improvement.

- BI is for understanding your customers.
- BI is for recognizing the opportunities to cross-sell and up-sell.
- BI is both an “art” and a “science”.
- BI has to be supported at all levels by both the IT and business executives.



### *Point Me (Books)*

- *Business Intelligence for dummies* – Swain Scheps.
- *Successful Business Intelligence: Secrets to Making Killer BI Applications* by Cindi Howson.



### *Test Me Exercises*

#### **Fill me**

1. For BI to be successful in an organization, both \_\_\_\_\_ and \_\_\_\_\_ executives must come together.
2. Operational BI might be required to interact directly with \_\_\_\_\_ systems.
3. Cross-selling is also called \_\_\_\_\_.

#### **Solution:**

- |  |
|--|
| <ol style="list-style-type: none"> <li>1. IT, business</li> <li>2. Transaction</li> <li>3. Market Basket Analysis</li> </ol> |
|--|

## **5.3 BI USERS**

One can broadly classify BI users into two broad categories:

- Casual users.
- Power users.

Table 5.2 distinguishes casual users from power users.

**Table 5.2** Distinction between casual users and power users

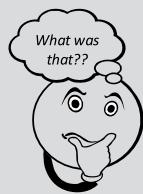
| Type of User                           | Example of Such Users  | Data Access   | Tools                                   | Sources  |
|--|--|---|---|--|
| Casual users/<br>Information consumers | Executives, managers, customers, suppliers, field/operation workers, etc                             | Tailor-made to suit the needs of their respective roles | Pre-defined reports/<br>dashboards      | Data warehouse/data marts                              |
| Power users/<br>Information producers  | SAS, SPSS developers, administrators, business analysts, analytical modelers, IT professionals, etc. | Ad hoc/exploratory                                      | Advanced analytical/<br>authoring tools | Data warehouse/data marts (both internal and external) |

### 5.3.1 Casual Users

These users are the consumers of information in pre-existing reports. They are usually executives, managers, field/operations workers, customers, or suppliers. They might base their decisions/actions on the acquired information. They do not create the reports. They make do with the reports/dashboards tailored to meet the needs of their respective roles and created by power users by sourcing data from data warehouse/data marts.

### 5.3.2 Power Users

These users are the producers of information. They produce information either for their own needs or to satisfy the information needs of others. Developers, administrators, business analysts, analytical modelers (SAS, SPSS developers, etc.), IT professionals, etc. belong to this category. The power users take decisions on issues such as “What information should be placed on the report?”, “What is the best way to present the information?”, “Who should see what information (access rights)?”, “How the information should be distributed (distribution channels)?”, etc. Developers can develop reports, write simple/complex queries, fix/tweak/optimize queries, slice/dice the information, and analyze information/data to gain better insights. They usually use powerful analytical and authoring tools to access data from data warehouses/data marts and other sources both inside and outside the organization.



#### Remind Me

- Power users decide on what information should be made available to other knowledge workers.
- Casual users or information consumers make do with pre-defined and pre-existing reports/dashboards.



### Test Me Exercises

**Fill me**

1. \_\_\_\_\_ users decide “who should see what information (access rights)”.
2. \_\_\_\_\_ users make do with reports created by developers.

**Solution:**

1. Power
2. Information consumers

## 5.4 BUSINESS INTELLIGENCE APPLICATIONS

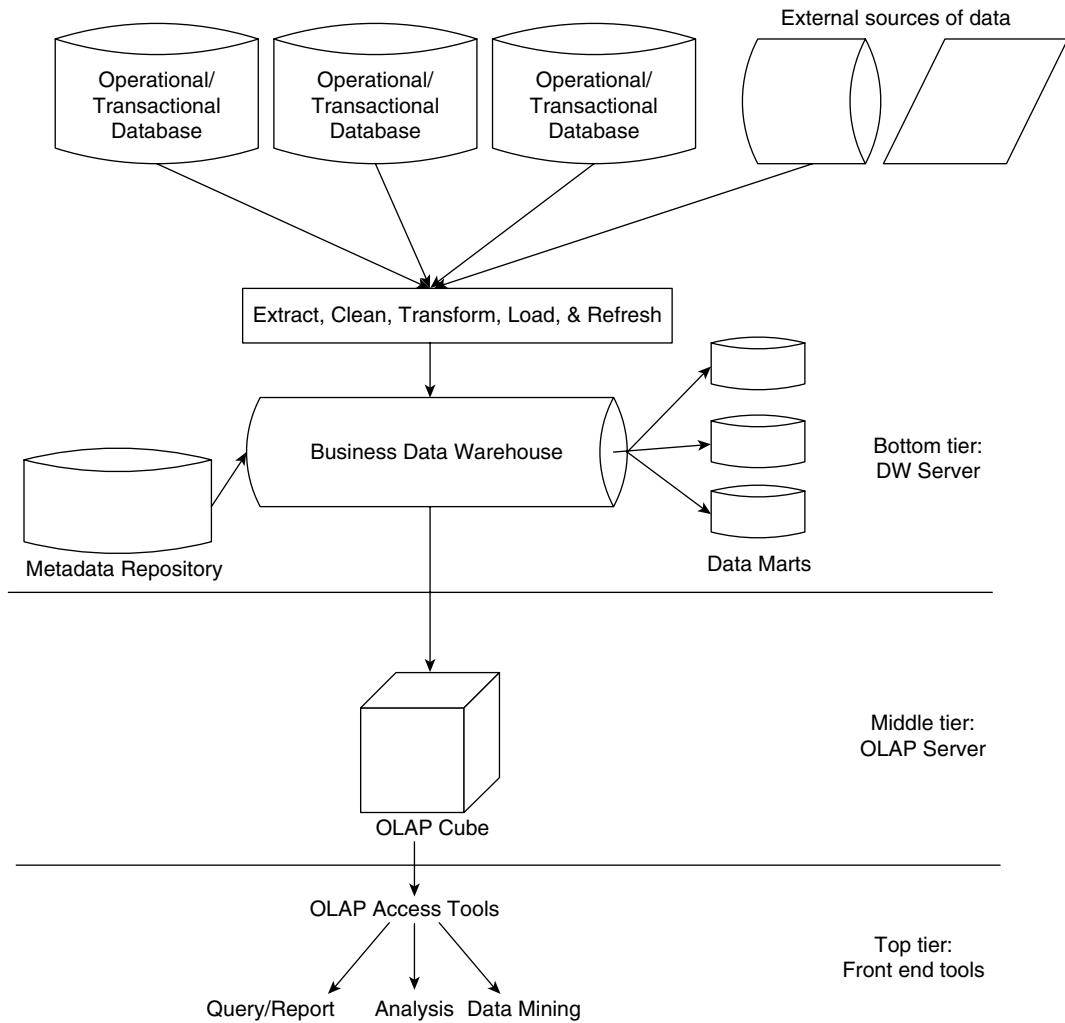
BI applications can be divided into two categories – *technology solutions* and *business solutions*.

- Technology solutions
  - DSS
  - EIS
  - OLAP
  - Managed query and reporting
  - Data mining
- Business solutions
  - Performance analysis
  - Customer analytics
  - Marketplace analysis
  - Productivity analysis
  - Sales channel analysis
  - Behavior analysis
  - Supply chain analytics

### 5.4.1 Technology Solutions

- **DSS** stands for Decision Support System. It supports decision making at the operational and tactical levels. DSS is discussed in Chapter 4.
- **EIS** stands for Executive Information System. It supports decision making at the senior management level. It is a specialized form of DSS. It provides relevant internal and external information to senior executives so that they can evaluate, analyze, compare, and contrast, recognize trends, identify opportunities and problems, etc. The emphasis with EIS has been to provide an easy to use graphical interface with strong reporting capabilities. EIS is discussed in Chapter 4.
- **OLAP** stands for On-Line Analytical Processing. The lifeline of OLAP systems is multidimensional data. OLAP access tools allow the slicing and dicing of data from varied perspectives. Read more about OLAP in Chapter 3.

Let us understand On-Line Analytical Processing as depicted in Figure 5.8. We start at the bottom tier. Let us label it as the data warehouse server layer. This tier houses the enterprise-wide data warehouse



**Figure 5.8** On-Line Analytical Processing.

(DW). This data warehouse is sourced from multiple heterogeneous internal data sources and a few external data sources. There are few tools (back-end tools and utilities) that are used at the layer to cleanse, transform, and load the data into the DW. The DW is also periodically refreshed to propagate the updates from the data sources to the DW. This layer also has the presence of a metadata repository that stores information about the data warehouse and its contents.

Now let us look at the middle tier. Let us label it as the OLAP server layer. This layer usually has a ROLAP (relational OLAP) or a MOLAP (multidimensional OLAP) server. The top tier is the client front-end layer. This layer is adequately supported by query, reporting, and analysis tools.

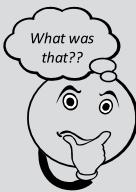
- **Managed query and reporting:** This tool includes predefined standard reports, report wizards, and report designer which are essentially used by developers to create reports. Then there is report builder, essentially used by business users to quickly create a report according to the given report template.

- **Data mining:** Data mining is about unravelling hidden patterns; spotting trends, etc. Anybody who has ever visited the Amazon site would have realized that the moment a person makes a selection, suggestions such as “those who bought this book also bought...” show up. This is possible owing to an analysis of customers buying behavior. Another example is that of market basket analysis. Those who bought bread also bought butter and eggs.

#### 5.4.2 Business Solutions

The business solutions mentioned earlier are described below:

- **Customer analytics:** Customer analytics plays a vital role in predicting the customer's behaviour, his/her buying pattern, etc. Essentially, customer analytics helps capture data about customer's behaviour and enabling businesses to make decisions such as direct marketing or improving relationships with customers (CRM). This analytics is supported by metrics to measure customer loyalty, customer satisfaction, etc.
- **Marketplace analysis:** This analysis helps understand the marketplace better. It is about understanding the customers, the competitors, the products, the changing market dynamics, etc. It is performed to answer questions such as “Whether the launch of product X in region A will be successful?”, “Will the customers be receptive to the launch of product X?”, “Should we discontinue item Z?”, “Where should items A and B be placed on the shop shelves?”, etc.
- **Performance analysis:** This analysis facilitates optimum utilization of employees, finance, resources, etc. Once you do the performance analysis of your employees, you will know about the employees that you will want to retain, the employees whom you want to reward, etc. The performance analysis of your business will provide clear insights into the areas that are a cause for concern and need your immediate attention.
- **Behavior analysis:** This analysis helps predict trends such as purchasing patterns, on-line buying patterns (digital consumers), etc.
- **Supply chain analytics:** This analysis helps optimize the supply chain from planning → manufacturing → sales. Supply chain analytics helps spot trends, identify problems and opportunities in the supply chain functions such as sourcing, inventory management, manufacturing, sales, logistics, etc.
- **Productivity analysis:** In economics, productivity is defined as the ratio of output produced per unit of input. Productivity can be affected by several factors such as workforce efficiency, product quality, process documentation, resources availability, some external influences, etc. This analysis is aimed at enhancing the enterprise profitability. In order to perform productivity analysis, the organization will need to collect data, perform aggregations/summarizations, compare the actual against the estimated or planned, etc. It is based on business metrics that enables the enterprise to increase its profitability. Productivity analysis goes a long way in evaluating the performance of the enterprise.
- **Sales channel analysis:** This analysis will help decide the best channel for reaching out your product/services for use or consumption by the customers/consumers. A good “marketing mix” comprises 4 Ps: Product, Price, Place (distribution), and Promotion. An enterprise on road to achieving profitability will consider all the 4 Ps. It will decide on “what product to produce”, “what services to offer”, etc. It will decide on the price – “when to increase/decrease the price of the product or service”, etc. It will decide on how to distribute/reach out to the customers with its products or service. It will decide on how to offer the product/service – “whether in a physical store or through an online virtual store”, etc. It will also look at promotion – “personal selling”, “advertising”, etc. The sales channel analysis provides insights that help decide which sales channel are the most profitable and which sales channel should be done away with.



### *Remind Me*

- BI applications can be divided into technology solutions and business solutions.
- Examples of technology solutions: DSS, EIS, managed query and reporting, data mining, etc.
- Examples of business solutions: customer analytics, marketplace analysis, behaviour analysis, sales channel analytics, productivity analysis, etc.
- Multidimensional data is the lifeline of OLAP systems.
- EIS is a specialized version of DSS, meant to assist senior executives in their decision making.
- Productivity analysis is performed to help the enterprise increase its profitability.
- Marketplace analysis helps understand the customers, the competitors and the market dynamics, etc.



### *Connect Me (Internet Resources)*

- <http://www.tdwi.org>
- [http://www.infosys.com/offerings/industries/high-technology/white-papers/  
Documents/supply-chain-analytics.pdf](http://www.infosys.com/offerings/industries/high-technology/white-papers/Documents/supply-chain-analytics.pdf)
- [http://en.wikipedia.org/wiki/Customer\\_analytics](http://en.wikipedia.org/wiki/Customer_analytics)



### *Test Me Exercises*

#### **Match me**

| <i>Column A</i>                                  | <i>Column B</i>        |
|--|------------------------|
| Essentially for senior management                | DSS                    |
| Helps in operational tasks                       | EIS                    |
| Helps optimize supply chain functions            | Customer analytics     |
| Is the ratio of output produced per unit of work | Supply chain analytics |
| Helps in improving relationship with customers   | Productivity           |

**Solution:**

| <i>Column A</i>                                  | <i>Column B</i>        |
|--|------------------------|
| Essentially for senior management                | EIS                    |
| Helps in operational tasks                       | DSS                    |
| Helps optimize supply chain functions            | Supply chain analytics |
| Is the ratio of output produced per unit of work | Productivity           |
| Helps in improving relationship with customers   | Customer analytics     |

## 5.5 BI ROLES AND RESPONSIBILITIES

BI roles can be broadly classified into two categories – *program roles* and *project roles* – as listed in Table 5.3. For a BI project to succeed, one requires executive level sponsorship. It should be backed by the senior leadership of the organization. This level of sponsorship will garner enough support from all the related project teams and will also be responsible to generate/allocate the requisite funds.

The program team lays down the strategy of how the BI project will execute. The program team members are responsible for coordination and integration. The project team executes the program team's strategy.

### 5.5.1 BI Program Team Roles

#### *BI Program Manager*

The BI program manager is responsible for several projects. He etches out the BI strategy. The following are the key responsibilities of the BI program manager:

**Table 5.3** Two categories of BI roles

| <i>Program Roles</i>   | <i>Project Roles</i>          |
|------------------------|-------------------------------|
|                        | Business Manager              |
| BI Program Manager     | BI Business Specialist        |
| BI Data Architect      | BI Project Manager            |
| BI ETL Architect       | Business Requirements Analyst |
| BI Technical Architect | Decision Support Analyst      |
| Metadata Manager       | BI Designer                   |
| BI Administrator       | ETL Specialist                |
|                        | Data Administrator            |

- The program manager has a clear understanding of the organization's strategic objective and aligns the objectives of the BI project with it.
- He explicitly defines metrics to measure and monitor the progress on each objective/goal.
- He plans and budgets the projects, distributes tasks, allocates/de-allocates resources, and follows up on the progress of the project.
- He identifies measures of success.
- He measures success/ROI.

### ***BI Data Architect***

The BI data architect

- Owns accountability for the enterprise's data.
- Ensures proper definition, storage, distribution, replication, archiving, and management of data.

He not only optimizes data for current usage but also takes into account the future needs in both design and content.

### ***BI ETL Architect***

For building a data warehouse, data needs to be extracted from multiple disparate sources. An ETL architect has the responsibility of

- Determining the optimal approach for obtaining data from different operational source systems/platforms and loading it into the data warehouse.
- Training project ETL specialists on data acquisition, transformation, and loading.

### ***BI Technical Architect***

The chief responsibilities of a BI technical architect are

- Interface with operations staff.
- Interface with technical staff.
- Interface with DBA staff.
- Evaluate and select BI tools (ETL, analysis and reporting tools, etc).
- Assess current technical architecture.
- Estimate system capacity to meet near- and long-term processing requirements.
- Define BI strategy/processes for
  - Definition of technical architecture for BI environment.
  - Hardware requirements.
  - DBMS requirements.
  - Middleware requirements.
  - Network requirements.
  - Server configurations.
  - Client configurations.
- Define the strategy and process for data back-up and recovery, and disaster recovery.

### ***Metadata Manager***

Metadata is "data about data". Metadata helps understand data. The metadata manager keeps track of the following activities:

- Structure of data (technical metadata).
- Level of details of data.
- When was the ETL job performed? (ETL audit)
- When was the data warehouse updated?
- Who accessed the data and when? (Application metadata)
- What is the frequency of access? (Application metadata)

### ***BI Administrator***

The BI administrator has the following set of responsibilities:

- Design and architect the entire BI environment.
- Architect the metadata layer.
- Monitor the health/progress of the entire BI environment.
- Manage the security of the BI environment.
- Monitor all scheduled jobs such as ETL jobs, scheduled reports for business users, etc.
- Monitor and tune the performance of the entire BI environment.
- Maintain the version control of all objects in the BI environment.

### **5.5.2 BI Project Team Roles**

#### ***Business Manager***

The business manager plays the role of monitoring the project team from the user group perspective. The roles performed by him/her can be listed as follows:

- To act as the sponsor representing the user group (the target customer of the project).
- Monitoring the activities of project team.
- Addressing the business issues identified by the project manager.

#### ***BI Business Specialist***

Each project team requires at-least one FTE (Full Time Employee) resource having expertise in the business area of focus. BI business specialist helps in identifying the suitable data usage and structure for the business functional area. The knowledge of BI specialist ensures that

- Information is identified correctly at the required level itself.
- All the modes of accessing and analyzing data are enabled.

The BI business specialist is also the lead in data stewardship and quality programs.

#### ***BI Project Manager***

The BI project manager takes up the responsibility of leading the project and ensuring delivery of all project needs. Also, the project manager translates business needs into technical terms and ensures adherence to all business standards and BI processes. It is the BI project manager's responsibility to

- Understand existing business processes.
- Analyze existing decision support and executive information systems to understand their functionality.
- Understand subject matter.

- Anticipate and judge what users will/may want.
- Manage expectations of the project.
- Scope an increment.
- Develop project plan for an increment/project.
- Motivate team members.
- Evaluate team members.
- Assess risk.
- Manage expectations.
- Understand information architecture.
- Understand technical architecture.
- Manage project.
- Coordinate with the program manager to ensure standards.
- Coordinate with other project managers.
- Understand organizational architecture.
- Implement warehousing specific standards.
- Communicate with all other team members.

### ***Business Requirements Analyst***

The business requirements analyst maintains a synchronized handshake between the end-users and the BI project team, and performs requirements gathering. It is the business requirements analyst's responsibility to

- Question the end-users to determine requirements keeping in view all the aspects like data, reports, analyses, metadata, performance, etc.
- Work with architects to transform requirements into technical specifications.
- Document requirements.
- Help identify and assess potential data sources.
- Recommend appropriate scope of requirements and priorities.
- Validate that BI meets requirements and service-level agreements.
- Coordinate prototype reviews.
- Gather prototype feedback.

### ***Decision Support Analyst***

A decision support (DS) analyst is an individual who helps encounter issues and supports DSS. The DS analyst is an expert on issues surrounding business objectives, questions, and problems, and in obtaining and presenting the required data to address the same issues.

The DS analyst creates data results using several techniques and tools from firing basic queries to multidimensional analyses and data mining, making new relations and/or derivations as may seem fit. He/she also makes an attempt at extracting the maximum amount of valid information. It is the DS analyst's responsibility to

- Educate users on warehousing capabilities.
- Analyze business information requirements.
- Design training infrastructure.
- Discover business transformation rules.
- Work with production data to validate business requirements.

- Map the requirements to suit the business functional model.
- Create state transformation models.
- Discover dimension hierarchies.
- Validate hierarchies with production data.
- Define business rules for state detection.
- Classify business users by type.
- Define and have an agreement with business users, and base it on service-level agreement.
- Develop security rules/standards.
- Create data results through several techniques and tools.
- Develop necessary reports.
- Develop decision support and EIS applications.
- Develop Internet and intranet delivery applications.
- Convert existing reporting applications to the environment.
- Develop new periodic report applications.
- Develop training materials.
- Write users' guide.
- Plan acceptance test.
- Execute acceptance test plan.
- Train BI users.
- Implement support plan.
- Assist users in finding the right information.
- Interface with process teams regarding business process reengineering.

### ***BI Designer***

The BI designer aims at designing the data structure for optimal access, performance, and integration. Designing essentially plays a key role while building new data sets as required for supporting the business needs.

The designer must maintain the balance of needs in both design and content, and must always keep both the current and future demands in mind to judge and foresee. The works of the BI designer and BI data architect are closely knit, which ensures compliance of all standards, consistency of the project and other information architecture deliverables. The responsibilities of the BI designer are

- Create a subject area model.
- Create or review the business enterprise model.
- Interpret requirements.
- Create a logical staging area model.
- Create a structural staging area model.
- Create a physical staging area model.
- Create logical distribution model.
- Create a structural distribution model.
- Create a physical distribution model.
- Create a logical relational model.
- Create a structural relational model.
- Create a physical relational model.
- Create a logical dimensional model.

- Create a structural dimensional model.
- Create a physical dimensional model.
- Validate models with production data.
- Develop processes to maintain and capture metadata from all BI components.

### ***ETL Specialist***

Before implementing any extraction, a proper and apt technique has to be finalized for the process. This is where the ETL specialist comes in to determine and implement the best technique for extracting. Collaboration between the ETL specialist and the ETL architect ensures compliance of all standards and deliverables which enhances the consistency and longevity of infrastructure plans. It is the ETL specialist's responsibility to

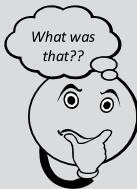
- Understand both the source and the target BI systems.
- Identify data sources.
- Assess data sources.
- Create source/target mappings.
- Apply business rules as transformations.
- Implement navigation methods/applications.
- Design and specify the data detection and extraction processes to be used.
- Design and develop transformation code/logic/programs for environment.
- Design and develop data transport and population processes for environment.
- Build and perform unit test data transformation processes for environment.
- Build and perform unit test source data transport and population processes for environment.
- Work with production data to handle and enhance data conditions and quality.
- Design data cleanup processes.
- Adapt ETL processes to facilitate changes in source systems and new business user requirements.
- Define and capture metadata and rules associated with ETL processes.
- Coordinate with the program-level ETL architect.

### ***Database Administrator***

The database administrator (DBA) is like a guardian of the data and data warehouse environment. He/she keeps check on physical data that is appended to the existing BI environment under current project cycle. The DBA works closely with the metadata manager to measure and speculate on the alterations done to BI and efficiency of the processes of the BI environment. It is the DBA's responsibility to

- Design, implement, and tune database schemas.
- Conduct regular performance testing and tuning.
- Manage storage space and memory.
- Conduct capacity planning.
- Create and optimize physical tables and partitions.
- Implement all models, including indexing strategies and aggregation.
- Manage user accounts and access privileges.
- Implement vendor software patches.
- Analyze usage patterns and downtime.
- Administer tables, triggers, etc.

- Log technical action reports.
- Document configuration and integration with applications and network resources.
- Maintain backup and recovery documentation.



### *Remind Me*

- BI roles can be broadly divided into two categories: program roles and project roles.
- The database administrator is like a guardian of the data and data warehouse environment.
- The ETL specialist determines and implements the best technique for data extraction.
- The BI designer aims at designing the data structure for optimal access, performance, and integration.

- The business requirements analyst maintains a synchronized handshake between the end-users and the BI project team and performs requirements gathering.
- The BI data architect owns the accountability for the organization's data.
- For a BI project to succeed, executive-level sponsorship is necessary.
- A BI program manager is responsible for several projects.



### *Connect Me (Internet Resources)*

- <http://www.tdwi.org>
- Beye Network

## **5.6 BEST PRACTICES IN BI/DW**

Following is a list of best practices adapted from an article in TDWI's FlashPoint e-newsletter of April 10, 2003:

- **Practice “user first” design:** We can gain value from BI/DW products only when they are used by the end-users for whom they have been designed. The first thing to cash in on will be an impeccable “user experience”. There ought to be “no compromises here!” BI solutions should be designed to fit seamlessly into the scheme of day-to-day business activities. The design has to help the end-users to be more productive. It has to improve the usability of the system and thereby greater acceptance by the users. In other words, the “user first” design has to change the design aim from “users should use it” to “users want to use it”.
- **Create new value:** BI projects are costly endeavors. Therefore, a BI project cannot be just left at extracting and storing the data in the enterprises data warehouse. It should lead to business

value addition. The key point here is that businesses should benefit by investing in BI/DW. A BI solution should be able to impact the business in a positive way. It has to have this positive impact on all strata of the organization. Besides improving the day-to-day business in terms of improved productivity, reduced operational costs, etc., it should also assist the organization in realizing their strategic objectives in a timely manner.

- **Attend to human impacts:** BI projects are known to be less technology-centric and more people-centric. BI supports decision making at the strategic, tactical, and operational levels. It starts with those who define and articulate strategies to those who help in carrying out day-to-day operations. For its successful implementation, it needs a persistent tight handshake between the two communities: business users and IT personnel. It requires new skills, an adaptive mindset, etc. It promises a new experience.
- **Focus on information and analytics:** The transformation of raw data to business benefits through BI may be depicted as

$$\text{Data} \rightarrow \text{Information} \rightarrow \text{Knowledge} \rightarrow \text{Insights} \rightarrow \text{Impact}$$

Raw data over time needs to be transformed into meaningful information. The information needs to be mined to provide adequate knowledge. The knowledge gained will lead to clearer insights into the running of the business. This will help in making correct and accurate decisions that can lead to huge impacts. Success of a BI/DW initiative relies heavily on analytics (obtaining an optimal and realistic decision based on existing data) and reporting (drill down, drill through, roll-ups, etc.). The basic querying, reporting, OLAP, and alert tools can provide answers to questions such as "What happened?", "Where does the problem lie?", "How many have been impacted?", "How often does it happen?", etc. Business analytics can provide answer to questions such as "Why is this happening?", "What will happen next (i.e., predicting)?", "What is the best that can happen (that is, optimizing)?", etc.

- **Practice active data stewardship:** Information is created out of data. Good data begets good meaningful information. And what is "good data". Good data is "data of quality". Good data is created by implementing sound data management and metadata management techniques. It is a result of a diligent act of data cleansing and purifying. Data governance is a quality regime that can be religiously implemented to ensure the accuracy, correctness, and consistency of data. Practice of active data stewardship rings in accountability and thereby enhances the trust and confidence on data.
- **Manage BI as a long-term investment:** BI projects should be looked at as something that yields good returns in the long term. Although, small wins are possible and often times experienced, it is a sustainable BI program that should be the focus. It is not a roller-coaster ride that one should be concerned with, but rather consistent supply of useful information, good analytics, and visible measurable impact.
- **Reach out with BI/DW solutions:** Extend the reach of BI/DW as far as possible. Each business function/process that uses BI is a potential candidate to provide business value. BI being a people-centric program, the same is true in the case of every business user or IT personnel who uses BI. BI should not be confined only to the boardroom. Organizations have benefitted and will benefit by implementing BI at all levels.
- **Make BI a business initiative:** BI projects are usually expensive. Businesses have to invest in BI. Besides requiring an executive-level sponsorship, BI relies heavily on accountability. There has to be accountability for data management, metadata management, data stewardship, etc. In

BI space, it is business first and technology second. Technology supports BI; it cannot create BI. The IT folks have to be brought around to appreciate the positive impact of BI. Focus should be on leveraging technology to support BI.

- **Measure results:** Businesses that have been able to reap benefits from data analytics have religiously employed effective metrics. They have unfailingly been measuring the TCO (total cost of ownership), TVO (total value of ownership), ROI (return on investment), ROA (return on asset), etc. and used the results to further enrich their business gains. Better metrics mean a better handle on the business and better management and control.
- **Attend to strategic positioning:** The need for BI was born out of a few necessities. BI was brought in to endure changing markets, changing workforce, changing technologies, changing regulations, etc. These external forces drove businesses to adopt BI. BI/DW programs were devised to provide a foundation to align information technology to business strategies. It is very important to understand the need for BI in an organization and leverage it for impactful business benefits.



### *Remind Me*

- BI is for all and not just for the senior management.
- Focus on leveraging technology to support BI initiatives.
- BI is for the long run. Make it into a business initiative.
- Good data is created by sound data management and metadata management.
- In BI space, it is business first and technology second.
- Better metrics mean a better handle on the business and better management and control.



### *Connect Me (Internet Resources)*

- <http://www.tdwi.org/>
- TDWI's FlashPoint e-newsletter of April 10, 2003.



### *Test Me Exercises*

#### **Fill me**

1. Data → Information → \_\_\_\_\_ → Insights  
→ \_\_\_\_\_.
2. In BI space it is \_\_\_\_\_ first and \_\_\_\_\_ next.
3. TCO is \_\_\_\_\_.
4. ROA is \_\_\_\_\_.
5. TVO is \_\_\_\_\_.
6. Success of BI relies on good \_\_\_\_\_ and \_\_\_\_\_.

**Solution:**

1. Knowledge, impacts
2. Business, technology
3. Total cost of ownership

4. Return on assets
5. Total value of ownership
6. Analytics and reporting

## 5.7 THE COMPLETE BI PROFESSIONAL

Let us look at what skills are required in a BI professional. We present below a list of disciplines associated with BI and we recommend that a budding professional should have a broad knowledge of all the disciplines, i.e. a view at the “big picture”, and should aim for specialization in a few of them. Here’s our list:

- Data modeling.
- Metadata management.
- Data quality.
- Data governance.
- Master data management.
- Data integration.
- Data warehousing.
- Content management.
- Enterprise information management.
- Business intelligence.
- Business analytics.
- Performance management.

For ease of understanding, we have grouped together disciplines which complement each other in Table 5.4.

**Table 5.4** Grouping of disciplines that complement each other

|   |  |
|---|--|
| To identify, understand, and discern data       | <b>Data modeling and metadata management:</b> Data modeling entails metadata management. Metadata, previously called Data Dictionary, is a collection of definitions and relationships that describe the information stored. It is “data about data”. The storage and access of metadata information is as important today as it was earlier.              |
| To govern data quality and information delivery | <b>Data quality and data governance:</b> Good decisions are based on quality data. Data governance is a quality regime that includes ensuring accuracy, consistency, completeness, and accountability of data.   |
| To consolidate data from disparate data sources | <b>Master data management, data integration, and data warehousing:</b> Consolidation and integration of data extracted from varied operational data sources spread across companies and across geographies and more likely existing in varied formats, transformation of data, and the eventual loading into a common repository is called data warehouse. |

(Continued)

**Table 5.4** (Continued)

|   |   |
|---|---|
| To manage the supply and demand of data   | <b>Content management and enterprise information management:</b> They are the means to manage the information that is available (supply), the information that is required (in demand), and the gap that exists between the demand and supply.  |
| To measure, monitor and manage performance: “You cannot manage what you cannot measure, and you cannot measure what you cannot define.” | <b>Business intelligence, business analytics, and performance management:</b> Good metrics help in measuring and managing performance. KPIs (Key Performance Indicators) are metrics or measures that help an organization measure its performance and its subsequent progress towards its strategic objectives. These KPIs can either be quantifiable or non-quantifiable. A few examples of good KPIs are: <ul style="list-style-type: none"> <li>• Market share</li> <li>• Market growth</li> <li>• Customer satisfaction</li> <li>• Customer churn rate</li> <li>• Customer retention</li> <li>• Customer profitability</li> </ul>                      |
| To gain insight and foresight   | <b>Data mining and predictive analytics:</b> Predictive analytics is about analyzing the present and historical data to make predictions for the future. It is about unravelling hidden patterns and trends. One of the very famous examples of predictive modeling is <i>Customer Credit Scoring</i> . Here, the scoring models take into consideration the customer's credit history, his loan application(s), other details about the customer to rank-order individuals by their likelihood of making their future credit card payments on time. Another example is spotting the “cross-sell” opportunities by analysing the customers buying patterns. |



### Remind Me

- Metadata is “data about data”.
- Data governance is quality control mechanisms employed for accessing, modifying, improving, monitoring, managing, assessing, protecting, using, and maintaining corporate data.
- KPIs can be quantifiable or non-quantifiable.
- Good metrics assist in measuring the performance of the organization.
- Predictive analytics is about predicting the future based on the analysis of current and historical data.
- Enterprise information management helps manage the supply and demand of data.



### Connect Me (Internet Resources)

- <http://tdwi.org/articles/2011/01/05/complete-bi-professional.aspx>



### Test Me Exercises

#### Fill me

1. \_\_\_\_\_ are measures that help an organization measure its performance.
2. \_\_\_\_\_ is a quality regime to ensure the accuracy, consistency, and completeness of data.

3. \_\_\_\_\_ helps unravel hidden patterns and spot trends in historical data.

#### Solution:

1. KPIs
2. Data governance
3. Data mining

## 5.8 POPULAR BI TOOLS

Some of the popular vendors for BI tools are mentioned below as well as depicted in Figure 5.9.

### RDBMS

- Netezza 4.6
- NCR Teradata 13
- Sybase IQ (DWH & BI RDBMS)
- Oracle 11g (Oracle 11g Release 2)
- Microsoft SQL Server 2008
- DB2 Connector 9.7

### ETL Tools

- Informatica 9
- IBM's Data Stage 8.5
- Ab Initio 3.0.2
- Microsoft SQL Server Integration Services SSIS 2008

### Analysis Tools

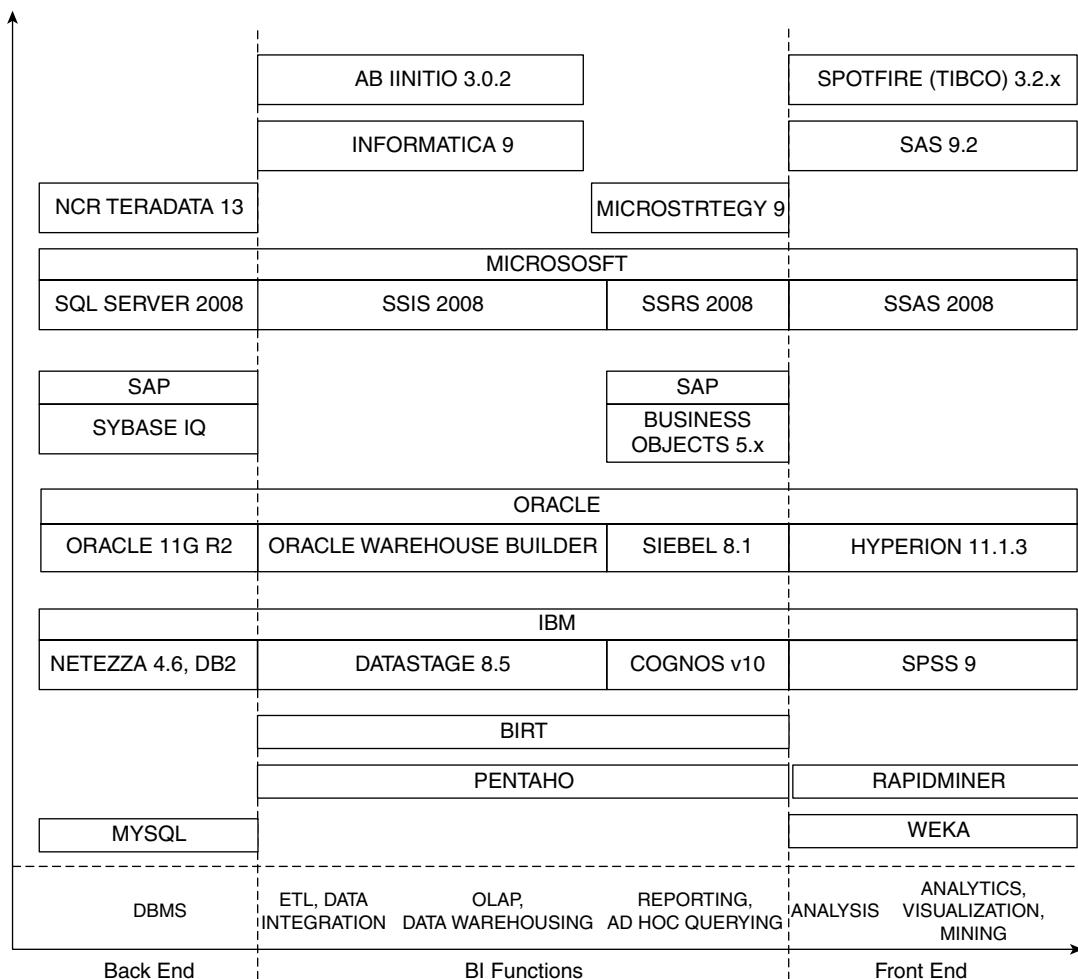
- IBM's SPSS 9
- SAS 9.2
- Microsoft SQL Server Analysis Services SSAS 2008
- Spotfire(tibco) 3.2.x
- Oracle's Hyperion 11.1.3
- Oracle's Essbase

### ***Reporting/Ad Hoc Querying Tools/Visualization***

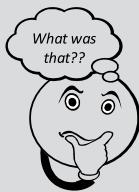
- Microstrategy 9
- SAP's Business Objects 5.x
- IBMS's Cognos V10
- Microsoft SQL Server Reporting Services SSRS 2008
- Siebel Answers (7.5.3) Siebel 8.1(2008)

### ***Open Source Tools***

|   |  |
|---|--|
| <b>RDBMS</b>  | MySQL, Firebird  |
| <b>ETL tools</b>                                      | Pentaho Data Integration (formerly called Kettle), SpagoBI |
| <b>Analysis tools</b>                                 | Weka, RapidMiner, SpagoBI                                  |
| <b>Reporting/Ad hoc querying tools/ Visualization</b> | Pentaho, BIRT, Actuate, Jaspersoft                         |



**Figure 5.9** Popular BI tools.



## *Remind Me*

- *Popular RDBMS:* Oracle 11g, Sybase, MS Sql Server 2008, Netezza, MySQL, etc.
- *Popular ETL tools:* Ab initio, Informatica, SSIS 2008, Data Stage, etc.
- *Popular Reporting tools:* Cognos, Microstrategy, Business Objects, SSRS 2008, etc.
- *Popular Analysis tools:* SPSS, Weka, RapidMiner, SSAS 2008, SAS, Spotfire, etc.



## *Connect Me (Internet Resources)*

- <http://www.squidoo.com/osbi>



## *Test Me Exercises*

### **Fill me**

1. \_\_\_\_\_ and \_\_\_\_\_ are open source tools for analysis.
2. \_\_\_\_\_ is an open source tool for ETL/data integration.
3. \_\_\_\_\_ is an open source tool for reporting.
4. \_\_\_\_\_ is an open source RDBMS.
5. \_\_\_\_\_ is IBM's tool for ETL/data integration.
6. \_\_\_\_\_ is IBM's tool for reporting.
7. Netezza is a product of \_\_\_\_\_ company.

### **Solution:**

1. Weka and RapidMiner
2. Pentaho Data Integration tool formerly called "Kettle"
3. Pentaho
4. MySQL
5. DataStage
6. Cognos
7. IBM

## **UNSOLVED EXERCISES**

1. Describe the business intelligence framework.
2. Assume you are a project manager who has been sent to collect business requirements for a retail chain. Give a few examples of business requirements that you would have collected.

3. Mention a few BI tools in each of the following categories:
  - a. ETL
  - b. Databases that can support data warehouse
  - c. Reporting
  - d. Business data analytics
4. What is metadata? Explain giving examples.
5. Why is maintaining the quality of data important?
6. How can BI be used to enhance customer experience? Explain with an example.
7. How can BI lead to performance enhancement. Explain with an example.
8. What do you understand by the term “business value”?
9. What is “Total Cost of Ownership (TCO)”?
10. Explain your understanding of the data warehouse.
11. Explain the various components of BI architecture.
12. Give examples to explain the various types of metadata: business metadata, application metadata, technical metadata, and process metadata.
13. What type of metadata is stored in the structure given in Table 5.5?

**Table 5.5** Unsolved Exercise 13

| Column Name      | Data Type and Length | Constraints |
|------------------|----------------------|-------------|
| ProjectID        | Integer              | Primary Key |
| ProjectName      | Varchar(35)          | Unique      |
| NoofPersons      | Integer              |             |
| ProjectManager   | Varchar(50)          | Not Null    |
| ProjectClient    | Varchar(35)          |             |
| ProjectLocation  | Varchar(35)          |             |
| ProjectStartDate | Date                 | Not Null    |
| ProjectEndDate   | Date                 | Not Null    |

14. Picture this scenario: There is an online application to buy and sell used cars. The website owner would like to have answers to the following :
  - a. How many visitors visit the website in a day?
  - b. What time of the day there is maximum traffic on the website?
  - c. What time of the day minimum hits happen on the website?
  - d. Do the visitors directly come in through the landing page?

What type of metadata is presented in the scenario above?
15. Picture this scenario: An enterprise has an enterprise-wide data warehouse. The data architect has the responsibility of maintaining the data warehouse. The data warehouse is periodically updated asynchronously. The data architect keeps track of the ETL process – When was it done? Was it an incremental update to the data warehouse?, etc. What according to you is the type of metadata that the data architect is maintaining?

- 16.** What is cross-sell? Explain with an example.
- 17.** What is up-sell? Explain with an example.
- 18.** Explain the following terms with an example each:
  - a.** Customer analytics
  - b.** Productivity analysis
  - c.** Supply chain analysis
- 19.** Provide your understanding of master data management.
- 20.** Differentiate between casual users and power users.
- 21.** State two BI applications.
- 22.** Give two key functionalities of the business requirements analyst.
- 23.** Explain three best practices to be followed in the BI space.
- 24.** Mention one open source tool for each of the following categories:
  - a.** ETL
  - b.** RDBMS
  - c.** Reporting
  - d.** Analysis
- 25.** What skills should a BI professional possess?

# 6



## Basics of Data Integration

---

### BRIEF CONTENTS

|   |   |
|---|---|
| What's in Store                                       | What is Data Integration?               |
| Need for Data Warehouse                               | Data Integration Technologies           |
| Definition of Data Warehouse                          | Data Quality                            |
| What is a Data Mart?                                  | Unsolved Exercises                      |
| What is Then an ODS?                                  | Data Profiling                          |
| Ralph Kimball's Approach vs.<br>W.H. Inmon's Approach | Summary                                 |
| Goals of a Data Warehouse                             | A Case Study from the Healthcare Domain |
| What Constitutes a Data Warehouse?                    | Solved Exercises                        |
| Extract, Transform, Load                              | Unsolved Exercises                      |

---

### WHAT'S IN STORE

We assume that you are familiar with the basics of RDBMS concepts and associated terminologies. We hope that the previous chapters have given you the necessary start in BI. This chapter deals with the general concepts of data integration with respect to data warehousing. It will familiarize you with the concept of ETL (Extract, Transform, Load) in the context of data warehousing, and the importance of data profiling and quality.

We have taken up a sample data set as a case study that will help you relate to the concepts being discussed in this chapter and gain a better understanding of the subject.

We recommend you to attempt the “Test Me” and “Challenge Me” exercises at the end of this chapter to re-enforce the concepts learnt in this chapter. A few books and on-line references have also been listed at the end. They may be referred to for additional reading.

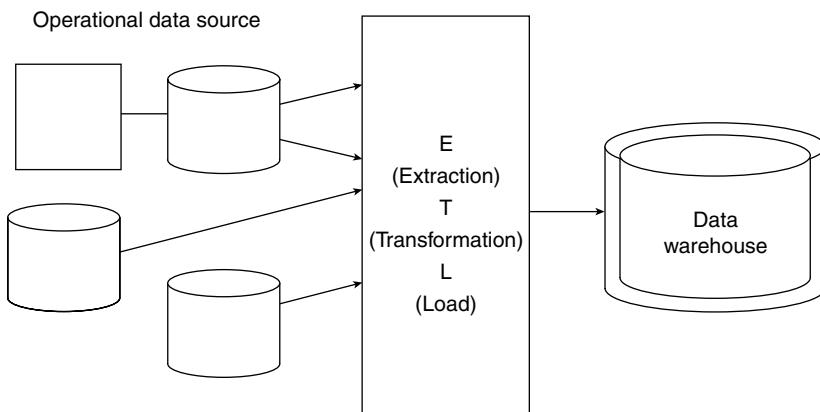
## 6.1 NEED FOR DATA WAREHOUSE

### Picture this scenario...

The Darbury Institute of Information Technology (DIIT) is an engineering institution that conducts engineering courses in Information Technology (IT), Computer Science (CS), System Engineering (SE), Information Science (IS), etc. Each department (IT, CS, SE, IS, etc.) has an automated library that meticulously handles library transactions and has good learning content in the form of books, CD/DVDs, magazines, journals, several online references, etc. DIIT is also looking at expansion and is likely to have its branches in all major cities.

The only downside of the library data is that it is stored differently by different departments. One department stores it in MS Excel spreadsheets, another stores it in MS Access database, and yet another department maintains a .CSV (Comma Separated Values) file. The DIIT administration is in need of a report that indicates the annual spending on library purchases. The report should further drill down to the spending by each department by category (books, CDs/DVDs, magazines, journals, etc.). However, preparing such a report is not easy because of different data formats used by different departments. Prof. Frank (an expert on database technology) was called upon to suggest a possible solution to the problem at hand. He feels it would be better to start archiving the data in a data warehouse/data mart. The arguments put forth by him in favor of a library data warehouse are

- Data from several heterogeneous data sources (MS Excel spreadsheets, MS Access database, .CSV file, etc.) can be extracted and brought together in a data warehouse as depicted in Figure 6.1.
- Even when DIIT expands into several branches in multiple cities, it still can have one data warehouse to support the information needs of the institution.
- Data anomalies can be corrected through an ETL package.
- Missing or incomplete records can be detected and duly corrected.
- Uniformity can be maintained over each attribute of a table.
- Data can be conveniently retrieved for analysis and generating reports (like the report on spending requested above).
- Fact-based decision making can be easily supported by a data warehouse.
- Ad hoc queries can be easily supported.



**Figure 6.1** Data from several heterogeneous data sources extracted and loaded in a data warehouse.

The need for the data warehouse is now clear. In a general sense, it is a system which can conveniently archive data. This data can then be analyzed to make business decisions and predict trends. So, data warehousing is quite important for organizations. However, there still remains one question unanswered: Should all organizations go for a data warehouse, and what is an opportune time for an organization to go for data warehousing?

Let us take the example of a fictitious company by the name “AllFinances”. The 40-year-old company is in the business of banking, finance, and capital markets. It has grown from 15 employees to 1,50,000 employees. The company has seen a stupendous rise in their use of enterprise applications – from 1 enterprise application two decades back to 250 plus such applications today. All applications are not on the same technology. Few applications are on Mainframe, a few on Dot Net, a few on AS/400, and yet a few on Java platform. Most of the applications have their independent databases which are also varied. A few have Oracle as the backend, and a few others have SQL Server 2008 or DB2 as the backend. We have a few others on MySQL. The organization has felt the need for homogeneity. It realizes the importance of integrating data from various enterprise applications existing in silos. The organization appreciates “single version of truth”. So, “AllFinances” very wisely has chosen to go in for data warehousing.

We list below a few issues/concerns of data usage, which often prompt organizations to go for data warehousing. If your organization has also faced one or more of these issues, then it is ready for a data warehouse:

- **Lack of information sharing:** Information though available is not being shared between the various divisions or departments. In our example of DIIT, assume there is a subject, “Fundamentals of software engineering” which is being offered by IT and IS departments. The students of IT and IS departments should be able to access the learning contents on this subject from the libraries of both the departments. This is possible only if their libraries share the information. In yet another example from the retail industry, assume there are two departments (one selling leather jackets and the other selling leather accessories) which have the same customers. Also assume that these departments don’t disclose their information to each other. Both the departments virtually elicit the same information from the customers. What could it mean for the business? Cross-selling opportunities in favor of the customers cannot be realized. And the quite obvious result: the frustrated customer and loss of improved customer comprehensibility.

On the flip side, let us look at a couple of benefits that can accrue from information sharing. One, it can provide cross-selling opportunities which in turn can lead to better customer relationship management. Example: You have just booked yourself on a flight from Delhi to Amsterdam. Within a matter of hours, you receive a call from a cab service, “CabHelp”, of Amsterdam, asking you if you would like to avail their services. In about the same time, “HotelLuxury” places a call to you offering you to stay with them at discounted rates. A clear example of cross-sell opportunities made possible because of your itinerary details being shared with “CabHelp” and “HotelLuxury”.

- **Lack of information credibility:** Picture a board meeting of “AllFinances” company.... Alfred, a senior executive, presents a business metric X. Richard, another senior executive, is also supposedly presenting the same business metric X. But the information provided by Alfred and Richard does not seem to agree. This is because there is no single place where the data is housed. Each (Alfred and Richard) is pulling the data from his own respective spreadsheets and each believes that his data is the most recent and accurate. The only entity confused is the BOARD!!!

- **Reports take a longer time to be prepared:** The general practice with the OLTP systems is to purge older data from transaction processing systems. This is done with the intention of controlling the expected response time. This purged data along with the current data make it to the data warehouse where there is presumably less requirement to control expected response time. The warehouse data is then used to meet the query and reporting needs. Therefore it becomes difficult, if not impossible, to furnish a report based on some characteristic at a previous point in time. For example, it is difficult, if not impossible, to get a report that presents the salaries of employees at grade Level 6 as of the beginning of each quarter in 2003 because the transaction processing system has the data for the current fiscal year. This type of reporting problem can be easily resolved if the enterprise implements data warehouse that can handle what is called the “slowly changing dimension” issue. The “slowly changing dimension” is discussed in greater detail in Chapter 7, “Multidimensional Data Modeling”.
- **Little or no scope for ad hoc querying or queries that require historical data:** The operational systems of records do not archive historical data. The queries demanding the usage and analysis of historical data either cannot be satisfied or take a long time. There are some pre-defined/pre-existing reports that work well but the transaction processing system has little or no support for ad hoc querying, particularly those requiring historical data. *Picture this...* A meeting of senior managers is in progress. Decision has to be made on the launch of a new product in an existing market in the next quarter, i.e. quarter IV of the fiscal year. The senior executives are going through reports. The reports provide them with information on the sales figures of the existing market in the quarter gone by, i.e. quarter III. One senior manager requests data for the last five years to compare the sales trends in quarter III. This has come as an ad hoc requirement and there is no report to support this. A report to satisfy the above query can be generated but it is sure to take time in the absence of a data warehouse.

## 6.2 DEFINITION OF DATA WAREHOUSE

---

According to William H. Inmon, “*A data warehouse is a subject-oriented, integrated, time variant and non-volatile collection of data in support of management's decision making process.*” Let us look at each of the terms used in the above definition more closely:

- **Subject-oriented:** A data warehouse collects data of subjects such as “customers”, “suppliers”, “partners”, “sales”, “products”, etc. spread across the enterprise or organization. A data mart on the other hand deals with the analysis of a particular subject such as “sales”.
- **Integrated:** A typical enterprise will have a multitude of enterprise applications. It is not unlikely that these applications are on heterogeneous technology platforms. It is also not unlikely that these applications use varied databases to house their data. Few of the applications may exist in silos. Few others may be sharing a little information between them. A data warehouse will serve to bring together the data from these multiple disparate (meaning differing in the format and content of data) sources after careful cleansing and transformation into a unified format to serve the information needs of the enterprise.
- **Time-variant:** A data warehouse keeps historical data while an OLTP (On-Line Transaction Processing) system will usually have the most up-to-date data. From a data warehouse, one can retrieve data that is 3 months, 6 months, 12 months, or even older. For example, a transaction system may hold the most recent address of a customer, whereas a data warehouse can hold all addresses associated with a customer recorded, say, over the last five years.

- **Non-volatile:** We have learnt earlier that transaction processing, recovery, and concurrency control mechanisms are usually associated with OLTP systems. A data warehouse is a separate physical store of data transformed from the application data found in the operational environment.

## 6.3 WHAT IS A DATA MART?

---

### Picture this...

The “GoodsForAll” enterprise has successfully implemented an enterprise-wide data warehouse. This data warehouse has data collected for all the customers and sales transactions from every unit/division and subsidiary in the business. The data warehouse true to its nature provides a homogenized, unified, and integrated view of information. It has proved very useful to the “GoodsForAll” enterprise. The market research wing of the “GoodsForAll” enterprise wishes to access the data in the data warehouse. They have plans to execute a predictive analytics application on the data stored in the data warehouse and look at how the analysis can help provide better business gains. The data architect of the “GoodsForAll” enterprise has decided to create a data mart for the market research unit. A data mart is meant to provide single domain data aggregation that can then be used for analysis, reporting, and/or decision support. Data marts can be sourced from the enterprise-wide data warehouse or can also be sourced directly from the operational/transactional systems. These data marts can also perform transformations and calculations on the data housed within. When compared to the data warehouse, data marts are restricted in their scope and business purpose.

Is it a good idea to go for a data mart for virtually every business process/event? The answer is “No”. This could result in several disparate and independent data marts. Chances are that it will become a challenge to ensure the single version of truth.

## 6.4 WHAT IS THEN AN ODS?

---

An “operational data store” (ODS) is similar to a data warehouse in that several systems around the enterprise feed operational information to it. The ODS processes this operational data to provide a homogeneous, unified view which can then be utilized by analysts and report-writers alike for analysis and reporting.

An ODS differs from an enterprise data warehouse in that it is not meant to store and maintain vast amounts of historical information. An ODS is meant to hold current or very recent operational data. Why is this required? Sometimes it is required to perform an instant analysis on the more recent data to allow one to respond immediately to a given situation.

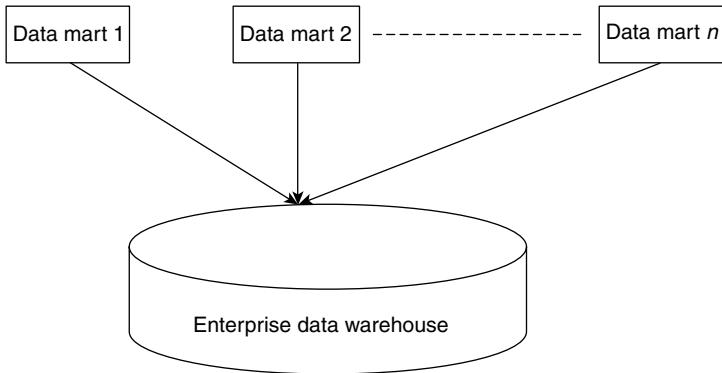
There are cases where some enterprises use the ODS as a staging area for the data warehouse. This would mean that the integration logic and processes are built into the ODS. On a regular basis, the data warehouse takes the current processed data from the ODS and adds it to its own historical data.

## 6.5 RALPH KIMBALL'S APPROACH VS. W.H. INMON'S APPROACH

---

There are two schools of thought when it comes to building a data warehouse.

According to Ralph Kimball, “*A data warehouse is made up of all the data marts in an enterprise.*” This is a bottom-up approach which essentially means that an enterprise-wide data warehouse is a confluence of all the data marts of the organization (Figure 6.2). Quite opposite to Kimball’s approach is the top-down approach prescribed by W.H. (Bill) Inmon. According to Inmon, “*A data warehouse is*



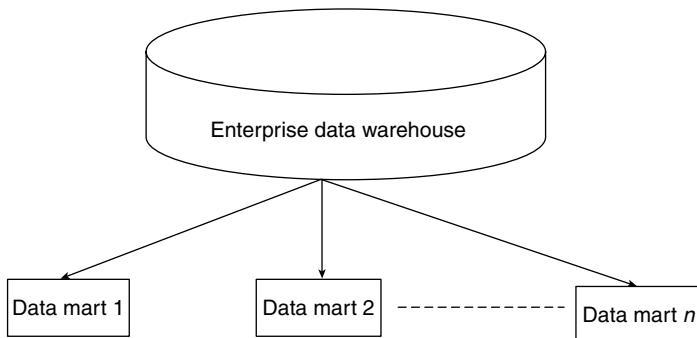
**Figure 6.2** Ralph Kimball's approach to building a data warehouse.

*a subject-oriented, integrated, non-volatile, time-variant collection of data in support of management's decisions."* Figure 6.3 depicts Inmon's top-down approach.

Now, the question is: What decides whether an organization should follow Ralph Kimball's approach or should it go the Inmon's way when it comes to building the enterprise data warehouse? Does the size of the organization play a role in this decision? It has been found that small organizations will benefit by building the data warehouse following the Kimball approach, whereas large organizations will find Inmon's approach extremely lucrative. Let us dwell into the reason.

Kimball's approach is faster, cheaper, and less complex. Contrary to this, Inmon's approach is more expensive and is a time-consuming slower process involving several complexities. However, it is able to achieve the "single version of truth" for large organizations. Therefore, it is worth investment of time and efforts. The single version of truth might be compromised in Kimball's approach, and the reason is very obvious. If you have a large organization, you will have several independent data marts, with each data mart proclaiming to have the genuine corporate data. The confused entity here is the end-user!!! He has absolutely no idea which data mart to turn to.

Let us explain this with an example. Assume you are running a small business from your home. You can make do with the help of a single personal computer. Yes, you do realize the importance of keeping data organized and also the need to maintain historical data. There is, however, no question of the single



**Figure 6.3** Bill Inmon's approach to building a data warehouse.

version of truth when a single person is running the show. Integration is also hardly the requirement. Picture another scenario wherein we have a large corporate house, almost 2,00,000 employees, several business divisions within the enterprise, multifarious requirements, multiple perspectives, and multiple decisions to make. There is a clear need for integration of data. Large organizations rely heavily on the single version of truth for effective decision making. Therefore, Inmon's architecture looks more alluring to large organizations.

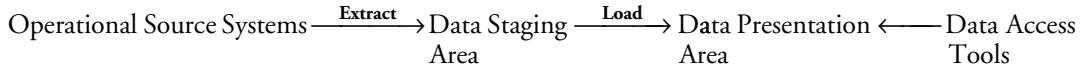
## 6.6 GOALS OF A DATA WAREHOUSE

The prime goal of a data warehouse is to *enable* users' appropriate access to a *homogenized* and *comprehensive* view of the organization. This in turn will support the *forecasting* and *decision-making* processes at the *enterprise* level. Described below are the main goals of a data warehouse.

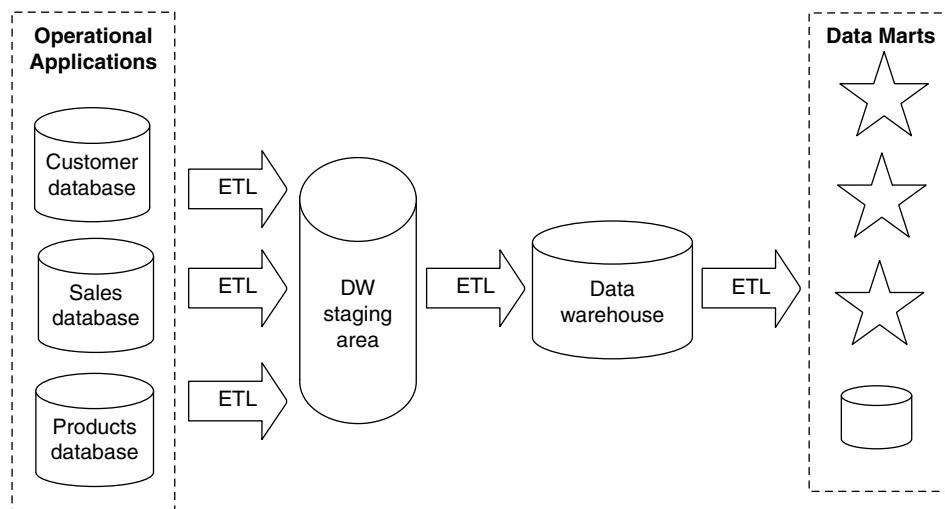
- **Information accessibility:** Data in a data warehouse must be easy to comprehend, both by the business users and developers alike. It should be properly labelled to facilitate easy access. The business users should be allowed to slice and dice the data in every possible way (slicing and dicing refers to the separation and combination of data in infinite combinations).
- **Information credibility:** The data in the data warehouse should be credible, complete, and of desired quality. Let us go back to the board meeting of "AllFinances" mentioned earlier. Suppose in this meeting Alfred presents a business metric X. If Richard also presents the same business metric X, then the information provided by both Alfred and Richard should be consistent.
- **Flexible to change:** Business situations change, users' requirements change, technology changes, and tools to access data may also change. The data warehouse must be adaptable to change. Addition of new data from disparate sources or new queries against the data warehouse should not invalidate the existing information in the data warehouse.
- **Support for more fact-based decision making:** "Manage by fact" seems to be the buzzword these days. The data warehouse should have enough pertinent data to support more precise decision making. What is also required is that the business users should be able to access the data easily.
- **Support for the data security:** The data warehouse maintains the company's confidential information. This information falling into wrong hands will do more damage than not having a data warehouse at all. There should be mechanisms in place to enable the provision of information in the required format to only those who are supposed to receive it.
- **Information consistency:** Information consistency is about a single/consistent version of truth. A data warehouse brings data from disparate data sources into a centralized repository. Users from across the organization make use of the data warehouse to view a single and consistent version of truth.

## 6.7 WHAT CONSTITUTES A DATA WAREHOUSE?

Data from the operational systems flow into the staging area where it undergoes transformation and is then placed in the presentation area from where it can be accessed using data access tools. Refer to Figure 6.4.



- **Operational source systems:** These systems maintain transactional or operational data. They are outside the data warehouse. There could be any number of such systems (similar or disparate) feeding data to the data warehouse. They may maintain little historical data. The queries against such systems generally return an answer set (also called record set or result set) of one or few records.
- **Data staging area:** The data staging area comprises storage space for the data that has been extracted from various disparate operational sources. It also consists of a set of processes related to data quality. There are three major processes popularly referred to as extraction, transformation, and loading. The data staging area is off-limits from the business users and is not designed to answer queries however simple they may be, or to offer presentation services.
- **Data presentation area:** Data staging area is off-limits to the business users. But data presentation area is the interface or the front face of the data warehouse with which the business community interacts via the data access tools. It is just a collection of integrated data marts. What does the term “integrated” imply? Consider an example of a bank system consisting of deposit and withdraw subsystems. It is likely that customer A (a person by the name Connors) of the deposit system and Customer B of the loan system are one and the same person. Without integration there is no way to know this. What are data marts? They are just a subset of the data maintained in the data warehouse, which is of value to a specific group of users. Furthermore, data marts can be either independent or dependent. Independent data marts are sourced directly from one or more operational systems, or can be sourced from external information providers, or can be sourced from data generated locally from within a department or unit or function. Dependent data marts, on the other hand, are sourced from enterprise data warehouses.



**Figure 6.4** Operational Data Sources → Data Warehouse → Data Marts.

- **Data access tools:** Data access tools can be ad hoc query tools used to query the data presentation area. A data access tool can also be a reporting tool or a data modelling/mining application (for trend analysis or prediction, etc.).

### 6.7.1 Data Sources

- Data sources (transaction or operational or external information providers) are the sources from which we extract data.
- In data warehousing, we extract data from different disparate sources (heterogeneous sources such as text files, .CSV files, .XLS files, .MDB files, etc.), transform this data into a certain format (unified/data warehouse format), and then load the data in data warehouse.
- The raw material for any data integration is provided by data sources.
- Data sources refer to any of the following types of source:
  - Data storage Media (flat file, DBMS, etc.).
  - Data organization (linked data in COBOL mainframes, normalized forms in RDBMS).
- The data in these data sources can be present in any format.

## 6.8 EXTRACT, TRANSFORM, LOAD

ETL (Extract, Transform, and Load) is a three-stage process in database usage, especially in data warehousing. It allows integration and analysis of data stored in different sources. After collecting the data from multiple varied sources (extraction), the data is reformatted (from host format to warehouse format) and cleansed (to detect and rectify errors) to meet the information needs (transformation) and then sorted, summarized, consolidated, and loaded into desired end target (loading). Put simply, ETL allows creation of efficient and consistent databases. So we can say, ETL is

- Extracting data from different data sources.
- Transforming the extracted data into a relevant format to fit information needs.
- Loading data into the final target database, usually a data warehouse.

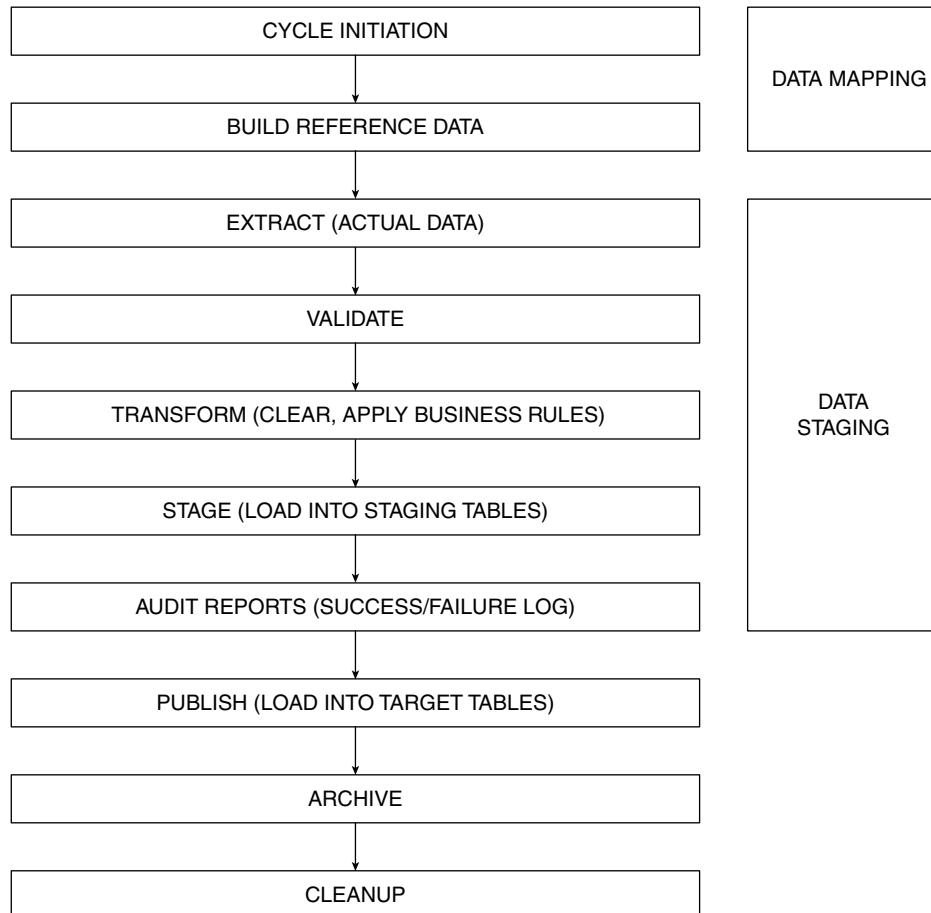
Figure 6.5 shows the intermediate stages of ETL.

### 6.8.1 Data Mapping

- It is a process of generating data element mapping between two distinct data models.
- It is the first process that is performed for a variety of data integration tasks which include
  - Data transformation between data source and data destination.
  - Identification of data relationships.
  - Discovery of hidden sensitive data.
  - Consolidation of multiple databases into a single database.

### 6.8.2 Data Staging

- A data staging area can be defined as an intermediate storage area that falls between the operational/transactional sources of data and the data warehouse (DW) or data mart (DM).



**Figure 6.5** Intermediate stages of ETL.

- A staging area can be used, among others, for the following purposes:
  - To gather data from different sources ready to be processed at different times.
  - To quickly load information from the operational database.
  - To find changes against current DW/DM values.
  - To cleanse data.
  - To pre-calculate aggregates.

### ***Data Extraction***

It is a process of collecting data from different data sources. In other words, it is the consolidation of data from different sources having different formats. Flat files and relational databases are most common data sources. Depending upon the type of source data, the complexity of extraction may vary. The storage of intermediate version of data is very necessary. This data is required to be backed up and archived. The area where the extracted data is stored is called staging area.

## ***Data Transformation***

- A series of rules or functions is applied to the data extracted from the source to obtain derived data that is loaded into the end target.
- Depending upon the data source, manipulation of data may be required. If the data source is good, its data may require very less transformation and validation. But data from some sources might require one or more transformation types to meet the operational needs and make data fit in the end target.
- Some transformation types are
  - Selecting only certain columns to load.
  - Translating a few coded values.
  - Encoding some free-form values.
  - Deriving a new calculated value.
  - Joining together data derived from multiple sources.
  - Summarizing multiple rows of data.
  - Splitting a column into multiple columns.
- Data transformation is the most complex and, in terms of production, the most costly part of the ETL process. Let us explain this point with an example: The government has announced a “Student Exchange Program”. The organizing committee of the “Student Exchange Program” has requested reputed universities of the country to furnish data on their top 10 students in the final semester of their graduation degree. There are some universities that maintain the students score on a grade scale of 0–10. There are other universities that maintain the students score on a grade scale of “A” to “E”. To add to the complications, there are universities that maintain the students score on a grade scale of 0% to 100% and some others that follow the percentile system. Clearly, once the data is received from the various universities, it will need to be transformed for some homogeneity to set in before being loaded into the data warehouse.

## ***Data Loading***

- The last stage of the ETL process is loading which loads the extracted and transformed data into the end target, usually the data warehouse.
- Data can also be loaded by using SQL queries.

We will understand the concept of data loading using the case study of DIIT library that we mentioned earlier. Assume for the sake of simplicity that each department (IT, CS, IS, and SE) stores data in separate sheets in an Excel workbook, except the Issue\_Return table which is in an Access database. Each sheet contains a table. There are five tables in all:

1. Book (Excel)
2. Magazine (Excel)
3. CD (Excel)
4. Student (Excel)
5. Issue\_Return (Access)

Previews of each of these tables are given in Figures 6.6–6.10.

|   | A        | B                    | C             | D                  | E          | F            |
|---|----------|----------------------|---------------|--------------------|------------|--------------|
| 1 | Book_ID  | Book_Name            | Author        | Publisher          | Technology | No of Copies |
| 2 | BLRB2001 | WPF PROGRAMING       | SAMS WILLEY   | WILLEY PUBLICATION | .NET       | 2            |
| 3 | BLRB2002 | C# PROGRAMING        | JACK WILCH    | SAMS PUBLICATION   | .NET       | 10           |
| 4 | BLRB2003 | ADODE FLEX 2         | MARILDA WHITE | TMH PUBLICATION    | JEE        | 10           |
| 5 | BLRB2004 | THE FINANCE HANDBOOK | VIKRAM PANDEY | THM PUBLICATION    | GENERAL    | 10           |
| 6 | BLRB2005 | SAP R/3 HANDBOOK     | JIMY ARNOLD   | PEARSON EDUCATION  | SAP        | 10           |

**Figure 6.6** A part of the Book Excel sheet.

|   | A           | B              | C          | D            | E      |
|---|-------------|----------------|------------|--------------|--------|
| 1 | Magazine_ID | Title          | Technology | No of Copies | Period |
| 2 | BLRM2001    | READERS DIGEST | GENERAL    | 2            | MON    |
| 3 | BLRM2002    | VISUAL STUDIO  | .NET       | 3            | MON    |
| 4 | BLRM2003    | JAVA FOR ME    | JEE        | 2            | MON    |
| 5 | BLRM2004    | SQL SERVER     | .NET       | 2            | MON    |
| 6 | BLRM2005    | MSDN           | .NET       | 1            | YRL    |

**Figure 6.7** A part of the Magazine Excel sheet.

|   | A        | B                     | C          | D            |
|---|----------|-----------------------|------------|--------------|
| 1 | CD_id    | Title                 | Technology | No of Copies |
| 2 | BLRC2001 | WPF PROGRAMMING       | .NET       | 2            |
| 3 | BLRC2002 | C# PROGRAMMING        | .NET       | 10           |
| 4 | BLRC2003 | SAP CRM CONFIGURATION | SAP        | 5            |
| 5 | BLRC2004 | STEP BY STEP SERVLET  | JEE        | 10           |
| 6 | BLRC2005 | ORACLE 10G            | DB         | 10           |

**Figure 6.8** A part of the CD Excel sheet.

|    | A       | B               | C                | D              | E                         | F                | G            | H          |
|----|---------|-----------------|------------------|----------------|---------------------------|------------------|--------------|------------|
| 1  | Stud_id | Stud_First_name | Stud_Middle_name | Stud_Last_name | Email                     | Address          | Phone Number | Birth Date |
| 2  | 20001   | Bhanu           | PRASAD           | Galgotia       | B_galgotia@yahoo.com      | Saharanpur       | 9952392126   | 21-08-1978 |
| 3  | 20002   | Parul           |                  | Dalakoti       | parul_rocks@gmail.com     | Chudiala         | 9555533325   | 03-03-1987 |
| 4  | 20003   | Sushma          | Billo            | Sharma         | sush_sweets@hotmail.com   | IQBALPUR         | 9554477886   |            |
| 5  | 20004   | girish          | Chandra          | RAWAT          | giri_calls@rocketmail.com | Baliya           | 9325489654   |            |
| 6  | 20005   | Fakir           |                  | Das            | das_fakir@yahoo.co.in     | Sunehti Khadkadi | 9654238552   | 1-1-1971   |
| 7  | 20006   | Lovepreet       | Singh            | Chadha         | sensation_star@gmail.com  | Patiala          | (8345)225689 | 18/03/1984 |
| 8  | 20007   | Jaishankara     | paRIMAL          | Chakshu        | Jai.chakshu@gmail.com     | Khatauli         | 78562 31848  | 17/04/1982 |
| 9  | 20008   | Sathyaranayanan | Samay            | siNgh          | Sathy.rules@gmail.com     | Modinagar        | 9565622261   |            |
| 10 | 20009   | Gaganmeet       |                  | Hansra         | gaggu_pranks@gmail.com    | Varanasi         | 98789-54655  | 11.10.1975 |
| 11 | 20010   | Bhavya          |                  | dhingra        | Dhingra.bhavya@gmail.com  | BAREILLY         | 99988-84569  | 22/10/1988 |
| 12 | 20011   | Abhinav         | Kumar            | patnala        | Teddy_patnala@yahoo.co.in | Mysore           | 7700598648   | 20/06/1979 |
| 13 | 20012   | ABHUIT          |                  | Ramcharandas   | charan_me_aaape@gmail.com | Pune             | 9982485698   | 1.1.1987   |
| 14 | 20013   | Mukulika        |                  | senroy         | Sweet.muku@gmail.com      | CHEnnai          | 9900665598   | 15-08-1984 |
| 15 | 20014   | JAYARAM         |                  | SAINI          | Jayaram_s@hotmail.com     | New Jalpaiguri   | 9874652365   |            |
| 16 | 20015   | RanCHORDAS      | Shyamaldas       | Chhanchhad     | wise_brains@gmail.com     | Ahmedabad        | 9879879876   | 15/10/1985 |

**Figure 6.9** A part of the Student Excel sheet.

| Issue_Return   |         |          |            |           |             |             |
|----------------|---------|----------|------------|-----------|-------------|-------------|
| Transaction_ID | Stud_ID | Item_ID  | Issue_Date | Due_Date  | Return_Date | Damage_fine |
| 101001         | 20004   | BLRB2002 | 15-May-08  | 15-Jun-08 | 14-Jun-08   |             |
| 101002         | 20007   | BLRB2001 | 15-May-08  | 15-Jun-08 | 18-Jun-08   |             |
| 101003         | 20014   | BLRM2002 | 18-May-08  | 18-Jun-08 | 18-Jun-08   |             |
| 101004         | 20015   | BLRB2004 | 22-May-08  | 22-Jun-08 | 22-Jun-08   |             |
| 101005         | 20001   | BLRC2005 | 22-Jul-08  | 22-Aug-08 | 23-Aug-08   | 110         |
| 101006         | 20015   | BLRB2001 | 29-Jul-08  | 29-Aug-08 | 27-Aug-08   |             |
| 101007         | 20010   | BLRB2004 | 29-Jul-08  | 29-Aug-08 | 31-Aug-08   |             |
| 101008         | 20007   | BLRM2003 | 4-Aug-08   | 4-Sep-08  | 4-Sep-08    | 400         |
| 101009         | 20014   | BLRC2003 | 7-Aug-08   | 7-Sep-08  | 7-Sep-08    |             |
| 101010         | 20015   | BLRC2001 | 10-Aug-08  | 10-Sep-08 | 3-Sep-08    |             |

**Figure 6.10** A part of the Issue\_Return MS Access database table.

Of all these tables, the **Issue\_Return** table is the table that keeps records of all transactions (issue and return) made in the library by registered users (students). Registered users' data is stored in the **Student** table. The rest of the three tables contain data about the items (books, CDs, and magazines) that can be issued from the library.

Let us again look at the requirements of DIIT administration. “The DIIT administration is in need of a report that indicates the annual spending on library purchases. The report should further drill down to the spending by each department by category (books, CDs/DVDs, magazines, journals, etc.).” Prof. Frank had suggested the building of a data warehouse. In order to implement Prof. Frank’s solution, the following requirements should be satisfied:

- Data in all tables should be in proper format and must be consistent.
- The three tables – **Book**, **Magazine**, and **CD** – should be combined into a single table that contains details of all these three categories of items.
- The student table should have a single column containing the full name of each student.
- Phone numbers must be stored in a numeric (integer) column.
- Date columns should be in a uniform format, preferably as a datetime data type.
- String columns (Fullname, Address) should be in uniform case (preferably uppercase).

To achieve the stated requirements, and to clean the library data of other minor (but significant) inconsistencies, Prof. Frank decided to first profile the data (manually and/or using a profiling tool) to check for quality issues, and then perform ETL to migrate it to the data staging area, and finally to the data warehouse. This process involves some very important processes, namely **data profiling** and **data integration** (through ETL).

## 6.9 WHAT IS DATA INTEGRATION?

- It is the integration of data present in different sources for providing a unified view of the data. Refer to Figure 6.11.
- It is the ability to consolidate data from several different sources while maintaining the integrity and reliability of the data.
- For example, in the library scenario:
  - The item (Book, CD, Magazine) and student (Student) data is maintained in Excel files, whereas the transactional data (Issue\_Return) is stored in the Access database.
  - The library application would need to integrate all this data and present it in a unified manner to the end-user while maintaining the integrity and reliability of data.
  - Data integration would play a vital role in this scenario.

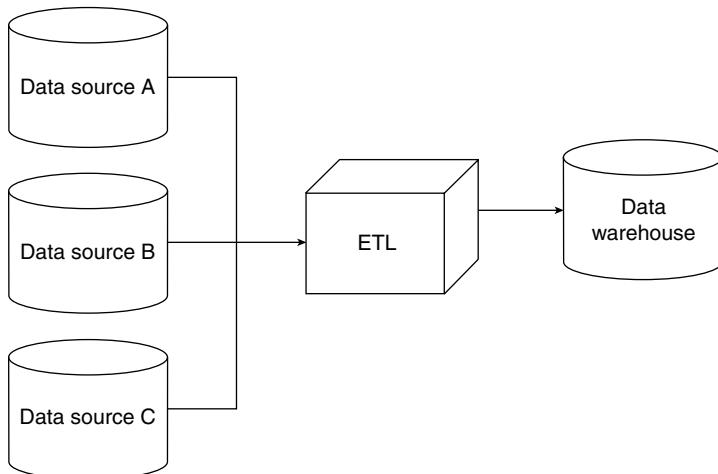
### 6.9.1 Two Main Approaches to Data Integration

The two main approaches to data integration are – schema integration and instance integration.

#### **Schema Integration**

Multiple data sources may provide data on the same entity type. The main goal is to allow applications to transparently view and query the data as though it is one uniform data source. This is done using various mapping rules to handle structural differences.

“Schema integration is developing a unified representation of semantically similar information, structured and stored differently in the individual databases.”



**Figure 6.11** Data integration.

## Picture this...

A retail outlet has branches spread across the country. There are some 20 different schemas that need to be integrated. One of the branches, “Collingwood branch”, stores the transaction details such as “CustID”, “TransactionID”, “ProductID”, “UnitQuantity”, etc. Another branch, “Trembly Park branch”, stores the transaction details such as “CustomerNumber”, “TransactionID”, “ProductID”, “UnitQuantity”, etc.

### “Collingwood branch”

| <i>CustID</i> | <i>TransactionID</i> | <i>ProductID</i> | <i>UnitQuantity</i> | _____ |
|---------------|----------------------|------------------|---------------------|-------|
| C101          | T1001                | P1010            | 10                  |       |

### “Trembly Park branch”

| <i>CustomerNumber</i> | <i>TransactionID</i> | <i>ProductID</i> | <i>UnitQuantity</i> | _____ |
|-----------------------|----------------------|------------------|---------------------|-------|
| C201                  | T1007                | P1111            | 22                  |       |

The data analyst is looking at merging the data from both the branches, “Collingwood branch” and “Trembly Park branch”. How can the data analyst make sure that the data from the database of one of the branches under the column “CustID” and the data from the database of another branch under the column “CustomerNumber” are mapped to the same column in the target database? The answer is by looking up the metadata information. The metadata is data about data. It defines each column, its data type, the length, the possible constraints, the range of values permitted for the attribute, the rules of dealing with NULL, Zero and blank values, etc.

Let us assume that adequate metadata was available and schema integration was successful. The result will be the following target schema.

### Target Schema:

| <i>CustID</i> | <i>TranID</i> | <i>ProductID</i> | <i>UnitQuantity</i> | _____ |
|---------------|---------------|------------------|---------------------|-------|
| C101          | T1001         | P1010            | 10                  | _____ |
| C201          | T1007         | P1111            | 22                  | _____ |

### Instance Integration

Data integration from multiple heterogeneous data sources has become a high-priority task in many large enterprises. Hence to obtain the accurate semantic information on the data content, the information is being retrieved directly from the data. It identifies and integrates all the instance of the data items that represent the real-world entity, distinct from the schema integration.

## Picture this...

A corporate house has almost 10,000 employees working for it. This corporate house does not have an ERP system. It has few enterprise applications, all existing in silos. There is a “ProjectAllocate” application that stores the project allocation details of the employees. There is “EmployeeLeave” application that stores the details about the leaves availed by the employees. There is yet another “EmployeeAttendance”

application that has the attendance data of the employees. And, of course, there is an “EmployeePayroll” application that stores the salary details of the employees. The management has decided to consolidate all the details of every employee in a data warehouse.

Let us look at the challenges. Assume there is an employee by the name, “Fred Aleck”. His name is stored as “Fred Aleck” by the “ProjectAllocate” application, as “Aleck Fred” by the “EmployeeLeave” application, as “A Fred” by the “EmployeeAttendance” application, and to add to the challenge as “F Aleck” by the “EmployeePayroll” application.

You see it now... The same person and his name has been stored in four different ways.

### “ProjectAllocate”

| <i>EmployeeNo</i> | <i>EmployeeName</i> | <i>SocialSecurityNo</i> | _____ |
|-------------------|---------------------|-------------------------|-------|
| 10014             | Fred Aleck          | ADWPP10017              | _____ |

### “EmployeeLeave”

| <i>EmployeeNo</i> | <i>EmployeeName</i> | <i>SocialSecurityNo</i> | _____ |
|-------------------|---------------------|-------------------------|-------|
| 10014             | Aleck Fred          | ADWPP10017              | _____ |

### “EmployeeAttendance”

| <i>EmployeeNo</i> | <i>EmployeeName</i> | <i>SocialSecurityNo</i> | _____ |
|-------------------|---------------------|-------------------------|-------|
| 10014             | A. Fred             | ADWPP10017              | _____ |

### “EmployeePayroll”

| <i>EmployeeNo</i> | <i>EmployeeName</i> | <i>SocialSecurityNo</i> | _____ |
|-------------------|---------------------|-------------------------|-------|
| 10014             | F. Aleck            | ADWPP10017              | _____ |

Let us look at how instance integration can solve this problem. One possible solution is presented here. Look up all the records of the employee using the “EmployeeNo” or the “SocialSecurityNo”. And then, replace the value in the “EmployeeName” column with one consistent value such as “Fred Aleck” in the example above.

### “ProjectAllocate”

| <i>EmployeeNo</i> | <i>EmployeeName</i> | <i>SocialSecurityNo</i> | _____ |
|-------------------|---------------------|-------------------------|-------|
| 10014             | Fred Aleck          | ADWPP10017              | _____ |

### “EmployeeLeave”

| <i>EmployeeNo</i> | <i>EmployeeName</i> | <i>SocialSecurityNo</i> | _____ |
|-------------------|---------------------|-------------------------|-------|
| 10014             | Fred Aleck          | ADWPP10017              | _____ |

### **“EmployeeAttendance”**

| <i>EmployeeNo</i> | <i>EmployeeName</i> | <i>SocialSecurityNo</i> | _____ |
|-------------------|---------------------|-------------------------|-------|
| 10014             | Fred Aleck          | ADWPP10017              | _____ |

### **“EmployeePayroll”**

| <i>EmployeeNo</i> | <i>EmployeeName</i> | <i>SocialSecurityNo</i> | _____ |
|-------------------|---------------------|-------------------------|-------|
| 10014             | Fred Aleck          | ADWPP10017              | _____ |

## **6.9.2 Need and Advantages for Data Integration**

Data integration is the need of the hour.

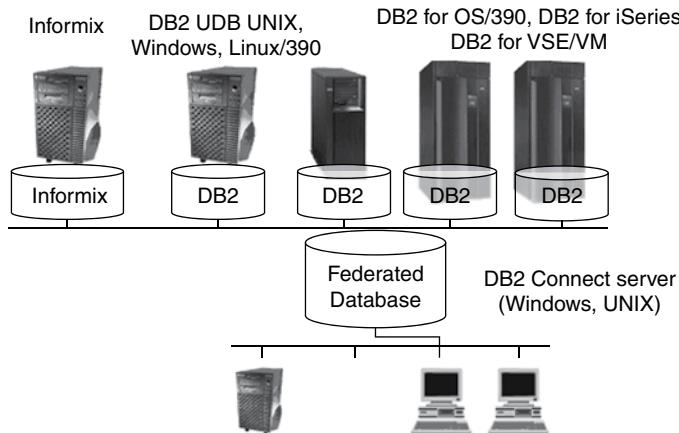
- It is of benefit to decision makers who will be able to quickly access information based on a key variable along with the query against existing data from past studies in order to gain meaningful insights.
- It helps reduce costs, overlaps, and redundancies, and business will be less exposed to risks and losses.
- It helps in better monitoring of key variables such as trending patterns and consumer behavior across geographies, which would alleviate the need to conduct more studies and surveys, bringing about reduced spending on R&D and minimization of redundancy.

## **6.9.3 Common Approaches of Data Integration**

### ***Federated Databases***

Several database systems are combined/integrated into one single FDBS. Figure 6.12 shows the FDBS. Described below are some features of the FDBS.

- A **federated database system (FDBS)** can be stated as a category of database management system (DBMS) that integrates multiple disparate and autonomous database systems into one single **federated database**. Constituent databases may be geographically decentralized and interconnected via computer network. Since the constituent databases remain more or less autonomous, having an FDBS is a contrastable, but feasible, alternative to the task of merging several disparate databases. A federated database can also be called as **virtual database**. It is a fully integrated, logical composite result of all of its constituent databases.
- The federated database system was first defined by McLeod and Heimbigner. According to them, the FDBS is the one which defines the architecture and interconnects the databases supporting partial sharing and coordination among database systems.
- A Uniform User Interface can be provided through data abstraction which will enable users and clients to retrieve data from multiple non-contiguous databases (that are combined into a federated database) using a single query (even if the constituent databases are heterogeneous).
- An FDBS should have the ability to decompose a larger/complex query into smaller sub-queries before submitting them to the relevant constituent DBMSs, after which the system combines the individual result sets of each sub-query. Since various DBMSs work on different query languages, the FDBS can apply wrappers to sub-queries, so that they can be translated to the corresponding query languages.



**Figure 6.12** A federated database system.

- A federated database may consist of a collection of heterogeneous databases. In such a case, it allows applications to look at data in a relatively more unified way without the need to duplicate it across individual databases or make several smaller, multiple queries and combine the results manually.

### Data Warehousing

Data integration is an important, fundamental part of data warehousing. It is a process vital to the creation of a robust, yet manageable and highly informative data resource whose purpose is to deliver business intelligence (BI) solutions. Data integration includes the necessary activities:

- Acquire data from sources (*extract*).
- Transform and cleanse the data (*transform*).
- Load the transformed data into the data store, which may be a data warehouse or a data mart (*load*).

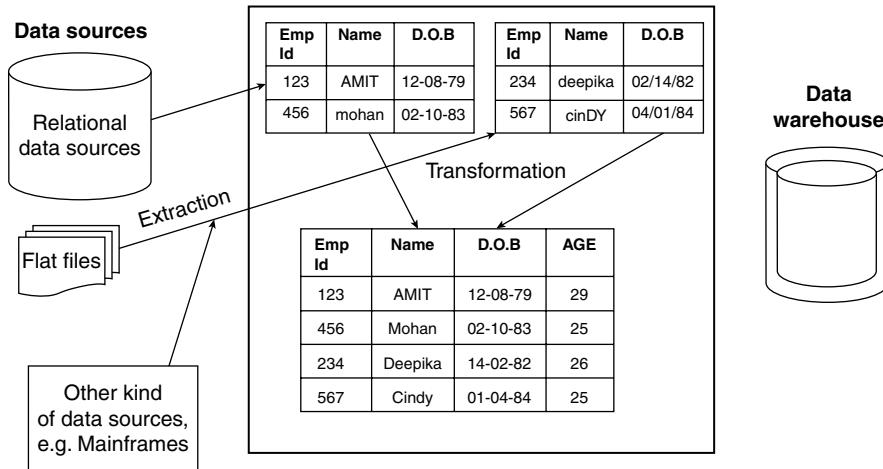
The body of knowledge required for data integration includes

- Concepts and skills required to analyze and qualify source data.
- Data profiling.
- Source-to-target mapping.
- Data cleansing and transformation.
- ETL development.

Using data warehousing method, data is first pulled (extracted) from various data sources. Next, the extracted data is converted into a common format, so that data sets are compatible with one another (transformed). Lastly, this new data is loaded into its own database (loaded). When a user runs a query on the data warehouse, the required data is located, retrieved, and presented to the user in an integrated view.

For example, a data warehouse may contain updated information on drive-in restaurants and road maps of a certain town in its tables. Once you submit a query on the warehouse, it would integrate the two (restaurant locations/timings and road maps) together and send the result set view back to you. The various primary concepts used in data warehousing are

- ETL (extract, transform, load).
- Component-based (data mart).

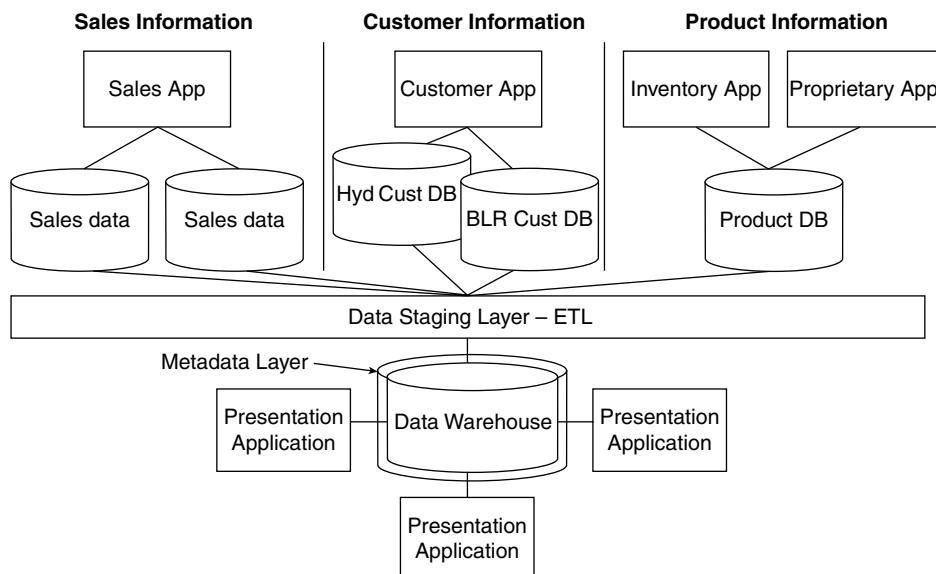


**Figure 6.13** Operational data sources (heterogeneous) to a unified data warehouse.

- Dimensional models and schemas.
- Metadata-driven.

As shown in Figure 6.13, data from several different/disparate sources is extracted, transformed, and loaded into a single database, which can then be queried as a single schema. Let us understand this with the help of an example depicted in Figure 6.14.

From Figure 6.14, it is clear that the sales information, product information, and customer information have been extracted from different sources. The extracted information then undergoes the ETL process where the data is stored in data staging area. Finally, the data from the staging database is loaded in the data warehouse.



**Figure 6.14** Operational data sources (heterogeneous) to a unified data warehouse.

**Table 6.1** Difference between a federated database and a data warehouse

| <i>Federated</i>  | <i>Data Warehouse</i>   |
|---|---|
| Preferred when the databases are present across various locations over a large area (geographically decentralized). | Preferred when the source information can be taken from one location. |
| Data would be present in various servers.   | The entire data warehouse would be present in one server.             |
| Requires high speed network connection.   | Requires no network connection.                                       |
| It is easier to create as compared to data warehouse.   | Its creation is not as easy as that of the federated database.        |
| Requires no creation of new database.   | Data warehouse must be created from scratch.                          |
| Requires network experts to set up the network connection.  | Requires database experts such as data Steward.                       |

### ***Memory-Mapped Data Structure***

- It is useful when one needs to do in-memory data manipulation, and the data structure is large. It's mainly used in the dot net platform and is always performed with C# or using VB.NET
- It is a much faster way of accessing the data than using memory stream.

## **6.10 DATA INTEGRATION TECHNOLOGIES**

The following are the technologies used for data integration:

### **1. Data Interchange**

- a. It is the structured transmission of organizational data between two or more organizations through electronic means; used for the transfer of electronic documents from one computer to another (i.e., from one corporate trading partner to another).
- b. Data interchange must not be seen merely as email. For instance, organizations might want to do away with bills of lading (or even checks), and use appropriate EDI messages instead.

### **2. Object Brokering**

- a. An ORB (Object Request Broker) is a certain variety of middleware software. It gives programmers the freedom to make calls from one computer to another over a computer network.
- b. It handles the transformation of in-process data structure to and from the byte sequence.

### **3. Modeling Techniques:** There are two logical design techniques:

- a. **ER Modeling:** Entity Relationship (ER) Modeling is a logical design technique whose main focus is to reduce data redundancy. It is basically used for transaction capture and can contribute in the initial stages of constructing a data warehouse. The reduction in the data redundancy solves the problems of inserting, deleting, and updating data but it leads to yet another problem. In our bid to keep redundancy to the minimum extent possible, we end up creating a whole lot of tables. These huge numbers of tables imply dozens of joins between them. The result is a massive spider web of joins between tables.

*What could be the problems posed by ER Modeling?*

- End-users find it difficult to comprehend and traverse through the ER model.
- Not too many software exist which can query a general ER model.
- ER Modeling cannot be used for data warehousing where the focus is on performance access and satisfying ad hoc, unanticipated queries.

Example: Consider a library transaction system of a department of DIIT. Every transaction (issue of book to a student or return of book by a student) are recorded. Let us draw an ER model to represent the above-stated scenario.

Steps to drawing an ER model:

- Identify entities.
- Identify relationships between various entities.
- Identify the key attribute.
- Identify the other relevant attributes for the entities.
- Draw the ER diagram.
- Review the ER diagram with business users and get their sign-off.

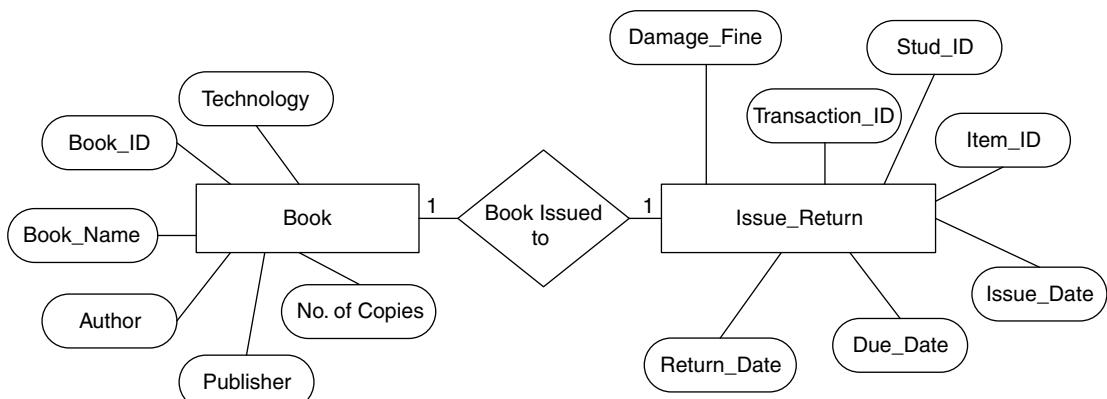
Figure 6.15 considers 2 entities: Book and Issue\_Return.

Book entity has the attributes: Book\_ID (key attribute), Technology, Book\_Name, Author, Publisher, NoOfCopies.

Issue\_Return entity has the attributes: Stud\_ID, Item\_ID, Transaction\_ID (key attribute), Issue\_Date, Due\_Date, Return\_Date, Damage\_Fine, etc.

The relationship between the two entities (Book, Issue\_Return) is 1:1.

**b. Dimensional Modeling:** It is a logical design technique, the main focus of which is to present data in a standard format for end-user consumption. It is used for data warehouses having either a Star schema or a Snowflake schema. Every dimensional model is composed of one large table called the fact table and a number of relatively smaller tables called the dimensional tables. The fact table has a multipart primary key. Each table has a single-part primary key. The dimension



**Figure 6.15** ER diagram between Book and Issue\_Return entities.

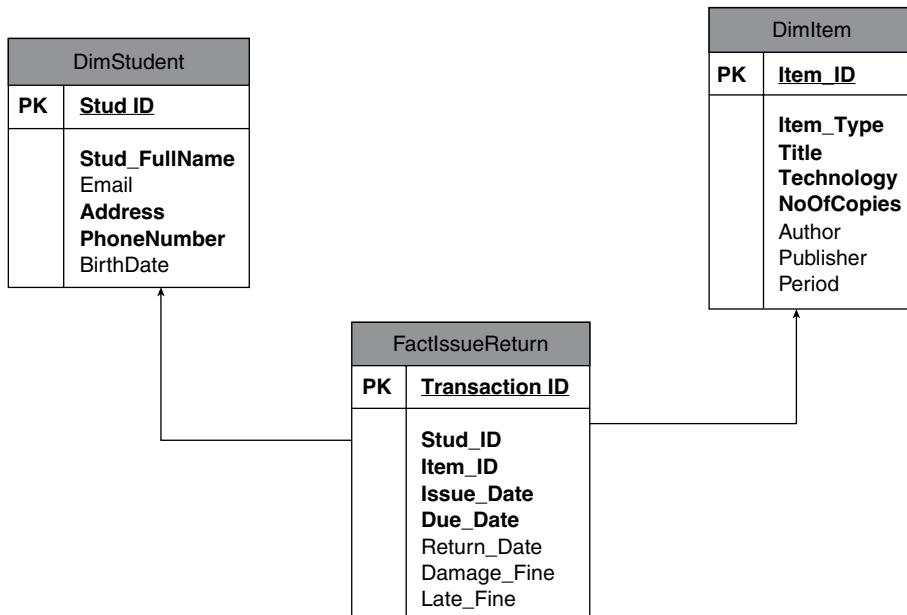
primary key corresponds to precisely one component of the multipart key of the fact table. In addition to the multipart primary key, the fact table also contains a few facts which are numeric and additive. The dimension table generally contains textual data. The fact table maintains many-to-many relationships.

*What are the perceived benefits of the dimensional modeling?*

- End-users find it easy to comprehend and traverse/navigate through the model.
- If designed appropriately, it can give quick responses to ad hoc query for information.
- A lot of tools such as ad hoc querying tools, report generation tools, data mining applications, etc. are available which can be used on the data warehouse to satisfy the decision support requirements of business users.

Consider the DIIT library case study. Prof. Frank designed a dimensional schema structure for the data warehouse as shown in Figure 6.16. The DIIT library was given permission to access the central college server machine that housed Microsoft SQL Server. Hence, the above relational schema could be implemented as an SQL Server database. Now, the major steps that were taken to transfer data from the Excel spreadsheets and Access database to SQL Server database were as follows:

1. Profiling the source data (identifying natural key columns, identify data anomalies, corrections to be made, etc.)
2. Results of profiling were:
  - a. Identification of natural keys (that would naturally be **unique** and **not null**):



**Figure 6.16** Dimensional data structure for the data warehouse of DIIT.

- *Book*: Book\_ID
  - *Magazine*: Magazine\_ID
  - *CD*: CD\_ID
  - *Issue\_Return*: Transaction\_ID
  - *Student*: Stud\_ID
- b. Removal of leading and trailing blanks wherever they occur.
- c. Removal of special characters from “PhoneNumber” column.
- d. Transforming the “Birthdate” column to a standard date format.
- e. Capitalization in “Address” column is not appropriate, hence convert all letters to capitals.
3. Choosing an appropriate ETL tool (SSIS, Kettle, Informatica, etc.) and creating ETL packages to transform and transport the data from the source database to the destination database.
4. Some major rules, transformations, and data-cleaning activities that were identified and implemented were:
- Merging together the tables **Book**, **Magazine**, and **CD** into a single table “**DimItem**”.
  - Data-type conversion.
  - Combining Stud\_First\_name, Stud\_Middle\_name and Stud\_Last\_name and creating a new column “FullName” (fully capitalized) in DimStudent, and “Late\_Fine” in FactIssueReturn (Late\_Fine) would be calculated as
- (Return\_Date – Due\_Date) × 10 (in Rs.)**
- Removing leading and trailing blanks wherever they occur.
  - Removing special characters from “PhoneNumber” column and its subsequent conversion to numerical format.
  - Standardizing the “BirthDate” column format and conversion to “DateTime” datatype.
  - Transforming all entries in the “Address” column to capitals.
  - In the fine columns (Damage\_Fine and Late\_Fine), each NULL entry be recorded as zero (0).
5. As per general practice, the Dimension tables (DimItem and DimStudent) are populated with data first, followed by the Fact table(s) (FactIssueReturn). This is because of referential constraints of fact on the dimensions.

After performing ETL, the tables in the database looked like those shown in Figures 6.17–6.19.

|    | Transaction_ID | Stud_ID | Item_ID  | Issue_Date              | Due_Date                | Return_Date             | Damage_fine | Late_fine |
|----|----------------|---------|----------|-------------------------|-------------------------|-------------------------|-------------|-----------|
| 1  | 101001         | 20004   | BLRB2002 | 2008-05-15 00:00:00.000 | 2008-06-15 00:00:00.000 | 2008-06-14 00:00:00.000 | 0           | 0         |
| 2  | 101002         | 20007   | BLRB2001 | 2008-05-15 00:00:00.000 | 2008-06-15 00:00:00.000 | 2008-06-18 00:00:00.000 | 0           | 30        |
| 3  | 101003         | 20014   | BLRM2002 | 2008-05-18 00:00:00.000 | 2008-06-18 00:00:00.000 | 2008-06-18 00:00:00.000 | 0           | 0         |
| 4  | 101004         | 20015   | BLRB2004 | 2008-05-22 00:00:00.000 | 2008-06-22 00:00:00.000 | 2008-06-22 00:00:00.000 | 0           | 0         |
| 5  | 101005         | 20001   | BLRC2005 | 2008-07-22 00:00:00.000 | 2008-08-22 00:00:00.000 | 2008-08-23 00:00:00.000 | 110         | 10        |
| 6  | 101006         | 20015   | BLRB2001 | 2008-07-29 00:00:00.000 | 2008-08-29 00:00:00.000 | 2008-08-27 00:00:00.000 | 0           | 0         |
| 7  | 101007         | 20010   | BLRB2004 | 2008-07-29 00:00:00.000 | 2008-08-29 00:00:00.000 | 2008-08-31 00:00:00.000 | 0           | 20        |
| 8  | 101008         | 20007   | BLRM2003 | 2008-08-04 00:00:00.000 | 2008-09-04 00:00:00.000 | 2008-09-04 00:00:00.000 | 400         | 0         |
| 9  | 101009         | 20014   | BLRC2003 | 2008-08-07 00:00:00.000 | 2008-09-07 00:00:00.000 | 2008-09-07 00:00:00.000 | 0           | 0         |
| 10 | 101010         | 20015   | BLRC2001 | 2008-08-10 00:00:00.000 | 2008-09-10 00:00:00.000 | 2008-09-03 00:00:00.000 | 0           | 0         |

**Figure 6.17** A part of the FactIssueReturn table.

|    | Stud_id | Stud_FullName                     | Email                     | Address           | PhoneNumber | BirthDate               |
|----|---------|-----------------------------------|---------------------------|-------------------|-------------|-------------------------|
| 1  | 20001   | BHANU PRASAD GALGOTIA             | B_galgotia@yahoo.com      | SAHARANPUR        | 9952392126  | 1978-08-21 00:00:00.000 |
| 2  | 20002   | PARUL DALAKOTI                    | parul_rocks@gmail.com     | CHUDIALA          | 9555533325  | 1978-03-03 00:00:00.000 |
| 3  | 20003   | SUSHMA BILLO SHARMA               | sush_sweets@hotmail.com   | IQBALPUR          | 9554477886  | NULL                    |
| 4  | 20004   | GIRISH CHANDRA RAWAT              | giri_calls@rocketmail.com | BALIYA            | 9325489654  | NULL                    |
| 5  | 20005   | FAKIR DAS                         | das_fakir@yahoo.co.in     | SUNEHTI KHADKHADI | 9654238552  | 1971-01-01 00:00:00.000 |
| 6  | 20006   | LOVEPREET SINGH CHADHA            | sensation_star@gmail.com  | PATIALA           | 8545225689  | 1984-03-18 00:00:00.000 |
| 7  | 20007   | JAISHANKARA PARIMAL CHAKSHU       | Jai.chakshu@gmail.com     | KHATAULI          | 7856231848  | 1982-04-17 00:00:00.000 |
| 8  | 20008   | SATHYANARAYAN SAMAY SINGH         | Sathyas_rules@gmail.com   | MODINAGAR         | 9565622261  | NULL                    |
| 9  | 20009   | GAGANMEET HANSRA                  | gaggub_pranks@gmail.com   | VARANASI          | 9878954655  | 1975-11-10 00:00:00.000 |
| 10 | 20010   | BHAVYA DHINGRA                    | Dhingra.bhavya@gmail.com  | BAREILLY          | 9998884569  | 1988-10-22 00:00:00.000 |
| 11 | 20011   | ABHINAV KUMAR PATNALA             | Teddy_patnala@yahoo.co.in | MYSORE            | 7700598648  | 1979-06-20 00:00:00.000 |
| 12 | 20012   | ABHIJIT RAMCHARANDAS              | charan_me_aapke@gmail.com | PUNE              | 9982485698  | 1987-01-01 00:00:00.000 |
| 13 | 20013   | MUKULIKA SENROY                   | Sweet.muku@gmail.com      | CHENNAI           | 9900665898  | 1984-08-15 00:00:00.000 |
| 14 | 20014   | JAYARAM SAINI                     | Jayaram_s@hotmail.com     | NEW JALPAIGURI    | 9874652365  | NULL                    |
| 15 | 20015   | RANCHHORDAS SHYAMALDAS CHHANCHHAD | wise_brains@gmail.com     | AHMEDABAD         | 9879879876  | 1985-10-15 00:00:00.000 |

**Figure 6.18** A part of the DimStudent table.

|    | Item_ID  | Item_Type | Title                | Technology | No Of Copies | Author        | Publisher          | Period |
|----|----------|-----------|----------------------|------------|--------------|---------------|--------------------|--------|
| 1  | BLRB2001 | Book      | WPF PROGRAMING       | .NET       | 2            | SAMS WILLEY   | WILLEY PUBLICATION | NULL   |
| 2  | BLRB2002 | Book      | C# PROGRAMING        | .NET       | 10           | JACK WILCH    | SAMS PUBLICATION   | NULL   |
| 3  | BLRB2003 | Book      | ADOBE FLEX 2         | JEE        | 10           | MARILDA WHITE | TMH PUBLICATION    | NULL   |
| 4  | BLRB2004 | Book      | THE FINANCE HANDBOOK | GENERAL    | 10           | VIKRAM PANDEY | THM PUBLICATION    | NULL   |
| 5  | BLRB2005 | Book      | SAP R/3 HANDBOOK     | SAP        | 10           | JIMY ARNOLD   | PEARSON EDUCATIKON | NULL   |
| 6  | BLRC2001 | CD        | WPF PROGRAMMING      | .NET       | 2            | NULL          | NULL               | NULL   |
| 7  | BLRC2002 | CD        | C# PROGRAMMING       | .NET       | 10           | NULL          | NULL               | NULL   |
| 8  | BLRC2003 | CD        | SAP CRM CONFIGURATIN | SAP        | 5            | NULL          | NULL               | NULL   |
| 9  | BLRC2004 | CD        | STEP BY STEP SERVLET | JEE        | 10           | NULL          | NULL               | NULL   |
| 10 | BLRC2005 | CD        | ORACLE 10G           | DB         | 10           | NULL          | NULL               | NULL   |
| 11 | BLRM2001 | Magazine  | READERS DIGEST       | GENERAL    | 2            | NULL          | NULL               | MON    |
| 12 | BLRM2002 | Magazine  | VISUAL STUDIO        | .NET       | 3            | NULL          | MON                | MON    |
| 13 | BLRM2003 | Magazine  | JAVA FOR ME          | JEE        | 2            | NULL          | MON                | MON    |
| 14 | BLRM2004 | Magazine  | SQL SERVER           | .NET       | 2            | NULL          | MON                | MON    |
| 15 | BLRM2005 | Magazine  | MSDN                 | .NET       | 1            | NULL          | YRL                | YRL    |

**Figure 6.19** A part of the DimItem table.**Table 6.2** Difference between ER modeling and dimensional modeling

| ER Modeling  | Dimensional Modeling  |
|--|---|
| Optimized for transactional data.                                    | Optimized for query ability and performance.  |
| Eliminates redundant data.   | Does not eliminate redundant data, where appropriate.                                     |
| Highly normalized (even at a logical level).                         | It aggregates most of the attributes and hierarchies of a dimension into a single entity. |
| It is a complex maze of hundreds of entities linked with each other. | It has logical grouped set of Star schemas.   |
| Useful for transactional systems.                                    | Useful for analytical systems.  |
| It is split as per the entities.                                     | It is split as per the dimensions and facts.  |

### ***Relationship Between Dimensional Modeling and Entity Relationship Modeling***

A single entity relationship diagram will be broken down into several fact table diagrams. The following are the steps to convert an entity relationship diagram into a set of dimensional models:

1. An ER diagram of an enterprise represents almost every business process of the enterprise on a single diagram. The first *step* is to separate out the various business processes and represent each as a separate dimensional model.
2. The second step is to constitute the fact tables. To do this, identify all many-to-many relationships in the ER diagram, containing numeric and additive non-key attributes, and construct them into fact tables.
3. The third step is to constitute the dimension tables. To do this, de-normalize all the remaining tables into single-part key tables. These single-part key tables connect to the fact table which essentially is multipart. The fact table generally has two or more foreign keys that refer to the single-part key of the respective dimension tables. In the cases where the same dimension table connects to more than one fact table, the dimension table is represented in each schema, and the dimension tables are referred to as “conformed” between the two-dimensional models.

## **6.11 DATA QUALITY**

Today is a world of heterogeneity. We have different technologies from databases (relational to multidimensional), RFID tags to GPS units, etc. We operate on different platforms – from personal computers to server machines to cell phones, etc. We have several vendors such as Informatica, IBM, Microsoft, etc. to name a few. We have hordes of data being generated every day in all sorts of organizations and enterprises. And we do have problems with data. Data is often duplicated, inconsistent, ambiguous, and incomplete. We do realize that all of this data needs to be collected in one place and cleaned up. Why? Well, the obvious reason is bad data leads to bad decisions and bad decisions lead to bad business.

### **6.11.1 Why Data Quality Matters?**

Let us look at few instances where bad data quality can/has cost a fortune...

- A CD company announced an introductory offer of free CDs. A man defrauded the company by availing of this introductory offer almost 1600 times over. How? Each time he managed to register a different account with a different address by clever use of punctuation marks, slightly different spellings, fictitious apartment numbers, fictitious street numbers, etc.
- Cited here is an instance where a cell phone company has lost out on cross-sell opportunities and has also lost out on potential prospective customers. How? A customer lives with his family (spouse and children) in a separate house, his parents stay in another house at a different address, and his brother is in yet another house at a different physical address. The cell phone company had no way of pooling this information together to relate those individuals. If they had one, they could probably look at cashing in on the power of the customer to influence the buying and holding decisions of someone in one of those other, related households. This brings us to a key question, “How does one define good customer data, and how do you think companies can achieve it?”
- Here is another case, where a cable company lost \$500,000 because a mislabelled shipment resulted in a cable being laid at a wrong location.

- Referring back to the DIIT library System case study, it was seen that the “PhoneNumber” column had certain inconsistencies in the data. Consider a situation, where an application was developed that could fetch a phone number from the database and make an IP phone call to the number from the desktop computer itself. If the “PhoneNumber” column had alphabets or special characters (as was the case in the Excel sheets the library maintained), the application would definitely throw an error. Hence, it became absolutely essential to clean the dirty data of such inconsistencies and populate the fields in the database accordingly, as per practical requirements.

Another field that had a wide difference in the format of date used is the “BirthDate” field. Here, the date of birth of the students had been entered in three different date formats (such as dd-mm-yyyy, dd/mm/yyyy, dd.mm.yyyy). Also, most records had the birth date entered as a *string* data type, while some entries were in a *date* format. Now, in this situation, performing ETL on the source data would result in some records being excluded from the data flow, which would not appear in the database. Hence, it became essential to convert the different formats into a single standard date format and store it in the database as a *date* data type.

### 6.11.2 What is Data Quality?

One of the issues that hinder progress in addressing polluted data is the narrow definition that is applied to the concept of data quality. It has been seen that most attention is paid to data integrity, which has a narrower scope than data quality. Hence, we now need to take a look at the following two definitions:

- Definition of Data Integrity.
- Definition of Data Quality.

#### ***Definition of Data Integrity***

Data integrity reflects the degree to which the attributes of data associated with a certain entity accurately describe the occurrence of that entity. Examples of data integrity:

- **Primary Key:** A column or a collection of columns designated as primary key imposes the “Unique” and “Not Null” constraint.
- **Foreign Key:** A column designated as the foreign key column means it can only have values that are present in the primary key column of the same or different table that it refers to. A foreign key column can have a null or duplicate value.
- **Not Null:** A Not Null constraint on the column means that the column must be given a value. It cannot have a Null value (a missing or unknown value). A space or zero or carriage return or line feed is not a Null value).
- **Check Constraint:** A check constraint allows imposing a business rule on a column or a collection of column.

Let us explain data integrity with an example. We consider here a table, “Employee” having four columns: EmpNo (Primary key), DeptNo (Foreign Key), EmpName (Not Null), and the Age column (with a check constraint imposed on it). The table definition of “Employee” table is as follows:

## Employee

| Column Name | Data Type and Length | Constraint   |
|-------------|----------------------|--|
| EmpNo       | Numeric              | Primary Key ( <b>Entity Integrity Constraint</b> )   |
| DeptNo      | Varchar(5)           | Foreign Key ( <b>Referential Integrity Constraint</b> )<br>DeptNo column of Employee table refers to DeptNo column of Department table |
| EmpName     | Varchar(50)          | Not Null ( <b>Not Null Constraint</b> )  |
| Age         | Numeric              | Age cannot be > 18 or < 65 ( <b>Check Constraint</b> )   |

A column (DeptNo) of the Employee table refers to a column (DeptNo) of the Department table. This is a case of referential integrity constraint. Employee table is the referencing table and Department table is the referenced table. Assume the Department table has the following records in it.

## Department

| DeptNo | DeptName        |
|--------|-----------------|
| D01    | Finance         |
| D02    | Purchase        |
| D03    | Human Resources |
| D04    | Sales           |

Let us take a quick look at the records in the Employee table, particularly the values in the DeptNo column. All the values in the DeptNo column (D01, D02) are present in the DeptNo column of the Department table. Here, the referential integrity constraint is being followed.

## Record Set (Records in the Table, Employee)

| EmpNo | DeptNo | EmpName            | Age |
|-------|--------|--------------------|-----|
| 1001  | D01    | John Mathews       | 25  |
| 1010  | D02    | Elan Hayden        | 27  |
| 1011  | D01    | Jennifer Augustine | 23  |

## Definition of Data Quality

Data quality is measured with reference to appropriateness for purpose as defined by the business users of data and conformance to enterprise data quality standards as formulated by systems architects and administrators. Therefore, it is evident that the concept of data integrity is local to the domain of database technology. However, the concept of data quality has a wider scope and is rooted in the business. It is the latter definition that conveys real efforts in the context of data warehousing.

Data quality is not linear. It is described by several dimensions such as accuracy/correctness, completeness, consistency, and timeliness.

- **Correctness/Accuracy:** Accuracy of data is the degree to which the captured data correctly reflects/describes the real world entity/object or an event. Examples of data accuracy:
  - The address of customer as maintained in the customer database is the real address.
  - The temperature recorded in the thermometer is the real temperature.
  - The air pressure as recorded by the barometer is the real air pressure.
  - The bank balance in the customer's account is the real value customer deserves from the Bank.
  - The age of the patient as maintained in the hospital's database is the real age.

- **Consistency:** This is about the single version of truth. Consistency means data throughout the enterprise should be in sync with each other. Let us look at a few examples of inconsistent data:
  - A customer has cancelled and surrendered his credit card. Yet the card billing status reads as "due".
  - An employee has left the organization. Yet his email address is still active.

Let us look at yet another example. Consider a financial organization, "EasyFinance". This organization has several departments. There are departments that work in silos and there are departments that work in tandem. A customer comes over to deposit a cheque to his account. The cheque is collected by the "cheque collection department". The status of the cheque is designated as "cleared". However, the money is yet to be credited to the account. The reason being, only after the transactions are closed for the day, the processing will begin. For this brief time period, the data is inconsistent.

- **Completeness:** The completeness of data is the extent to which the expected attributes of data are provided. Examples of data completeness are as follows:
  - Data of all students of a University are available.
  - Data of all patients of a hospital are available.
  - All the data of all the clients of an IT organization is available.

Let us look at a case where all the data is not available yet it is considered complete. Example: In the feedback form of a restaurant, a customer provides his postal address (mandatory requirement) but does not mention his email address or his telephone number (optional requirement). Can this data still be taken as complete? The answer is "yes" subjected to the fact if it still meets the expectations of the user (in our example, the restaurant manager). The next question is "Can the data be complete but inaccurate?" The answer is "Yes". In our example, the customer did provide a postal address, so the data is complete because it meets the expectations of the restaurant staff; but the customer could have provided an incorrect address – a typical case of complete yet inaccurate data.

- **Timeliness:** The timeliness of data is extremely crucial. Right data to the right person at the right time is important for business. Delayed supply of data could prove detrimental to the business. No matter how important the data is, if it is not supplied at the right time, it becomes inconsequential and useless. Few examples are as follows:

- The airlines are required to provide the most recent information to their passengers.
- There can be no delay in the publishing of the quarterly results of an enterprise. The results cannot be announced the next year.

An example where data is not timely:

- The population census results are published two years after the census survey is completed.
- **Metadata:** We have discussed few of the dimensions of data quality such as completeness, accuracy, consistency, and timeliness. What we have not touched upon is yet another dimension of data quality and that is “metadata”. We all know metadata is data about data. To quote a few examples: if we consider relational databases, table definitions along with column definitions (data type and length), constraint descriptions, business rules, etc. constitute metadata. Other detailed examples include the conceptual and logical models as discussed in Chapter 7, “Multidimensional Data Modeling”. One key role of metadata that is often overlooked is its role in determining data usage.

A famous quote goes as: **A man with one watch knows what time it is. A man with two watches is never sure.**

And have we witnessed “the two many watches” phenomenon? The answer is yes.

### Picture this...

A manufacturing company has made it mandatory for its employees to attend 5 days of training program every quarter. The count of the number of training days attended by the employees is tracked by the team leader, the training coordinator, and by the individual (employee) who undergoes the training program. Let us take the case of William Welsh, an employee with the finance department. William has just finished his 5 days of training from March 29, 2011 to April 2, 2011. Here is a sample of data that is maintained by the team leader, the training coordinator and by William Welsh.

- *Team Leader:* 3 days for quarter IV (Jan–March) and 2 days for quarter I (Apr–June).
- *Training Coordinator:* 5 days for quarter IV (Jan–March) and zero days for quarter I (Apr–June). The reason given by him is the training commenced on 29 March and therefore counted in the same quarter.
- *William Welsh:* Zero days for quarter IV (Jan–March) and 5 days for quarter I (Apr–June). The reason given by him is the training ended on 2 April and therefore counted in the next quarter.

All the three data are correct but from different perspectives. Clearly, it is a case of “too many watches”. How could this problem be solved? This problem could be resolved by having the metadata clearly state so. Let us look at some typical business questions such as “How many customers do we have?”, “How many alliance partnerships do we have?”, “How many vendor partnerships do we have?”, “How many products do we have?”, “How much expenditure did we incur?”, etc. In the context of these questions, another example of metadata is providing clear definitions of what the terms “customers”, “vendors”, “alliance partners”, “products”, “expenditure”, etc. actually mean.

Let us consider from the business domain perspective. Here, the metadata changes to what they expect in high level reports. Let us look at it in the light of an example. In a transactional system, data is usually maintained to an accuracy of say a precision value of 6 (6 places after decimal) and that makes sense. However, the same data when pulled in a report might be rounded off, first to 2 places after decimal and then completely rounded off to the nearest integer in a still higher level report. However, if there are millions of records, there could be a huge discrepancy in the amount quoted because of this rounding off and data is labelled “poor quality”.

### 6.11.3 How Do We Maintain Data Quality?

From a technical standpoint, data quality results from the process of going through the data and scrubbing it, standardizing it, and de-duplicating records, as well as doing some of the data enrichment.

- Clean up your data by standardizing it using rules.
- Use fancy algorithms to detect duplicates which are obvious by just looking at the strings. For example, “ICS” and “Informatics Computer System” do not look similar. But if they have the same address, same number of employees, etc., then you can say they are the same.
- Let us look at another instance of de-duplication. A retail outlet has several branches in the city. The same customer might happen to visit a few of these. It is very likely that his name is spelled differently while invoicing at the varied branches. Assuming the name of the customer is “Aleck Stevenson”. In one of the invoices, his name appears as “Mr. A. Stevenson”, in another it is spelled as “Mr. Stevenson Aleck”, in another case it is “Mr. S. Aleck”, and so on. Different formats of the name but the same customer. We require an algorithm that could look up another detail of the customer such as the social security number which will be common in all transactions made by the customer and replace the name of the customer with just one consistent format.

#### Inconsistent data concerning Mr. Aleck Stevenson before cleaning up

##### Invoice 1:

| <i>BillNo</i> | <i>CustomerName</i> | <i>SocialSecurityNo</i> | _____ |
|---------------|---------------------|-------------------------|-------|
| 101           | Mr. Aleck Stevenson | ADWPS10017              | _____ |

##### Invoice 2:

| <i>BillNo</i> | <i>CustomerName</i> | <i>SocialSecurityNo</i> | _____ |
|---------------|---------------------|-------------------------|-------|
| 205           | Mr. S. Aleck        | ADWPS10017              | _____ |

##### Invoice 3:

| <i>BillNo</i> | <i>CustomerName</i> | <i>SocialSecurityNo</i> | _____ |
|---------------|---------------------|-------------------------|-------|
| 314           | Mr. Stevenson Aleck | ADWPS10017              | _____ |

##### Invoice 4:

| <i>BillNo</i> | <i>CustomerName</i> | <i>SocialSecurityNo</i> | _____ |
|---------------|---------------------|-------------------------|-------|
| 316           | Mr. Alec Stevenson  | ADWPS10017              | _____ |

## Consistent data after cleaning up

### Invoice 1:

| <i>BillNo</i> | <i>CustomerName</i> | <i>SocialSecurityNo</i> | _____ |
|---------------|---------------------|-------------------------|-------|
| 101           | Mr. Aleck Stevenson | ADWPS10017              | _____ |

### Invoice 2:

| <i>BillNo</i> | <i>CustomerName</i> | <i>SocialSecurityNo</i> | _____ |
|---------------|---------------------|-------------------------|-------|
| 205           | Mr. Aleck Stevenson | ADWPS10017              | _____ |

### Invoice 3:

| <i>BillNo</i> | <i>CustomerName</i> | <i>SocialSecurityNo</i> | _____ |
|---------------|---------------------|-------------------------|-------|
| 314           | Mr. Aleck Stevenson | ADWPS10017              | _____ |

### Invoice 4:

| <i>BillNo</i> | <i>CustomerName</i> | <i>SocialSecurityNo</i> | _____ |
|---------------|---------------------|-------------------------|-------|
| 316           | Mr. Aleck Stevenson | ADWPS10017              | _____ |

- We can also look at using external data to clean up the data for de-duplication. For example, US Postal Service publishes a CD of every valid address in the US. One easy step to major de-duplication is for businesses to buy that CD and use that to convert all their address data to this standard format.

### 6.11.4 Key Areas of Study

Let us look at a few key areas of study in this area:

- **Data governance:** Good decisions are based on quality data. Data governance is a quality regime that includes ensuring accuracy, consistency, completeness, and accountability of data. There are policies that govern the use of data in an organization. This is done primarily to secure data from hackers and data from inadvertently leaking out. Data governance also ensures compliance with regulations.
- **Metadata management:** Metadata, previously called Data Dictionary, is a collection of definitions and relationships that describe the information stored. It is data about data. The storage and access of metadata information is as important today as it was earlier.
- **Data architecture and design:** Overall architecture – data storage, ETL process design, BI architecture, etc.
- **Database management:** This includes optimizing the performance of the database, backup, recovery, integrity management, etc.
- **Data security:** Who should have access? Which data needs to be kept private?

- **Data quality:** Lots of work is needed to ensure that there is a single version of truth in the data warehouse. Especially difficult for non-transactional data (i.e., data that is not there in a database). For example, Ashwin Shrivastava is the same as A. Shrivastava. Need fancy software that will do these transformations on the data.
- **Data warehousing and business intelligence:** To measure, monitor, and manage performance: “you cannot manage what you cannot measure and you cannot measure what you cannot define”.

## UNSOLVED EXERCISES

1. What is data quality?
2. Why is data quality important?
3. How do we maintain data quality? Explain giving examples.
4. How is data integrity different from data quality? Explain.
5. Data quality is defined by several dimensions like accuracy/correctness, timeliness, consistency, and completeness. Explain giving examples.
6. Data can be accurate yet inconsistent. Explain with an example.
7. Data is complete yet inaccurate. Explain with an example.
8. Why is the timeliness of data so important? Comment.
9. Is metadata a dimension of data quality? Comment.
10. Explain entity integrity and referential integrity constraint using an example.
11. Write about a situation from your life where you feel that data quality has been compromised.

## 6.12 DATA PROFILING

### 6.12.1 The Context

*A few questions...*

If you have ever been a part of the process of data warehouse design, has this happened to you?

- You were in a great hurry to start your project and create your database, and populate it with data; however, you later realized that there were several data inconsistencies in the source, like missing records or NULL values.
- Or, the column you chose to be the primary key column is not unique throughout the table.
- Or, you realized that the schema design was not coherent to the end user-requirements.
- Or, any other concern with the data or the database/data warehouse that made you wish you had fixed that concern right at the beginning.

Well, you are not alone! Every day, many database professionals (novices and experienced people alike) face such situations. Every single day, these people encounter many problems and issues with the quality of the data they are populating in the database/data warehouse.

To fix such data quality issues would mean making changes in your ETL data flow packages, cleaning the identified inconsistencies, etc. This in turn will lead to a lot of re-work to be done. Re-work will mean added costs to the company, both in terms of time and effort. What would you do in such a case? Well, analyze the data source more closely, of course! Also, you would want to understand the business rules and user requirements better, or know the final purpose for which the database will be used for.

### 6.12.2 The Problem

In case the quality of the data at the source is not good, the inconsistencies and errors tend to amplify themselves over the entire process of creating ETL packages, creating the data warehouse, and finally using that data warehouse for analysis or reporting purposes.

From the business point of view, one cannot expect the final data warehouse to deliver the right information in this case. As we are all aware, right information delivery is extremely important for making wise, informed decisions and hence meeting expected goals and achieving quality results.

From the developer's point of view, incorrect information resulting from poor quality data will mean *regressively analysing the source data for drawbacks and inconsistencies*. This means a huge amount of re-work to be done. Re-work would mean *more time consumed* and *more effort* required to be put into analysing what went wrong, then editing or re-creating ETL packages, and making suitable changes in the data warehouse. Certainly, all this is something we can do without!

| Phone Number |
|--------------|
| 9952392126   |
| 9555533325   |
| 9554477886   |
| 9325489654   |
| 9654238552   |
| (8545)225689 |
| 78562 31848  |
| 9565622261   |
| 98789-54655  |
| 99988-84569  |
| 7700598648   |
| 9982485698   |
| 9900665898   |
| 9874652365   |
| 9879879876   |

Consider the case of the library system of DIIT. The source data that was maintained in Excel sheets and Access database had a lot of inconsistencies that were pointed out by Prof. Frank. One of these inconsistencies was the “PhoneNumber” column in the “Student” table.

This column has certain entries that have special characters, like circular brackets () . Suppose, all these entries were to pass on as they are, into the database, without cleaning them of the special characters and spaces, etc. and imagine if an application used this database to refer to the phone number of a particular student and attempted to dial directly to that number. Since any of these numbers could be a string value with special characters, the application will fail to do the required job. In this case, it is of utmost importance that this column be maintained as a purely numerical (integer) column, cleansed of all non-integral characters.

Now, it is required of an ETL developer to keep a suspiciously vigilant eye for such errors. There may be thousands and lakhs of records in a table, and there may be very small, unnoticeable errors like these in the table. So, one cannot just proceed with performing ETL without looking out for such errors or inconsistencies and figuring how to deal with them.

### 6.12.3 The Solution

Instead of a solution to the problem, would it not be better to catch it right at the start before it becomes a problem? After all, we are all too familiar with the age-old saying: “Prevention is better than cure!”

In fact, this has been the tone of database designers, developers and administrators since the time databases came into being. However, with the exponential rise in the volume of data being stored, studying and analyzing data in the source manually, became a tedious task. Hence **data profiling software** came to the rescue!

As per Kimball, data profiling software is now used not only for initial analysis of the data source, but throughout the life-cycle of data warehouse development. They have proved to be of great help in identifying data quality issues with the source as well as in looking out for errors that crept in during the data integration phase.

### 6.12.4 What is Data Profiling?

Data profiling is the process of statistically examining and analysing the content in a data source, and hence collecting information about that data. It consists of techniques used to analyze the data we have for accuracy and completeness.

- Data profiling helps us make a thorough assessment of data quality.
- It assists the discovery of anomalies in data.
- It also helps us understand content, structure, relationships, etc. about the data in the data source we are analyzing.
- It helps us know whether the existing data can be applied to other areas or purposes too.
- It helps us understand the various issues/challenges we may face in a database project much before the actual work begins. This enables us to make early decisions and act accordingly.
- It helps add quality to data by converting from the format in which it is stored (i.e., assigning more meaning to it), or categorizing it.
- It helps assess the risks associated with integrating data with other applications.
- Data profiling is also used to assess and validate metadata. This helps determine if the source is suitable enough for the intended purpose; and if so, then it is used to identify initial problems so that an appropriate solution may be designed for them.

According to Brian Marshall, data profiling can be either *data quality profiling* or *database profiling*.

- **Data quality profiling:** It refers to analyzing the data from a data source or database against certain specified business rules or requirements. This enables us identify and thus assess the major problems/issues associated with the quality of the data being analyzed. The analysis can be represented as *summaries* or *details*.
  - *Summaries:* Counts and percentages that give information on the completeness of data sets, the problem distribution in a data set, the uniqueness of a column, etc.
  - *Details:* Involves lists that contain information about missing data records or data problems in individual records, etc.
- **Database profiling:** It refers to the procedure of analysis of a database, with respect to its schema structure, relationships between tables, columns used, data-type of columns, keys of the tables, etc.

Generally, database profiling is the initial step that defines the data quality rules and lays the basic groundwork for the task of data quality profiling.

### 6.12.5 When and How to Conduct Data Profiling?

It is a good practice to schedule profiling tasks right before going into the details of designing the system/warehouse (though profiling is also conducted throughout the data warehousing process). Generally, data profiling is conducted in two ways:

- Writing SQL queries on sample data extracts put into a database.
- Using data profiling tools.

#### *When to Conduct Data Profiling?*

- **At the discovery/requirements gathering phase:** As soon as source data systems have been identified and business requirements have been laid out, an initial data profiling exercise must take place to examine the source data for anomalies, or any other such drawbacks that do not

concur with the given business requirements, or are detrimental to the designing of the data warehouse. This phase involves more of data quality profiling. It helps identify most potential issues at an early stage, and hence avoid corrections and re-work.

- **Just before the dimensional modeling process:** This is the stage where intensive data profiling is done. This stage involves more of database profiling, and also some of data quality profiling. In this phase, we analyze the most appropriate schema design for the dimensional data warehouse, and decide on the best method we can employ for conversion of the source data system to the dimensional model.
- **During ETL package design:** Certain data profiling analyses can be done during the ETL process too. The advantage is: it will help us identify possible errors and problems that may creep into the data due to ETL transformations applied on it. Data profiling during ETL also enables us identify what data to extract and what filters to apply, besides giving information on whether transformations have been applied correctly or not. This phase hence involves more of data quality profiling.

### ***How to Conduct Data Profiling?***

As mentioned before, data profiling involves statistical analysis of the data at source and the data being loaded, as well as analysis of metadata. These statistics may be used for various analysis purposes. The most common examples of analyses to be done are as under:

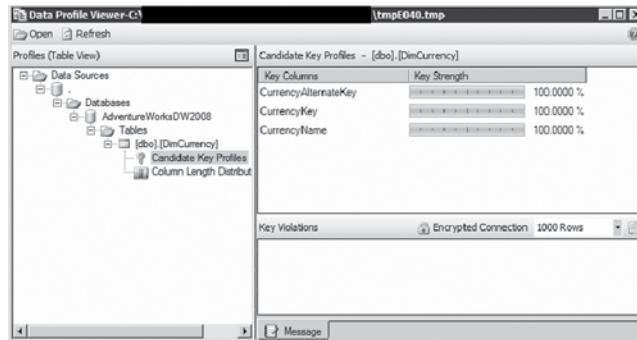
- **Data quality:** Analyze the quality of data at the data source. A column containing telephone numbers may be alpha-numeric in nature. Hence, the non-numeric characters need to be removed.
- **NULL values:** Look out for the number of NULL values in an attribute.
- **Candidate keys:** Analysis of the extent to which certain columns are distinct (percentage-wise) will give the developer useful information with respect to selection of candidate keys.
- **Primary key selection:** To check whether the candidate key column does not violate the basic requirements of not having NULL values or duplicate values.
- **Empty string values:** A string column may contain NULL or even empty string values. During cube analysis, these empty string or NULL values may create problems later. For example, if a cube is created in Microsoft SSAS, columns containing empty string values cause the cube to fail deployment.
- **String length:** An analysis of the largest and shortest possible length as well as the average string length of a string-type column can help us decide what data type would be most suitable for the said column.
- **Numeric length and type:** Similarly, assessing the maximum and minimum possible values of a numeric column can help us decide what numeric data type would be best suited for that column. Also, if it does not have decimal values, one should consider declaring it as an integer-type column instead of float-type.
- **Identification of cardinality:** The cardinality relationships (like one-to-one, one-to-many, and many-to-many) are important for inner and outer join considerations with regard to several BI tools. They are also important for the design of the fact-dimension relationships in the data warehouse.
- **Data format:** Sometimes, the format in which certain data is written in some columns may not be user-friendly. For example, if there is a string column that stores the marital status of a person in the form of “M” for “Married” and “U” for “Unmarried”, we might consider changing the short forms to the full lengthened forms (i.e., change to “Married” and “Unmarried”, respectively) in the data warehouse, since they are easier to comprehend.

... and many more cases like these.

### **Common Data Profiling Software**

Most of the data-integration/analysis softwares have data profiling functions built into them. Alternatively, various independent data profiling tools are also available. Some popular ones (as shown on [www.topbits.com](http://www.topbits.com)) have been listed below:

- **Trillium Enterprise Data Quality:** Powerful, yet very user-friendly software.
  - It scans all the data systems you require it to manage.
  - Automatically runs continual scans periodically to check whether all the data is consistently updated.
  - Removes duplicate records.
  - Provides for the separation of data into categories to allow easier data management.
  - Generates statistical reports about the data systems regularly.
- **Datiris Profiler:** This tool is very flexible and can manage your data without inputs from the user. Certain defaults can be set, and the tool manages your data automatically. It has many powerful features (listed below) that set it at a level higher than other profiling tools:
  - A powerful metric system.
  - Very good compatibility with other applications.
  - Domain validation.
  - Command-line interface.
  - Pattern analysis.
  - Real time data viewing.
  - Batch profiling.
  - Conditional profiling.
  - Profile templates and spreadsheets.
  - Data definitions.
- **Talend Data Profiler:** It is a free, open-source software solution to data profiling, which is now slowly becoming popular. It may not be as powerful as Datiris or other paid profilers, but it is good enough for small businesses and non-profit organizations.
- **IBM Infosphere Information Analyzer:** A powerful profiling tool developed by IBM, it does a deep-scan of the system in a very short time-period. Various other software features integrated with this tool are
  - IBM Infosphere security framework.
  - Scanning scheduler.
  - Reports.
  - Source system profiling and analysis.
  - Rules analysis.
  - User annotations.
  - Consistent common metadata across all other IBM Infosphere products.
- **SSIS Data Profiling Task:** This data profiling tool is not an independent software tool. It is integrated into the ETL software called SQL Server Integration Services (SSIS) provided by Microsoft. It can provide useful statistics about the source data as well as the transformed data that is being loaded into the destination system.
- **Oracle Warehouse Builder:** Oracle warehouse builder is not strictly a data-profiling software tool. It has the necessary functionality to let a person with zero programming knowledge build



a data warehouse from scratch. One of its features is the data profiling functionality which helps analyze source systems and hence provide “clean” data.

## SUMMARY

We have come to the end of this chapter, and by now we can appreciate (to a certain extent) the basic concepts of data integration, the ETL process, issues and concerns about data quality, and the enormous benefits of profiling data intensively at every stage of the data warehousing life-cycle.

Given below is the list of major steps that are undertaken during the data warehousing process:

- Identify data sources.
- Design schema for data warehouse.
- Profile the source data.
- Formulate business rules and strategy for data integration.
- Perform ETL and populate the data into the data warehouse.
- Perform tests and checks (validation) on the transported data and the related metadata.

The major advantages of data profiling are listed below:

- Potential threats, errors, or inconsistencies can be detected quite early at the start of the project life-cycle.
- It saves time and effort.
- It increases the quality of data, hence the information derived is of the highest quality too.
- Scheduling data profiling tasks periodically will help detect and eliminate erroneous missing or duplicate data immediately, thus maintaining a high level of accuracy and consistency.
- Helps add deeper meaning to data.

Even though data profiling may seem to be a tedious activity, it has benefits that far outweigh the efforts put into profiling. It adds value to the database and creates great value for the business as well.

## A CASE STUDY FROM THE HEALTHCARE DOMAIN

“HealthyYou” is a home healthcare products and services company of the “GoodLife HealthCare Group” that provides various easy-to-use at home, medical products for patients suffering from various

ailments, delivered at their doorstep. This company provides three types of medical products categorized as Disposable-Items (cotton pack, syringe, saline bottle, hand gloves, bandage, etc.), Reusable-Items (wheelchair, walker, donut cushion, thermometer, blood glucose meter, adjustable beds, etc.), and Medicines (Avayatin, Aquasure, Acerint, Instadril, Antidrome, Mutacine, etc.). There is a department for each category such as DisposableItem (DI), ReusableItem (RI), and Medicines (Med) department. "HealthyYou" has expansion plans and is likely to have branches in different locations nationwide.

Rita (receptionist) sits at the reception desk and attends to calls and makes note of the products required/requested/ordered by the patients. The customer can either place an order on call or can even come down to the company's center and purchase the products. If an order is made on call then payment has to be made online. However, cash/debit card payment is accepted when one directly purchases at the center. Data is stored and maintained by each department in MS Excel spreadsheet. Rita, the receptionist, maintains the details of callers/patients in an MS Access Database. However, the orders placed by the caller/patient are maintained by her in an MS Excel spreadsheet. The management of the company is in need of a report that will indicate the products that are in demand and are sold off-the-shelf very quickly. This is because the company would like to optimize their manufacturing processes to produce products that are in demand in larger volumes. Likewise they would want to discontinue the manufacturing of products which are no longer in demand. A report is also required that will show the annual revenue generated by the sale of products of all categories. This report should further drill down to the earnings on each item/product. The data is stored in different formats and to get a unified view of the data, a lot of work will be required. Robert (Manager) was called upon to suggest a possible solution to the problem. In Robert's opinion, it would be better if the company starts to archive the data in a data warehouse/data mart. The arguments put forth by him in favour of the company's data warehouse were:

- Data from several heterogeneous data sources (MS Excel spreadsheets, MS Access database, .CSV file, etc.) can be extracted and brought together in a data warehouse.
- Even when the company expands into several branches in multiple cities nationwide, it still can have one data warehouse to support the information needs of the company.
- Data anomalies can be corrected through an ETL package.
- Missing or incomplete records can be detected and duly corrected.
- Transformations can be made to convert the data into an appropriate format.
- Duplicate data can be removed.
- Uniformity can be maintained over each attribute of a table.
- Retrieval of data for analysis and reports can be done conveniently (like the report requested above).
- Fact-based decision making can be easily supported by a data warehouse.
- Ad hoc queries can be easily supported.

Let us assume that all the categories [DisposableItem, ReusableItem, Medicines, Purchase (showing the products purchased by the patients)] store the data in separate sheets in an Excel workbook, except the Patient table which is in an Access database. Each sheet contains a table. There are five tables in all:

- Patient (Access)
- Disposable (Excel)
- Reusable (Excel)
- Medicines (Excel)
- Purchase (Excel)

Previews of each of these tables are as follows:

## Patient

|    | A          | B                                 | C         | D             | E      | F              |
|----|------------|-----------------------------------|-----------|---------------|--------|----------------|
| 1  | Patient ID | Patient Name                      | Disease   | Date of Birth | Gender | Phone Number   |
| 2  | 10001      | A.V.Rao                           | Anemia    | 8/21/1971     | M      | 030-0074321    |
| 3  | 10002      | Raman Chandrashekhar              | Diabetes  | 3/3/1966      | M      | (5) 555-4729   |
| 4  | 10003      | Singhania M K                     | Arthritis | 4.4.1956      | M      | (5) 555-3932   |
| 5  | 10004      | Tarannum Naaz                     | Cough     | 6/11/1989     | F      | (171) 555-7788 |
| 6  | 10005      | Albert Goldberg                   | Fever     | 1/1/1989      | M      | 0921-12 34 65  |
| 7  | 10006      | Yogeshwar Bachupally              | Polio     | 3/18/2001     | M      | 0621-08460     |
| 8  | 10007      | Nikhil Shirish Agate              | Arthritis | 4/4/1962      | M      | 88.60.15.31    |
| 9  | 10008      | Charles Winslet                   | Fever     | 4.12.1994     | M      | (91) 555 22 82 |
| 10 | 10009      | Charlotte Beckinsale              | Diabetes  | 11/10/1968    | M      | 91.24.45.40    |
| 11 | 10010      | Niharika Tyagi                    | Body Ache | 10/10/1988    | F      | (604) 555-4729 |
| 12 | 10011      | Ankita Dhananjay Karandikar       | Cough     | 9/6/1963      | F      | (171) 555-1212 |
| 13 | 10012      | Bhuvan Vishwanath Deshpande       | Body Ache | 11.1.1982     | M      | (1) 135-5555   |
| 14 | 10013      | Allison Williams                  | Diabetes  | 7/8/1964      | M      | (5) 555-3392   |
| 15 | 10014      | N. Sirisha Naidu Peddinti         | Fracture  | 5/8/1988      | M      | 0452-076545    |
| 16 | 10015      | Sandeep Maroli Kini               | Diabetes  | 11.10.1953    | M      | (11) 555-7647  |
| 17 | 10016      | Manjunath Nagappa Krishna Murthty | Cough     | 1/4/1952      | M      | (171) 555-2282 |
| 18 | 10017      | Sophie Brown                      | Arthritis | 4/12/1954     | F      | 0241-039123    |
| 19 | 10018      | Hima Bindu Venkata Neelamraju     | Cold      | 11/10/1977    | F      | 40.67.88.88    |
| 20 | 10019      | Phunsuk Wangdu                    | Polio     | 2/10/1999     | M      | (171) 555-0297 |
| 21 | 10020      | Priyanka Badjatiya                | Arthritis | 12/6/1945     | F      | 7675-3425      |
| 22 | 10021      | Ho-Shi Zong                       | Fracture  | 1/1/1976      | M      | (11)555-9857   |

## Disposable

|   | A                  | B                    | C                |
|---|--------------------|----------------------|------------------|
| 1 | Disposable Item Id | Disposable Item Name | Unit Price (INR) |
| 2 | D001               | Cotton Pack          | Rs.45            |
| 3 | D002               | Syringe              | Rs.101.25        |
| 4 | D003               | Saline bottle        | Rs.330.75        |
| 5 | D004               | Hand gloves          | Rs.67.5          |
| 6 | D005               | Bandage              | Rs.134.10        |

## Reusable

|   | A                | B                   | C                |
|---|------------------|---------------------|------------------|
| 1 | Reusable Item Id | Reusable Item Name  | Unit Price (INR) |
| 2 | R001             | Wheel Chair         | Rs.6727.5        |
| 3 | R002             | Walker              | Rs.2250          |
| 4 | R003             | Adjustable Bed      | Rs.100400        |
| 5 | R004             | Donut Cushion       | Rs.855           |
| 6 | R005             | Thermometer         | Rs.506.25        |
| 7 | R006             | Blood Glucose Meter | Rs.1125          |

## Medicines

|   | A          | B            | C         | D              | E          |
|---|------------|--------------|-----------|----------------|------------|
| 1 | MedicineId | MedicineName | Disease   | UnitPrice(INR) | ExpiryDate |
| 2 | M001       | Avayatin     | Anemia    | Rs.123.75      | 10/20/2014 |
| 3 | M002       | Aquasure     | Diabetes  | Rs.90          | 9/25/2012  |
| 4 | M003       | Acerint      | Arthritis | Rs.94.5        | 2/4/2016   |
| 5 | M004       | Instadril    | Cough     | Rs.49.5        | 9/5/2014   |
| 6 | M005       | Antidrome    | Fever     | Rs.54          | 7/2/2013   |
| 7 | M006       | Bendrigin    | Polio     | Rs.81          | 8/3/2015   |
| 8 | M007       | Nitacine     | Body Ache | Rs.63          | 4/5/2016   |
| 9 | M008       | Mutacine     | Fracture  | Rs.126         | 8/6/2012   |

## Purchase

|    | A      | B      | C         | D               | E            | F            |
|----|--------|--------|-----------|-----------------|--------------|--------------|
| 1  | BillId | ItemId | PatientId | QuantityOrdered | PurchaseDate | Payment Mode |
| 2  | B001   | D001   | 10021     | 8               | 29/09/2010   | Cash         |
| 3  | B002   | D002   | 10009     | 3               | 20/10/2010   | Card         |
| 4  | B003   | D002   | 10002     | 4               | 29/10/2010   | Online       |
| 5  | B004   | D002   | 10015     | 2               | 30/10/2010   | Cash         |
| 6  | B005   | D002   | 10019     | 7               | 30/10/2010   | Card         |
| 7  | B006   | D002   | 10007     | 5               | 23/11/2010   | Online       |
| 8  | B007   | D002   | 10001     | 6               | 24/11/2010   | Cash         |
| 9  | B008   | D002   | 10009     | 5               | 22/11/2010   | Card         |
| 10 | B009   | D002   | 10017     | 4               | 30/11/2010   | Online       |
| 11 | B010   | D003   | 10001     | 5               | 20/12/2010   | Cash         |
| 12 | B011   | D003   | 10001     | 5               | 13/12/2010   | Card         |
| 13 | B012   | D004   | 10021     | 2               | 14/12/2010   | Online       |
| 14 | B013   | D005   | 10014     | 3               | 19/12/2010   | Cash         |
| 15 | B014   | M001   | 10001     | 12              | 31/12/2010   | Card         |
| 16 | B015   | M001   | 10001     | 12              | 15/1/2011    | Online       |
| 17 | B016   | M002   | 10002     | 10              | 21/1/2011    | Cash         |
| 18 | B017   | M003   | 10003     | 10              | 16/1/2011    | Card         |
| 19 | B018   | M006   | 10006     | 14              | 18/1/2011    | Online       |
| 20 | B019   | M008   | 10014     | 12              | 18/1/2011    | Cash         |
| 21 | B020   | R001   | 10014     | 1               | 19/1/2011    | Card         |
| 22 | B021   | R002   | 10017     | 1               | 17/01/2011   | Online       |
| 23 | B022   | R003   | 10001     | 1               | 20/01/2011   | Cash         |
| 24 | B023   | R004   | 10021     | 2               | 21/01/2011   | Card         |
| 25 | B024   | R005   | 10008     | 1               | 25/01/2011   | Online       |
| 26 | B025   | M005   | 10005     | 16              | 31/01/2011   | Card         |
| 27 | B026   | M004   | 10016     | 12              | 13/2/2011    | Card         |
| 28 | B027   | M007   | 10012     | 11              | 15/2/2011    | Card         |
| 29 | B028   | M004   | 10014     | 15              | 19/2/2011    | Cash         |
| 30 | B029   | M004   | 10011     | 16              | 18/2/2011    | Online       |

(Continued)

(Continued)

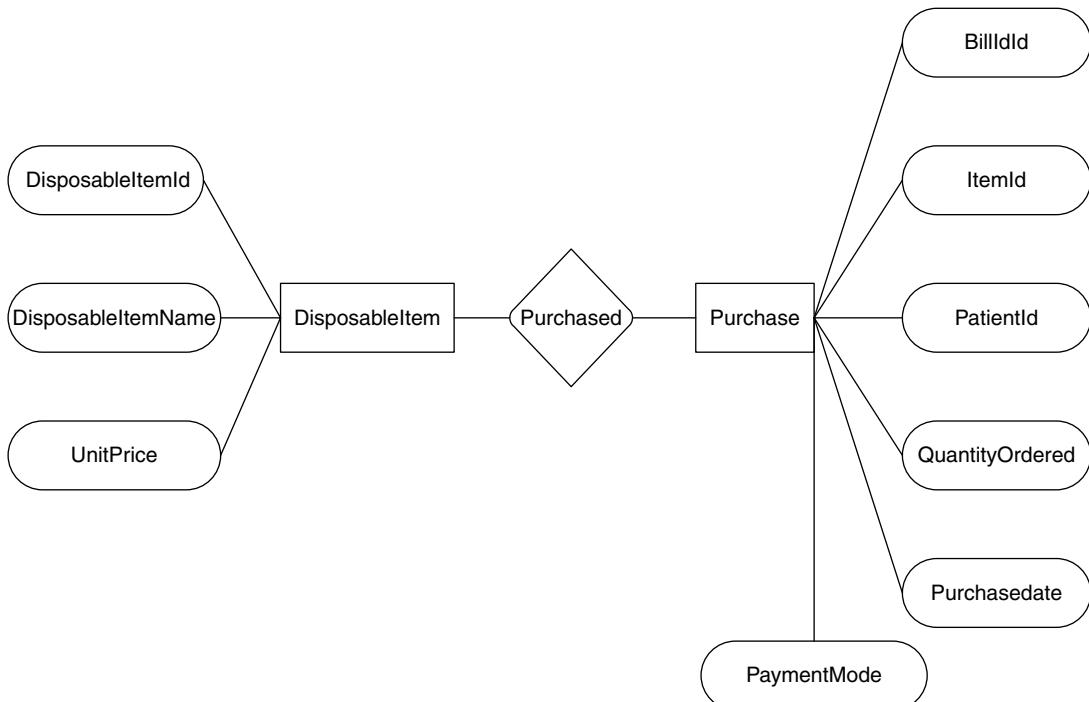
|    |      |      |       |  |    |           |        |
|----|------|------|-------|--|----|-----------|--------|
| 31 | B030 | D001 | 10014 |  | 12 | 16/2/2011 | Cash   |
| 32 | B031 | M004 | 10011 |  | 16 | 18/2/2011 | Online |
| 33 | B032 | D001 | 10014 |  | 12 | 16/2/2011 | Cash   |

The **Purchase** table keeps records of all transactions (purchases) made from the company by customers (patients). “Patients” data is stored in the **Patient** table. The rest of the three tables contain data about the products (**DisposableItem**, **ReusableItem**, **Medicines**) that can be purchased from the company. The requirements, as stated by Robert, are as follows:

- Data in all tables should be in proper format and must be consistent.
- UnitPrice column should be in a uniform format, preferably as a numeric data type.
- Date of birth should be in a uniform format (dd-mm-yyyy).
- Date columns should be in a uniform format, preferably as a datetime data type.
- The three tables – **DisposableItem**, **ReusableItem**, and **Medicines** – shall be combined into a single table that contains details of all these three categories of products.
- Extra white spaces should be removed from the product names.

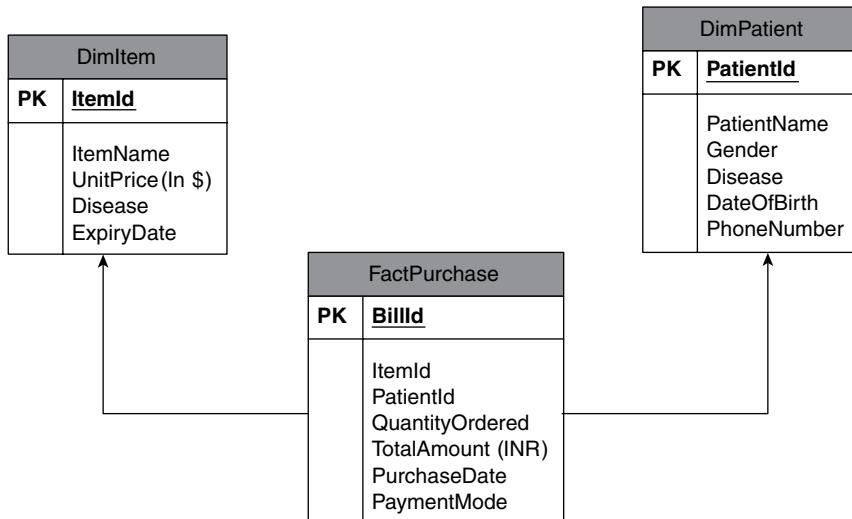
Let us look at the ER model for the above-stated scenario. Here, we consider two entities: **DisposableItem** and **Purchase**.

### Entity Relationship Model



Let us draw the dimensional model for this scenario.

### Dimensional Model



The above relational schema could be implemented as an SQL Server database. Now, the major steps that were taken to transfer data from the Excel spreadsheets and Access database to SQL Server database were as follows (as done in the last example):

1. Profiling the source data (identifying natural key columns, identifying data anomalies, corrections to be made, etc.)
2. Results of profiling were:
  - a. Identification of natural keys (that would naturally be **unique** and **not null**):
    - *Patient*: PatientId
    - *DisposableItem*: DisposableItemId
    - *ReusableItem*: ReusableItemId
    - *Medicines*: MedicineId
    - *Purchase*: BillId
  - b. Removal of leading and trailing blanks wherever they occur.
  - c. Removal of special character(Rs.) from the “UnitPrice” column.
  - d. Removal of special character(.,,(,)) from the “PhoneNumber” column.
  - e. Transforming the “DateOfBirth” column to a standard date format (dd-mm-yyyy).
  - f. Deriving a column “BillAmount” in for the Purchase table which is the multiplication of UnitPrice and QuantityOrdered.
  - g. The Gender column should contain “Male” and “Female” instead of “M” and “F,” respectively.
  - h. Removal of duplicate rows from the Purchase table.
3. Choose an appropriate ETL tool (SSIS, Kettle, Informatica, etc.) and create ETL packages to transform and transport the data from the source database to the destination database.
4. Some of the major rules, transformations and data-cleaning activities that were identified and implemented were:
  - a. Merging together the tables **DisposableItem**, **ReusableItem**, and **Medicines** into a single table “**DimItem**”.

- b. Data-type conversion.
- c. Removal of special character (Rs.) from the “UnitPrice” column.
- d. Creating a new column “TotalAmount” in FactPurchase. TotalAmount would be calculated as

$$\text{(QuantityOrdered)} \times \text{UnitPrice (INR)}$$

- e. Removing leading and trailing blanks wherever they occur.
  - f. Removal of special characters from the “PhoneNumber” (-,(,),.) column and its subsequent conversion to numerical format.
  - g. Removal of special characters from the “UnitPrice” (Rs.) column and its subsequent conversion to numerical format.
  - h. Standardizing the “BirthDate” column format (dd-mm-yyyy) and conversion to “Date-Time” datatype.
  - i. Removal of duplicate rows from the Purchase table.
  - j. Changing the values of the Gender column from “M” and “F” to “Male” and “Female,” respectively.
5. As per general practice, the Dimension tables (DimItem and DimPatient) are populated with data first, followed by the Fact table(s) (FactPurchase). This is because of referential constraints of fact on the dimensions.

After performing ETL, the tables in the database looked like the diagrams below:

### **DimItem**

|    | ItemId | ItemName            | UnitPrice(INR) | Disease   | ExpiryDate              |
|----|--------|---------------------|----------------|-----------|-------------------------|
| 1  | D001   | Cotton Pack         | 45.00          | NA        | NULL                    |
| 2  | D002   | Syringe             | 101.25         | NA        | NULL                    |
| 3  | D003   | Saline bottle       | 330.75         | NA        | NULL                    |
| 4  | D004   | Hand Gloves         | 67.50          | NA        | NULL                    |
| 5  | D005   | Bandage             | 134.10         | NA        | NULL                    |
| 6  | M001   | Avayatin            | 123.75         | Anemia    | 2014-10-20 00:00:00.000 |
| 7  | M002   | Aquasure            | 90.00          | Diabetes  | 2012-09-25 00:00:00.000 |
| 8  | M003   | Acerint             | 94.50          | Arthritis | 2016-02-04 00:00:00.000 |
| 9  | M004   | Instadril           | 49.50          | Cough     | 2014-09-05 00:00:00.000 |
| 10 | M005   | Antidrome           | 54.00          | Fever     | 2013-07-02 00:00:00.000 |
| 11 | M006   | Bendrigiin          | 81.00          | Polio     | 2015-08-03 00:00:00.000 |
| 12 | M007   | Nitacine            | 63.00          | Body Ache | 2016-04-05 00:00:00.000 |
| 13 | M008   | Mutacine            | 126.00         | Fracture  | 2012-08-06 00:00:00.000 |
| 14 | R001   | Wheel Chair         | 6727.50        | NA        | NULL                    |
| 15 | R002   | Walker              | 2250.00        | NA        | NULL                    |
| 16 | R003   | Adjustable Bed      | 100440.00      | NA        | NULL                    |
| 17 | R004   | Donut Cushion       | 855.00         | NA        | NULL                    |
| 18 | R005   | Thermometer         | 506.25         | NA        | NULL                    |
| 19 | R006   | Blood Glucose Meter | 1125.00        | NA        | NULL                    |

### DimPatient

|    | PatientId | PatientName             | Gender | Disease   | DateOfBirth             | PhoneNumber |
|----|-----------|-------------------------|--------|-----------|-------------------------|-------------|
| 1  | 10001     | A.V. Rao                | Male   | Anemia    | 1971-08-21 00:00:00.000 | 0300074321  |
| 2  | 10002     | Raman Chandrashekhar    | Male   | Diabetes  | 1966-03-03 00:00:00.000 | 55554729    |
| 3  | 10003     | Singhania M K           | Male   | Arthritis | 1956-04-04 00:00:00.000 | 55553932    |
| 4  | 10004     | Tarannum Naaz           | Female | Cough     | 1989-06-11 00:00:00.000 | 1715557788  |
| 5  | 10005     | Albert Goldberg         | Male   | Fever     | 1989-01-01 00:00:00.000 | 0921123465  |
| 6  | 10006     | Yogeshwar Bachupally    | Male   | Polio     | 2001-03-18 00:00:00.000 | 062108460   |
| 7  | 10007     | Nikhil Shirish Agate    | Male   | Arthritis | 1962-04-04 00:00:00.000 | 88601531    |
| 8  | 10008     | Charles Winslet         | Male   | Fever     | 1994-04-12 00:00:00.000 | 915552282   |
| 9  | 10009     | Charlotte Beckinsale    | Male   | Diabetes  | 1968-11-10 00:00:00.000 | 91244540    |
| 10 | 10010     | Niharika Tyagi          | Female | Body A... | 1988-10-10 00:00:00.000 | 6045554729  |
| 11 | 10011     | Ankita Dhananjay Ka...  | Female | Cough     | 1963-09-06 00:00:00.000 | 1715551212  |
| 12 | 10012     | Bhuvan Vishwanath...    | Male   | Body A... | 1982-11-01 00:00:00.000 | 11355555    |
| 13 | 10013     | Allison Williams        | Male   | Diabetes  | 1964-07-08 00:00:00.000 | 55553392    |
| 14 | 10014     | N. Sirisha Naidu Ped... | Male   | Fracture  | 1988-05-08 00:00:00.000 | 0452076545  |
| 15 | 10015     | Sandeep Maroli Kini     | Male   | Diabetes  | 1953-11-10 00:00:00.000 | 115557647   |
| 16 | 10016     | Manjunath Nagappa...    | Male   | Cough     | 1952-01-04 00:00:00.000 | 1715552282  |
| 17 | 10017     | Sophie Brown            | Female | Arthritis | 1954-04-12 00:00:00.000 | 0241039123  |
| 18 | 10018     | Hima Bindu Venkata...   | Female | Cold      | 1977-11-10 00:00:00.000 | 40678888    |
| 19 | 10019     | Phunsuk Wangdu          | Male   | Polio     | 1999-02-10 00:00:00.000 | 1715550297  |
| 20 | 10020     | Priyanka Badiatiya      | Female | Arthritis | 1945-12-06 00:00:00.000 | 76753425    |
| 21 | 10021     | Ho-Shi-Zong             | Male   | Fracture  | 1976-01-01 00:00:00.000 | 115559857   |

### FactPurchase

|    | BillId | ItemId | PatientId | QuantityOrdered | TotalAmount(INR) | PurchaseDate            | Payment Mode |
|----|--------|--------|-----------|-----------------|------------------|-------------------------|--------------|
| 1  | B001   | D001   | 10021     | 8               | 360.00           | 2010-09-29 00:00:00.000 | Cash         |
| 2  | B002   | D002   | 10009     | 3               | 315.00           | 2010-10-20 00:00:00.000 | Card         |
| 3  | B003   | D002   | 10002     | 4               | 405.00           | 2010-10-29 00:00:00.000 | Online       |
| 4  | B004   | D002   | 10015     | 2               | 225.00           | 2010-10-30 00:00:00.000 | Cash         |
| 5  | B005   | D002   | 10019     | 7               | 720.00           | 2010-10-30 00:00:00.000 | Card         |
| 6  | B006   | D002   | 10007     | 5               | 495.00           | 2010-11-23 00:00:00.000 | Online       |
| 7  | B007   | D002   | 10001     | 6               | 630.00           | 2010-11-24 00:00:00.000 | Cash         |
| 8  | B008   | D002   | 10009     | 5               | 495.00           | 2010-11-22 00:00:00.000 | Card         |
| 9  | B009   | D002   | 10017     | 4               | 405.00           | 2010-11-30 00:00:00.000 | Online       |
| 10 | B010   | D003   | 10001     | 5               | 1665.00          | 2010-12-20 00:00:00.000 | Cash         |
| 11 | B011   | D003   | 10001     | 5               | 1665.00          | 2010-12-13 00:00:00.000 | Card         |
| 12 | B012   | D004   | 10021     | 2               | 135.00           | 2010-12-14 00:00:00.000 | Online       |
| 13 | B013   | D005   | 10014     | 3               | 405.00           | 2010-12-19 00:00:00.000 | Cash         |
| 14 | B014   | M001   | 10001     | 12              | 1485.00          | 2010-12-31 00:00:00.000 | Card         |
| 15 | B015   | M001   | 10001     | 12              | 1485.00          | 2011-01-15 00:00:00.000 | Online       |
| 16 | B016   | M002   | 10002     | 10              | 900.00           | 2011-01-21 00:00:00.000 | Cash         |
| 17 | B017   | M003   | 10003     | 10              | 945.00           | 2011-01-16 00:00:00.000 | Card         |
| 18 | B018   | M006   | 10006     | 14              | 1125.00          | 2011-01-18 00:00:00.000 | Online       |

(Continued)

(Continued)

|    |      |      |       |    |           |                         |        |
|----|------|------|-------|----|-----------|-------------------------|--------|
| 19 | B019 | M008 | 10014 | 12 | 1530.00   | 2011-01-18 00:00:00.000 | Cash   |
| 20 | B020 | R001 | 10014 | 1  | 6750.00   | 2011-01-19 00:00:00.000 | Card   |
| 21 | B021 | R002 | 10017 | 1  | 2250.00   | 2011-01-17 00:00:00.000 | Online |
| 22 | B022 | R003 | 10001 | 1  | 100440.00 | 2011-01-20 00:00:00.000 | Cash   |
| 23 | B023 | R004 | 10021 | 2  | 1710.00   | 2011-01-21 00:00:00.000 | Card   |
| 24 | B024 | R005 | 10008 | 1  | 495.00    | 2011-01-25 00:00:00.000 | Online |
| 25 | B025 | M005 | 10005 | 16 | 855.00    | 2011-01-31 00:00:00.000 | Card   |
| 26 | B026 | M004 | 10016 | 12 | 585.00    | 2011-02-13 00:00:00.000 | Card   |
| 27 | B027 | M007 | 10012 | 11 | 675.00    | 2011-02-15 00:00:00.000 | Card   |
| 28 | B028 | M004 | 10014 | 15 | 765.00    | 2011-02-19 00:00:00.000 | Cash   |
| 29 | B029 | M004 | 10011 | 16 | 810.00    | 2011-02-18 00:00:00.000 | Online |
| 30 | B030 | D001 | 10014 | 12 | 540.00    | 2011-02-16 00:00:00.000 | Cash   |

Now the data is in a unified view and analysis can be done on the data as required. To know the sales of each product, we can write a query in SQL Server as:

```
CREATE TABLE ProductSales (
```

```
    ItemName varchar (50) PRIMARY KEY,  
    TotalQuantityOrdered int);
```

```
INSERT INTO ProductSales (ItemName, TotalQuantityOrdered)
```

```
Select B.ItemName, ISNULL (TotalQuantityOrdered, 0) AS TotalQuantityOrdered FROM  
((SELECT DimItem.ItemName, SUM (FactPurchase.QuantityOrdered) AS TotalQuantityOrdered  
FROM FactPurchase INNER JOIN  
DimItem ON FactPurchase.ItemId = DimItem.ItemId  
GROUP BY DimItem.ItemName) AS A  
RIGHT OUTER JOIN  
(Select ItemName from DimItem)AS B  
ON A.ItemName=B.ItemName);
```

The analyzed data can be shown as follows:

|    | ItemName      | TotalQuantityOrdered |
|----|---------------|----------------------|
| 1  | Instadril     | 43                   |
| 2  | Syringe       | 36                   |
| 3  | Avayatin      | 24                   |
| 4  | Cotton Pack   | 20                   |
| 5  | Antidrome     | 16                   |
| 6  | Bendrin       | 14                   |
| 7  | Mutacine      | 12                   |
| 8  | Nitacine      | 11                   |
| 9  | Acerint       | 10                   |
| 10 | Aquasure      | 10                   |
| 11 | Saline bottle | 10                   |
| 12 | Bandage       | 3                    |
| 13 | Donut Cushion | 2                    |

*(Continued)*

(Continued)

|    |                     |   |
|----|---------------------|---|
| 14 | Hand Gloves         | 2 |
| 15 | Adjustable Bed      | 1 |
| 16 | Thermometer         | 1 |
| 17 | Walker              | 1 |
| 18 | Wheel Chair         | 1 |
| 19 | Blood Glucose Meter | 0 |



### Remind Me

- Data warehouse is a system which can conveniently archive data.
- Data warehouse consists of four distinct components:
  - Operational source systems.
  - Data staging area.
  - Data presentation area.
  - Data access tools.
- ETL stands for Extract, Transform, Load.
- ETL is
  - Extracting data from different data sources.
  - Transforming the data into relevant format to fit operational needs.
  - Loading data into the final target database, usually a data warehouse.
- Data integration is the integration of data present in different sources and providing a unified view of the data.
- Common approaches of data integration:
  - Federated databases
  - Data warehousing
  - Memory-mapped data structure
- Data integration technologies:
  - Data interchange
  - Object brokering
  - Modeling techniques:
    - a. Entity Relationship (ER) Modeling
    - b. Dimensional Modeling
- Data profiling is the process of statistically examining and analyzing the content in a data source, and hence collecting information about that data.



### Point Me (Books)

- *The Data Warehouse ETL Toolkit* by Ralph Kimball.
- *The Data Warehouse Lifecycle Toolkit* by Ralph Kimball.



## Connect Me (Internet Resources)

- [http://en.wikipedia.org/wiki/Federated\\_database\\_system](http://en.wikipedia.org/wiki/Federated_database_system)
- [http://en.wikipedia.org/wiki/Data\\_integration](http://en.wikipedia.org/wiki/Data_integration)
- [http://www.telusplanet.net/public/bmarshal/dataqual.htm#DOC\\_TOP](http://www.telusplanet.net/public/bmarshal/dataqual.htm#DOC_TOP)
- <http://www.tech-faq.com/data-profiling.html>
- [http://it.toolbox.com/wiki/index.php/Data\\_profiling](http://it.toolbox.com/wiki/index.php/Data_profiling)
- [http://en.wikipedia.org/wiki/Data\\_profiling](http://en.wikipedia.org/wiki/Data_profiling)
- [www.rkimball.com/html/designtipsPDF/KimballDT59SurprisingValue.pdf](http://www.rkimball.com/html/designtipsPDF/KimballDT59SurprisingValue.pdf)
- <http://tdwi.org/Articles/2010/02/03/Data-Profiling-Value.aspx?Page=1>



## Test Me Exercises

### Match me

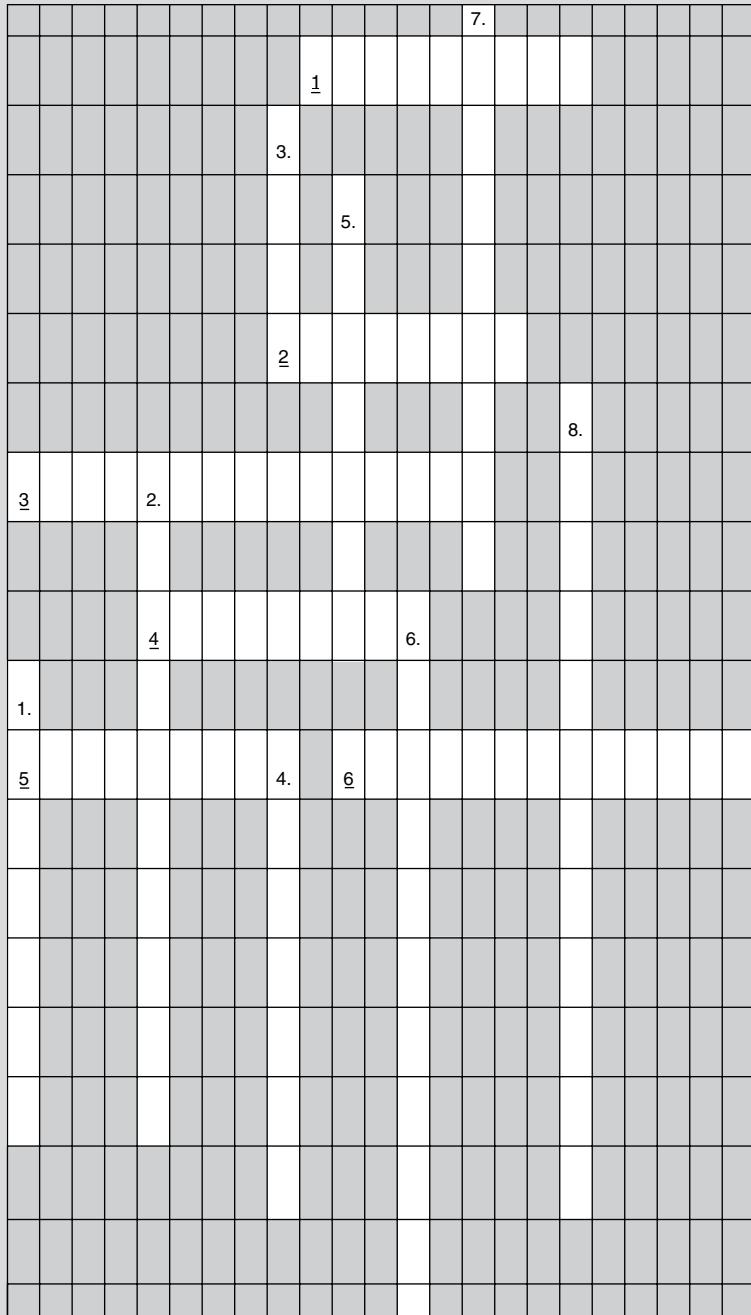
| Column A                               | Column B                |
|--|-------------------------|
| Data staging area                      | Reduce data redundancy  |
| Data warehouse                         | Object brokering        |
| ER Modeling                            | Data mapping            |
| Data Integration Technology            | Intermediate storage    |
| Data Integration Approach              | Data profiling software |
| Cycle initiation, build reference data | Archive data            |
| Datiris, Talend                        | Instance Integration    |

**Solution:**

| <i>Column A</i>                        | <i>Column B</i>         |
|--|-------------------------|
| Data staging area                      | Intermediate storage    |
| Data warehouse                         | Archive data            |
| ER Modeling                            | Reduce data redundancy  |
| Data Integration Technology            | Object brokering        |
| Data Integration Approach              | Instance integration    |
| Cycle initiation, build reference data | Data mapping            |
| Datiris, Talend                        | Data profiling software |



## BI Crossword



**ACROSS**

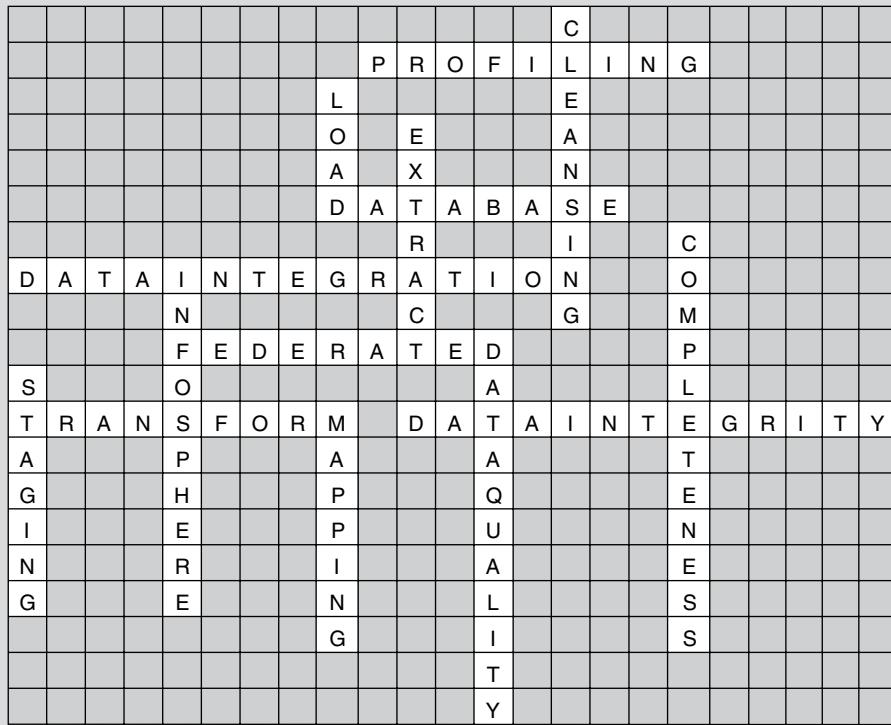
- 1** Statistical analysis of data content to assess quality of data (9)
- 2** A system that stores and organizes large amounts of data in digital form for easy management and retrieval of data (8)
- 3** The process of combining data from different sources to present it in a unified form, while maintaining the integrity of the data (15)
- 4** A virtual database that is the integrated view of multiple heterogeneous, autonomous and disparate databases (9)
- 5** The stage of data integration where conditions and rules are applied to make functional changes to data extracted from source (9)
- 6** The accuracy with which the attributes associated with data of an entity accurately describe the occurrence of that entity (13)

**DOWN**

1. Intermediate storage area, that falls in between data source and data mart (7)
2. Data profiling tool developed by IBM:  
\_\_\_\_\_ *Information Analyzer* (10)

- 3**. After transformation of source data, it goes into target database. This final process is called \_\_\_\_\_ (4)
- 4**. The process by which data elements are matched between two distinct data models for the purpose of identifying relationships, discovering hidden data, identifying required transformations etc. (7)
- 5**. Process by which data is pulled out from source for further transformation (7)
- 6**. The reliability, efficiency and overall integrity of data, and its appropriateness for the purpose of the business is called \_\_\_\_\_ (11)
- 7**. Process of correcting/removing inconsistencies, inaccuracy, irregularities etc. in the data from records in a database (9)
- 8**. Property of the data in a database system which measures the amount of data actually available as compared to the amount of data that is expected to be available (12)

### Solution:



# SOLVED EXERCISES

1. State the differences between data warehouse and data marts.

**Solution:**

| <i>Data Warehouse</i>  | <i>Data Mart</i>                                      |
|--|---|
| Focuses on one subject area.   | Focuses on all subject areas across the enterprise.   |
| There can be more than one data mart for an organization.<br>There can be one data mart for every function such as a data mart for finance, a data mart for marketing, a data mart for human resources, etc. | There can be only one enterprise-wide data warehouse. |
| Contains relatively less information.  | Contains relatively more information.                 |
| Is easy to understand and navigate.  | Is relatively tougher to understand and navigate.     |

2. What are the differences between the top-down approach and the bottom-up approach to building a data warehouse?

**Solution:**

| <i>Top-Down Approach</i>  | <i>Bottom-Up Approach</i>   |
|---|---|
| The focus is on data warehouse.   | The focus is on data marts.   |
| We start with designing the model for a data warehouse.   | We start with designing the model for a data mart.                                      |
| Data warehouse is enterprise-oriented while data marts are subject-specific.  | Data marts provide both enterprise and subject specific views.                          |
| Data warehouse is known to contain data at the atomic level whereas the data marts contain summarized data.               | Data marts contain both atomic and summary data.  |
| Both the data warehouse and data marts can be queried.  | Usually the data marts are queried.   |
| The top-down approach uses a multi-tier architecture comprising of staging area, data warehouse and dependent data marts. | The bottom-up approach uses a flat architecture comprising staging area and data marts. |

3. What are the various ETL tools available in the market?

**Solution:**

| <i>ETL Tool</i>                 | <i>Vendor</i>           |
|---------------------------------|-------------------------|
| Informatica                     | Informatica Corporation |
| SQL Server Integration Services | Microsoft               |
| Ab initio                       | Ab initio Corporation   |
| Oracle Warehouse Builder        | Oracle Corporation      |
| Data Stage                      | IBM Corporation         |

4. What do you understand by data cleansing?

**Solution:** Data cleansing is also called data scrubbing. It is a process that involves the detection and removal/of corrections from the database, which are caused by incomplete, inaccurate, redundant, obsolete, or improperly formatted data.

5. What is real time data warehousing?

**Solution:** In general, a data warehouse is an archive of historical data. Such a data warehouse is a window on the past. If such a data warehouse is queried, it will reflect the state of the business sometime in the past. Real time data warehouse is known to house real time business data. If such a data warehouse is queried, it will reflect the state of the business at the time the query was run.

## UNSOLVED EXERCISES

---

1. Why is there a need for a data warehouse? Explain.
2. What is a data warehouse? Explain.
3. Explain these terms in William H. Inmon's definition of a data warehouse: subject-oriented, integrated, non-volatile, and time-variant.
4. How is the Ralph Kimball's approach to building a data warehouse different from the William H. Inmon's approach to building a data warehouse? Explain.
5. When should an organization go for a data warehouse? Comment.
6. Should larger organizations adopt the William H. Inmon's approach to building a data warehouse? Explain with reasons.
7. What are the goals that a data warehouse is meant to achieve?
8. What is your understanding of data sources? Explain.
9. What is data integration? Why should data be integrated in a data warehouse? Explain.
10. What are the advantages of building a data warehouse? Explain.
11. What are the various approaches to data integration?
12. What constitutes a data warehouse? Explain.
13. Assume you are a teacher and you are required to collect the "Date of Birth" of your students. To your surprise, students provide "Date of Birth" in various formats such as "DD.MM.YY", "DD/MM/YYYY", "DD-MMM-YY", "MM/DD/YY", "YY.MM.DD". What will you do when faced with such a situation?
14. Explain schema integration and instance integration with an example.
15. What is your understanding of "data transformation"? Why is it required to transform data?
16. Explain the ETL process.
17. Explain the difference between "entity relationship modeling" and "dimensional modeling".
18. Explain the steps to convert an ER model to a dimensional model.
19. How is an "Operational Data Store" different from a Data Warehouse? Explain.
20. What is your understanding of the staging area? Explain.
21. What is the meaning of data profiling? Explain.
22. When and how is data profiling conducted? Explain.
23. Mention a few data profiling tools available in the market.
24. Mention a few ETL tools available in the market.
25. How is a data warehouse different from a data mart? Explain.
26. Can reports be pulled from an OLTP source or from an operation data store (ODS)? Explain.
27. What is "single version of truth"? Explain with an example.
28. How is a federated database different from a data warehouse? Explain.
29. What is your understanding of the "presentation area"? Explain.
30. Is it a good idea for enterprises to maintain independent data marts? Explain your answer with reasons.



# 7



# Multidimensional Data Modeling

---

## BRIEF CONTENTS

|                          |                                 |
|--------------------------|---------------------------------|
| What's in Store          | Dimension Table                 |
| Introduction             | Typical Dimensional Models      |
| Data Modeling Basics     | Dimensional Modeling Life Cycle |
| Types of Data Model      | Designing The Dimensional Model |
| Data Modeling Techniques | Solved Exercises                |
| Fact Table               | Unsolved Exercises              |

---

## WHAT'S IN STORE

You are already familiar with the concepts relating to basics of RDBMS, OLTP, and OLAP applications, role of ERP in the enterprise as well as “enterprise production environment” for IT deployment. We hope you have got a firm grasp on the concepts covered in the chapters “Types of Digital Data” (Chapter 2), “Introduction to OLTP and OLAP” (Chapter 3), “Getting Started with Business Intelligence” (Chapter 4), and “Basics of Data Integration (Chapter 6)”. With this background it’s time to think about “how data is modeled”.

Just like a circuit diagram is to an electrical engineer, an assembly diagram is to a mechanical engineer, and a blueprint of a building is to a civil engineer so is the data models/data diagrams to a data architect. But is “data modeling” only the responsibility of a data architect? The answer is “No”. A Business Intelligence (BI) application developer today is involved in designing, developing, deploying, supporting, and optimizing storage in the form of data warehouse/data marts. To be able to play his/her role efficiently, the BI application developer relies heavily on data models/data diagrams to understand

the schema structure, the data, the relationships between data, etc. In this chapter we will familiarize you with the basics of data modeling – How to go about designing a data model at the conceptual and logical levels? We will also discuss the pros and cons of the popular modeling techniques such as ER modeling and dimensional modeling.

We suggest you refer to some of the learning resources suggested at the end of this chapter and also complete the “Test Me” exercises. You will get deeper knowledge by interacting with people who have shared their project experiences in blogs. We suggest you make your own notes/bookmarks while reading through the chapter.

## 7.1 INTRODUCTION

---

### Refer to the Case Study Brief of “TenToTen Retail Stores”

A new range of cosmetic products has been introduced by a leading brand, which TenToTen wants to sell through its various outlets. In this regard TenToTen wants to study the market and the consumer's choice of cosmetic products. As a promotional strategy the group also wants to offer attractive introductory offers like discounts, buy one get one free, etc.

To have a sound knowledge of the cosmetics market, TenToTen Stores has to carry out a detailed study of the buying pattern of consumers' by geography, the sales of cosmetic products by outlet, most preferred brand, etc. and then decide on a strategy to promote the product.

To take right decisions on various aspects of business expansion, product promotion, consumer preferences, etc., TenToTen Stores has decided to go in for an intelligent decision support system. They approached “AllSolutions” to provide them with an automated solution. AllSolutions is one of the leading consulting firms of the world. They have been into business for about 15 years. Over the years they have become known for providing optimal IT solutions to the business problems of big and small enterprises alike.

After studying the requirements of TenToTen Stores, AllSolutions decided on building a data warehouse application. To construct a data model that would meet the business requirements put forth by TenToTen Stores, AllSolutions identified the following concerns that need to be addressed:

1. What are the entities involved in this business process and how are they related to each other?
2. What tables associated with those entities must be included in the data warehouse?
3. What columns have to be included into each table?
4. What are the primary keys for the tables that have been identified?
5. What are the relations that the tables have with each other and which is the column on which the relationship has to be made?
6. What should be the column definitions for the columns that have been identified?
7. What are the other constraints to be added into the tables?

Having zeroed down on the requirements for TenToTen Stores, we now proceed to build its data model. But, let us first understand “*What is a data model?*” and “*What are the steps involved in designing a data model?*”

## 7.2 DATA MODELING BASICS

Before getting down to the basics of data modeling, let us do a quick recap on a few database terms such as entity, attribute, cardinality of a relationship, etc.

### 7.2.1 Entity

Entity is a common word for anything real or *abstract* about which we want to store data. Entity types generally fall into five categories: roles (such as employee, executives, etc.), events (such as hockey match, cricket match, etc.), locations (such as office campus, auditorium, multiplex theatre, etc.), *tangible* things (such as book, laptop, etc.), or concepts (such as a project idea). For example, Harry, Luther, Connors, Jennifer, etc. could be instances of the “employee” entity.

### 7.2.2 Attribute

An attribute is a characteristic property of an entity. An entity could have multiple attributes. For example, for the “car” entity, attributes would be color, model number, number of doors, right or left hand drive, engine number, diesel or petrol, etc.

### 7.2.3 Cardinality of Relationship

This relationship defines the type of relationship between two participating entities. For example, one employee can work on only one project at any given point in time. One project, however, can have several employees working on it. So, the cardinality of relationship between employee and project is “many to one”. Here is an example of a “one to one” cardinality. One person can sit on only one chair at any point of time. One chair can accommodate only one person in a given point of time. So, this relationship has “one to one” cardinality.

To read more on entity, attribute, cardinality, etc. please refer:

- *An Introduction to Database Systems* (8th Edition), C.J. Date, Addison Wesley.
- *Database System Concepts* by Abraham Silberschatz, Henry Korth and S. Sudarshan, McGraw Hill.
- *Fundamentals of Database Systems* (6th Edition) by Ramez Elmasri and Shamkant Navathe, Addison Wesley.

With this understanding of a few database terms, let us proceed to understand data model.

## 7.3 TYPES OF DATA MODEL

A data model is a diagrammatic representation of the data and the relationship between its different entities. Although time consuming, the process of creating a data model is extremely important. It assists in identifying how the entities are related through a visual representation of their relationships and thus helps reduce possible errors in the database design. It helps in building a robust database/data warehouse. There are three types of data models:

- Conceptual Data Model.

- Logical Data Model.
- Physical Data Model.

### 7.3.1 Conceptual Data Model

The conceptual data model is designed by identifying the various entities and the highest-level relationships between them as per the given requirements. Let us look at some features of a conceptual data model:

- It identifies the most important entities.
- It identifies relationships between different entities.
- It does not support the specification of attributes.
- It does not support the specification of the primary key.

Going back to the requirement specification of TenToTen Stores, let us design the conceptual data model (Figure 7.1). In this case, the entities can be identified as

- **Category** (to store the category details of products).
- **SubCategory** (to store the details of sub-categories that belong to different categories).
- **Product** (to store product details).
- **PromotionOffer** (to store various promotion offers introduced by the company to sell products).
- **ProductOffer** (to map the promotion offer to a product).
- **Date** (to keep track of the sale date and also to analyze sales in different time periods).
- **Territory** (to store various territories where the stores are located).
- **MarketType** (to store details of various market setups, viz. “Hypermarkets & Supermarkets”, “Traditional Supermarket”, “Dollar Store”, and “Super Warehouse”).
- **OperatorType** (to store the details of types of operator, viz. company-operated or franchise).
- **Outlet** (to store the details of various stores distributed over various locations).
- **Sales** (to store all the daily transactions made at various stores).

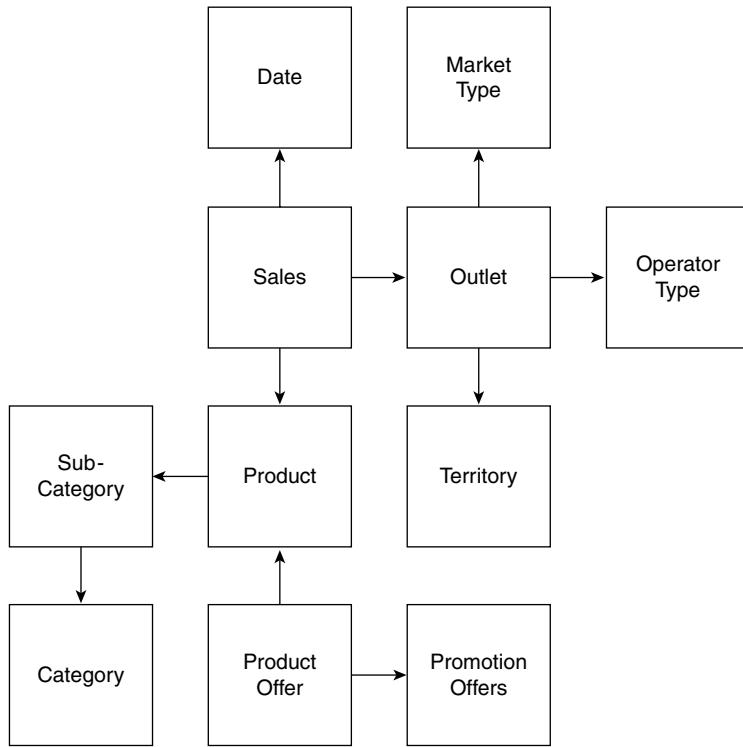
There can be several other entities, but for the current scenario we restrict ourselves to the entities listed above. Let us now define the relationships that exist between the various entities listed above.

- **Outlet** has a **MarketType**.
- **Outlet** has an **OperatorType**.
- **Outlet** belongs to a **Territory**.
- **SubCategory** belongs to a **Category**.
- **Product** belongs to a **SubCategory**.
- **ProductOffer** is an instance of **PromotionOffer** for a **Product**.
- **Sales** of a **Product**.
- **Sales** in a specific **Date** duration.
- **Sales** from an **Outlet**.

### 7.3.2 Logical Data Model

The logical data model is used to describe data in as much detail as possible. While describing the data, practically no consideration is given to the physical implementation aspect. Let us look at some features of a logical data model:

- It identifies all entities and the relationships among them.
- It identifies all the attributes for each entity.



**Figure 7.1** The conceptual data model for TenToTen Stores.

- It specifies the primary key for each entity.
- It specifies the foreign keys (keys identifying the relationship between different entities).
- Normalization of entities is performed at this stage.

Here is a quick recap of various normalization levels:

| Normal Form | Test   | Remedy (Normalization)  |
|-------------|--|---|
| 1NF         | Relation should have atomic attributes. The domain of an attribute must include only atomic (simple, indivisible) values.  | Form new relations for each non-atomic attribute.   |
| 2NF         | For relations where the primary key contains multiple attributes (composite primary key), non-key attributes should not be functionally dependent on a part of the primary key.  | Decompose and form a new relation for each partial key with its dependent attribute(s). Retain the relation with the original primary key and any attributes that are fully functionally dependent on it. |
| 3NF         | Relation should not have a non-key attribute functionally determined by another non-key attribute (or by a set of non-key attributes). In other words, there should be no transitive dependency of a non-key attribute on the primary key. | Decompose and form a relation that includes the non-key attribute(s) that functionally determine(s) other non-key attribute(s).   |

**For more on normalization, we recommend the following books:**

- *An Introduction to Database Systems* (8th Edition), Addison Wesley.
- *Database System Concepts* by Abraham Silberschatz, Henry Korth and S. Sudarshan, McGraw Hill.

Before introducing you to the physical data model, we get back to our case study. Having identified the entities, now we shall identify the various attributes for the entities and thus design the logical model for TenToTen Stores database.

### Category

| Columns           | Description  | Constraint         |
|-------------------|--|--------------------|
| ProductCategoryID | The ProductCategory ID is used to uniquely identify different types of categories. | <b>Primary Key</b> |
| CategoryName      | The name of the corresponding category.  |                    |

### SubCategory

| Columns              | Description  | Constraint                                 |
|----------------------|--|--|
| ProductSubCategoryID | The Product SubCategory ID uniquely identifies the sub-categories that belong to each product. | <b>Primary Key</b>                         |
| SubCategoryName      | The name of the corresponding sub-category.  |  |
| ProductCategoryID    | The Product Category ID to which the sub-category belongs.                                     | Refers to the Category (ProductCategoryID) |

### Product

| Columns                | Description   | Constraint          |
|------------------------|---|---------------------|
| ProductID              | The product code for the respective product, and will be the primary key for the Product Table. | <b>Primary Key</b>  |
| ProductName            | The name of the product.  |                     |
| ProductDescription     | Gives a brief description about the product.  |                     |
| SubCategoryID          | Describes the sub-category the product belongs to.  |                     |
| DateOfProduction       | Date when the corresponding product was manufactured.   | “dd/mm/yyyy” format |
| LastDate<br>OfPurchase | The last date the shipping was made to the warehouse.   |                     |
| CurrentInventoryLevel  | The number of products that are currently present in the inventory of the product.              |                     |
| StandardCost           | The standard price for the product.   |                     |
| ListPrice              | The listed price for the product.   |                     |

(Continued)

(Continued)

| Columns          | Description   | Constraint          |
|------------------|---|---------------------|
| Discontinued     | Used to find whether the manufacturing of a particular product has been discontinued. | "Y" or "N"          |
| DiscontinuedDate | The date when the product manufacture was discontinued.                               | "dd/mm/yyyy" format |

### PromotionOffers

| Columns              | Description  | Constraint                              |
|----------------------|--|---|
| PromotionOfferID     | It is used to uniquely identify the various promotion offers.      | <b>Primary Key</b>                      |
| PromotionType        | The type of offers given.  | (Discounts, Buy One Get One Free, etc.) |
| DiscountPercentage   | The percentage of discount given on the List Price of the product. |   |
| ComplimentaryProduct | Complimentary products that might be offered with a product.       |   |
| DateOfOfferExpiry    | The date when the offer will expire.                               | "dd/mm/yyyy" format                     |

### ProductOffer

| Columns          | Description  | Constraint   |
|------------------|--|--|
| ProductID        | Refers to a product from the product table to which the offer is to be made. | <b>Primary Key</b> ,<br>Refers to Product (ProductID)                |
| PromotionOfferID | Refers to the promotion offer that is to be given with a product.            | <b>Primary Key</b> ,<br>Refers to PromotionOffers (PromotionOfferID) |

### Date

| Columns     | Description   | Constraint   |
|-------------|---|--|
| DateID      | The Date ID for each date for uniquely identifying each date. | <b>Primary Key</b>                                     |
| Date        | The corresponding date.                                       | "dd/mm/yyyy" format                                    |
| DayOfWeek   | The name of the day of the week.                              | (Monday, Tuesday, Wednesday, Thursday, Friday, Sunday) |
| WeekOfMonth | The week number w.r.t. the month.                             | (1, 2, 3, 4)   |

(Continued)

(Continued)

| Columns     | Description                                    | Constraint                         |
|-------------|--|------------------------------------|
| WeekOfYear  | Describes the category the product belongs to. | (1, 2, 3, ..., 52)                 |
| MonthName   | Contains the month name of the year.           | (January, February, ..., December) |
| MonthOfYear | The month number in the year.                  | (1, 2, 3, ..., 12)                 |
| Quarter     | The quarter period of the corresponding year.  | (1, 2, 3, 4)                       |
| Year        | Contains the corresponding year.               | “yyyy” format                      |

### MarketType

| Columns     | Description  | Constraint  |
|-------------|--|---|
| MarketID    | Uniquely identifies the store setup.               | <b>Primary Key</b>  |
| Market_Name | The store setup based on which the store operates. | (Hypermarkets & Supermarkets, Traditional Supermarket, Dollar Store, Super Warehouse) |

### Operator\_Type

| Columns     | Description                              | Constraint                           |
|-------------|--|--------------------------------------|
| Operator_ID | Identifies the type of operation.        | <b>Primary Key</b>                   |
| Operator    | The type of working of the outlet/store. | (Company Operated, Franchises, etc.) |

### Outlets

| Columns      | Description  | Constraint                                |
|--------------|--|---|
| Store ID     | Each store has a respective Store ID. It will be used to uniquely identify various stores. | <b>Primary Key</b>                        |
| MarketID     | The market type of the store.  | Refers to the Market (MarketID)           |
| OperatorID   | The working type of the store.   | Refers to the Operator_Type (Operator_ID) |
| TerritoryID  | It contains the territory that the store belongs to.                                       | Refers to the Territory Table             |
| Opening Time | The time when the corresponding store opens.   | 24 hours format                           |

*(Continued)*

(Continued)

| Columns       | Description                                   | Constraint           |
|---------------|---|----------------------|
| Closing Time  | The time when the corresponding store closes. | 24 hours format      |
| StoreOpenDate | Contains the date when the store was opened.  | “dd/mm/yyyy” format. |

## Territory

| Columns       | Description  | Constraint         |
|---------------|--|--------------------|
| TerritoryCode | Identifies each territory uniquely in the Territory Table.     | <b>Primary Key</b> |
| Territory     | The name of the territory corresponding to the territory code. |                    |
| City          | The city in which the territory is present.                    |                    |
| State         | The state to which the city belongs to.                        |                    |

## SalesTransaction

| Columns         | Description   | Constraint  |
|-----------------|---|---|
| TransactionID   | Identifies every sale that was made.                                      | <b>Primary Key</b>  |
| TransactionDate | Contains the date the sales were made.                                    | Refers to the Time Table                                  |
| StoreID         | Contains the ID of the store from where the sale was made.                | Refers to the Stores Table                                |
| ProductID       | Contains the ID of the product that was sold.                             | Refers to the Product Table                               |
| QuantityOrdered | Contains the quantity of the product sold.                                |   |
| TotalAmount     | Contains the total amount of the sale made for the corresponding product. | $\text{SalesAmount} = \text{ItemSold} * \text{UnitPrice}$ |

To conclude the logical data modeling:

- We have identified the various entities from the requirements specification.
- We have identified the various attributes for each entity.
- We have also identified the relationship that the entities share with each other (Primary key – Foreign Key).

The logical data model for TenToTen Stores, based on the above-identified entities, is depicted in Figure 7.2.

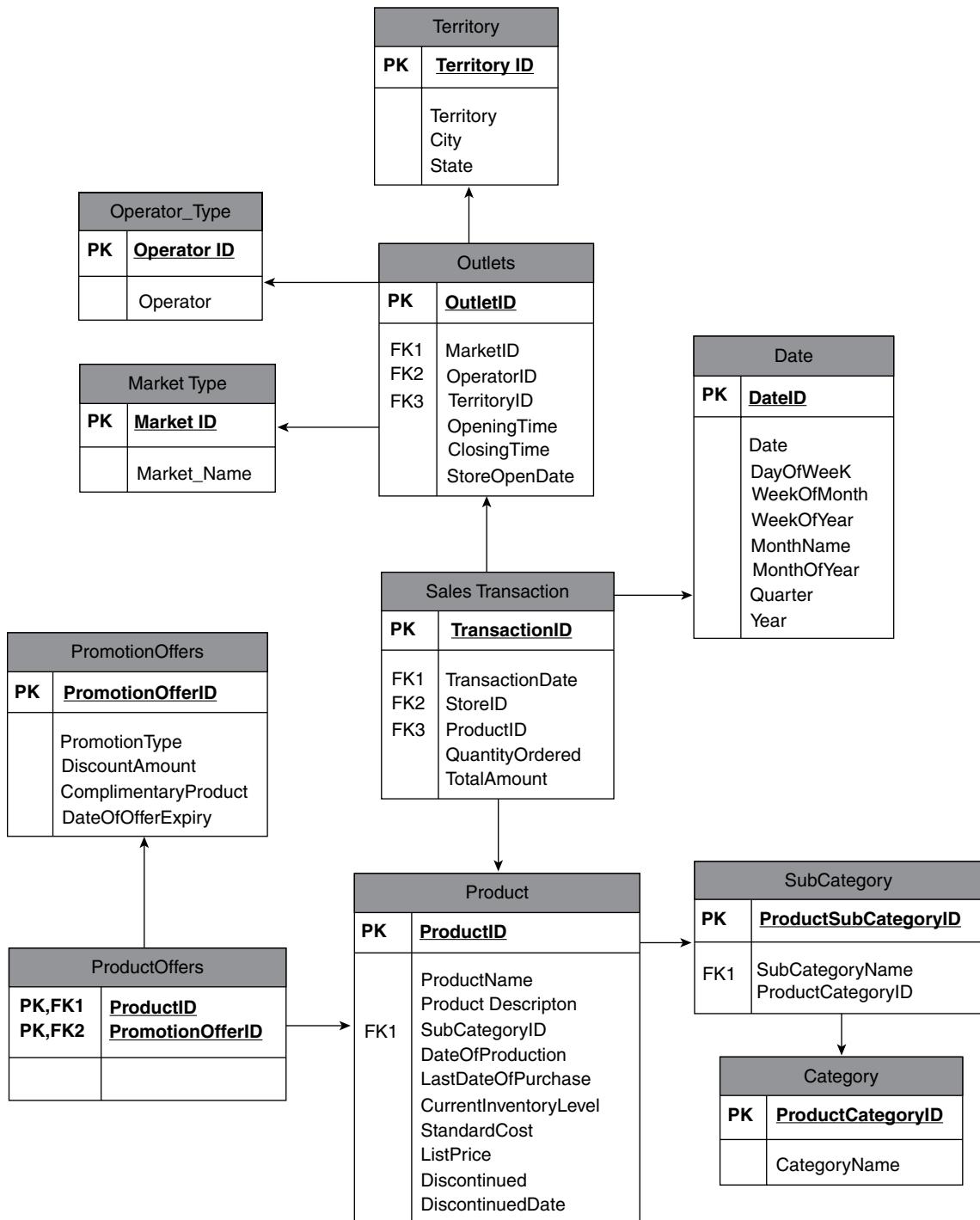


Figure 7.2 Logical data model for TenToTen Stores.

Now that we understand conceptual and logical data model, let us perform a quick comparison between the two:

- All attributes for each entity are specified in a logical data model, whereas no attributes are specified in a conceptual data model.
- Primary keys are present in a logical data model, whereas no primary key is present in a conceptual data model.
- In a logical data model, the relationships between entities are specified using primary keys and foreign keys, whereas in a conceptual data model, the relationships are simply stated without specifying attributes. It means in a conceptual data model, we only know that two entities are related; we don't know which attributes are used for establishing the relationship between these two entities.

### 7.3.3 Physical Model

A physical data model is a representation of how the model will be built in the database. A physical database model will exhibit all the table structures, including column names, columns data types, column constraints, primary key, foreign key, and the relationships between tables. Let us look at some features of a physical data model:

- Specification of all tables and columns.
- Foreign keys are used to identify relationships between tables.
- While logical data model is about normalization, physical data model may support de-normalization based on user requirements.
- Physical considerations (implementation concerns) may cause the physical data model to be quite different from the logical data model.
- Physical data model will be different for different RDBMS. For example, data type for a column may be different for MySQL, DB2, Oracle, SQL Server, etc.

The steps for designing a physical data model are as follows:

- Convert entities into tables/relations.
- Convert relationships into foreign keys.
- Convert attributes into columns/fields.

Let us look at the physical data model in the light of our TenToTen Stores case study. Having created the logical data model, now the physical model shall be created for the respective entities by adding the column definition for the attributes. The detailed descriptions of the table are as shown below (we use MS SQL Server 2008 Database here):

#### Category

|                    |                   |             |
|--------------------|-------------------|-------------|
| iProductCategoryID | INT IDENTITY(1,1) | PRIMARY KEY |
| vCategoryName      | VARCHAR(50)       | NOT NULL    |

## SubCategory

|                       |                   |                              |
|-----------------------|-------------------|------------------------------|
| iProductSubCategoryID | INT IDENTITY(1,1) | PRIMARY KEY                  |
| vSubCategoryName      | VARCHAR(50)       | NOT NULL                     |
| iProductCategoryID    | INT               | Category(iProductCategoryID) |

## Product

|                        |                   |                                    |
|------------------------|-------------------|------------------------------------|
| iProductID             | INT IDENTITY(1,1) | PRIMARY KEY                        |
| vProductName           | VARCHAR(50)       | NOT NULL                           |
| vProductDescription    | VARCHAR(250)      | NOT NULL                           |
| iSubCategoryID         | INT               | SubCategory(iProductSubCategoryID) |
| dDateOfProduction      | DATE              | NOT NULL                           |
| dLastDateOfPurchase    | DATE              | NOT NULL                           |
| iCurrentInventoryLevel | INT               | NOT NULL                           |
| mStandardCost          | MONEY             | NOT NULL                           |
| mListPrice             | MONEY             | NOT NULL                           |
| cDiscontinued          | CHAR(1)           | CHECK "Y" or "N"                   |
| dDiscontinuedDate      | DATE              | NULL                               |

## PromotionOffers

|                       |                   |             |
|-----------------------|-------------------|-------------|
| iPromotionOfferID     | INT IDENTITY(1,1) | PRIMARY KEY |
| vPromotionType        | VARCHAR(30)       | NOT NULL    |
| tiDiscountPercent     | TINYINT           | NULL        |
| vComplimentaryProduct | VARCHAR(50)       | NULL        |
| dDateOfOfferExpiry    | DATE              | NOT NULL    |

## ProductOffer

|                   |     |  |
|-------------------|-----|--|
| iProductID        | INT | PRIMARY KEY Product (iProductID)               |
| iPromotionOfferID | INT | PRIMARY KEY PromotionOffers (PromotionOfferID) |

## Date

|              |                   |             |
|--------------|-------------------|-------------|
| iDateID      | INT IDENTITY(1,1) | PRIMARY KEY |
| dDate        | DATE              | NOT NULL    |
| vDayOfWeek   | VARCHAR(10)       | NOT NULL    |
| iWeekOfMonth | INT               | NOT NULL    |

(Continued)

(Continued)

|              |             |          |
|--------------|-------------|----------|
| iWeekOfYear  | INT         | NOT NULL |
| vMonthName   | VARCHAR(10) | NOT NULL |
| iMonthOfYear | INT         | NOT NULL |
| iQuarter     | INT         | NOT NULL |
| iYear        | INT         | NOT NULL |

**MarketType**

|              |                   |             |
|--------------|-------------------|-------------|
| iMarketID    | INT IDENTITY(1,1) | PRIMARY KEY |
| vMarket_Name | VARCHAR(30)       | NOT NULL    |

**Operator\_Type**

|              |                   |             |
|--------------|-------------------|-------------|
| iOperator_ID | INT IDENTITY(1,1) | PRIMARY KEY |
| vOperator    | VARCHAR(50)       | NOT NULL    |

**Territory**

|                |                   |             |
|----------------|-------------------|-------------|
| iTerritoryCode | INT IDENTITY(1,1) | PRIMARY KEY |
| vTerritory     | VARCHAR(30)       | NOT NULL    |
| vCity          | VARCHAR(6)        | NOT NULL    |
| vState         | VARCHAR(6)        | NOT NULL    |

**Outlets**

|                |                   |                                      |
|----------------|-------------------|--------------------------------------|
| iStoreID       | INT IDENTITY(1,1) | PRIMARY KEY                          |
| iMarketID      | INT               | Market(iMarketID)                    |
| iOperatorID    | INT               | Operator_Type(iOperator_ID)          |
| iTerritoryID   | INT               | Territory(iTerritory_ID)             |
| vOpeningTime   | VARCHAR(5)        | NOT NULL,<br>[0-2][0-9][:][0-9][0-9] |
| vClosingTime   | VARCHAR(5)        | NOT NULL,<br>[0-2][0-9][:][0-9][0-9] |
| dStoreOpenDate | DATE              | NOT NULL                             |

**SalesTransaction**

|                  |                   |               |
|------------------|-------------------|---------------|
| iTransactionID   | INT IDENTITY(1,1) | PRIMARY KEY   |
| iTransactionDate | INT               | Date(iDateID) |

*(Continued)*

(Continued)

|                  |       |                     |
|------------------|-------|---------------------|
| iStoreID         | INT   | Outlets (iOutletID) |
| iProductID       | INT   | Product(iProductID) |
| iQuantityOrdered | INT   | NOT NULL            |
| mTotalAmount     | MONEY | NOT NULL            |

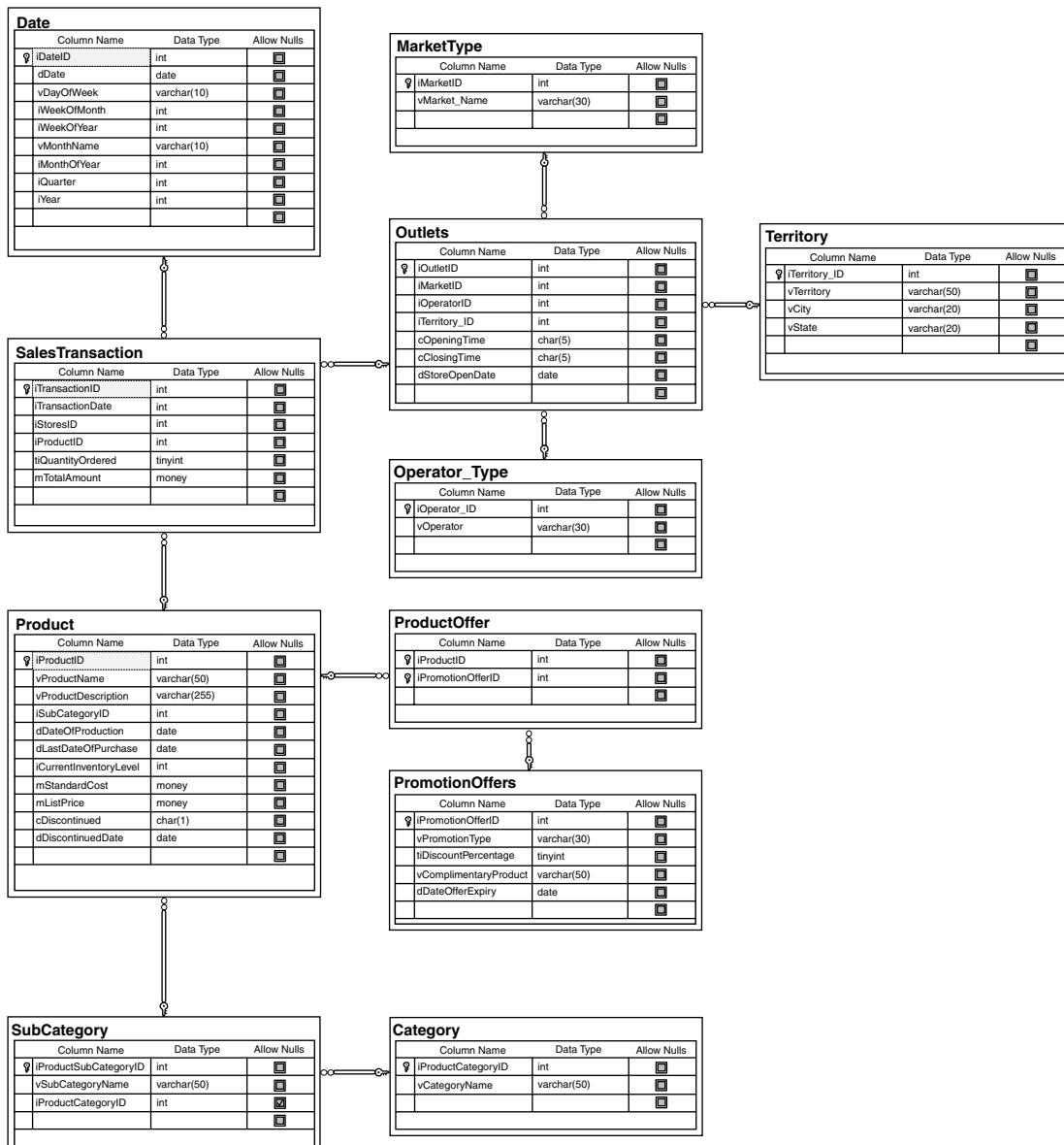
**Figure 7.3** Physical data model for TenToTen Stores.

Figure 7.3 depicts the physical model for TenToTen Stores. Let us look at a few points of difference between the logical data model and the physical data model:

- The entity names of the logical data model are table names in the physical data model.
- The attributes of the logical data model are column names in the physical data model.
- In the physical data model, the data type for each column is specified. However, data types can differ depending on the actual database (MySQL, DB2, SQL Server 2008, Oracle, etc.) being used. In a logical data model, only the attributes are identified without going into details about the data type specifications.

## 7.4 DATA MODELING TECHNIQUES

### 7.4.1 Normalization (Entity Relationship) Modeling

As learnt earlier, a table/relation in Third Normal Form (3NF) has no transitive dependency of a non-key attribute on the primary key. This is achieved by decomposing and forming a relation that includes the non-key attribute(s) that functionally determine(s) other non-key attribute(s). The Entity Relationship (ER) Model makes use of the third normal form (3NF) design technique. This form of normalization is very useful as it takes care of the insert, delete, and update anomalies. Here is an example of ER modeling:

An industry service provider, “InfoMechanists”, has several Business Units (BUs) such as Financial Services (FS), Insurance Services (IS), Life Science Services (LSS), Communication Services (CS), Testing Services (TS), etc. Each BU has a BU Head, who is most certainly an employee of the company. A BU head can head at the most one BU. A BU Head, as a manager, has many employees reporting to him. Each employee has a current residential address. There are also cases where a couple (both husband and wife) are employed either in the same BU or a different one. In such a case, they (the couple) have the same address. An employee can be on a project, but at any given point in time, he or she can be working on a single project only. Each project belongs to a client. There could be chances where a client has awarded more than one project to the company (either to the same BU or different BUs). A project can also be split into modules which can be distributed to BUs according to their field of specialization. For example, in an insurance project, the development and maintenance work is with Insurance Services (IS) and the testing task is with Testing Services (TS). Each BU usually works on several projects at a time.

Given the above specifications, let us see how we will proceed to design an ER model. Enumerated below is a list of steps to help you arrive at the ER diagram:

1. Identify all the entities.
2. Identify the relationships among the entities along with cardinality and participation type (total/partial participation).
3. Identify the key attribute or attributes.
4. Identify all other relevant attributes.
5. Plot the ER diagram with all attributes including key attribute(s).
6. The ER diagram is then reviewed with the business users.

#### Step 1: Identify all the entities

- Business Units

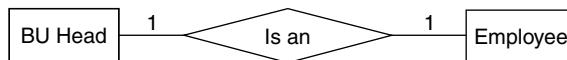
- BU Head
- Employee
- Address
- Project
- Client

### Step 2: Identify the relationships among the entities

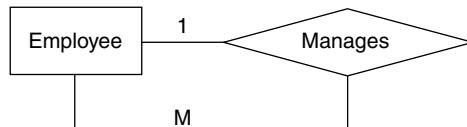
- One BU will have only one head. Hence the cardinality is “one to one”.



- A BU Head is also an employee. Hence the cardinality is “one to one”. The following diagram illustrates a case of partial and total participation. Each BU Head is an employee, but every employee is not a BU Head. The participation of the entity BU Head in the relationship “Is an” is “Total”, but the participation of the entity Employee in the relationship “Is an” is “Partial”.



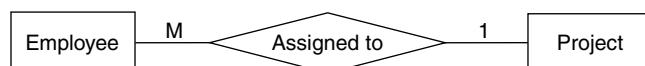
- An employee in the role of a manager may have several employees reporting into him/her. Hence, the cardinality is “one to many” (self-referential).



- One employee resides at a single address but many employees (in the case of spouses) can have the same address. Hence the cardinality is “many to one”.



- One employee can work on one project only at any point in time but one project can have several employees assigned to it. Hence the cardinality is “many to one”.



- One client could have awarded several projects to the company. Hence the cardinality is “one to many”.



- Each BU can have many projects allocated to it, but there are chances that different modules of the same project may be allocated to different units. Hence the cardinality is “many to one”.



### Step 3: Identify the key attribute or attributes

- BU Name** is the key attribute for the entity “Business Units”, as it identifies the business units uniquely.
- EmployeeID** is the key attribute for the entity “Employee” which is a foreign key for the “BU Head” entity.
- ProjectID** is the key attribute for the “Project” entity.
- ClientID** is the key attribute for the “Client” entity.
- HouseNumber** is the key attribute for the “Address” entity.

### Step 4: Identify all other relevant attributes

- Business Units(Domain).
- Employee(EmployeeName, EmailID, PhoneNumber).
- Project(ProjectName, StartDate, EndDate).
- Client(ClientName).
- Address(Street, City, State, Country).

### Step 5: Draw the ER diagram. Figure 7.4 shows the diagram of the ER model for “InfoMechanists”.

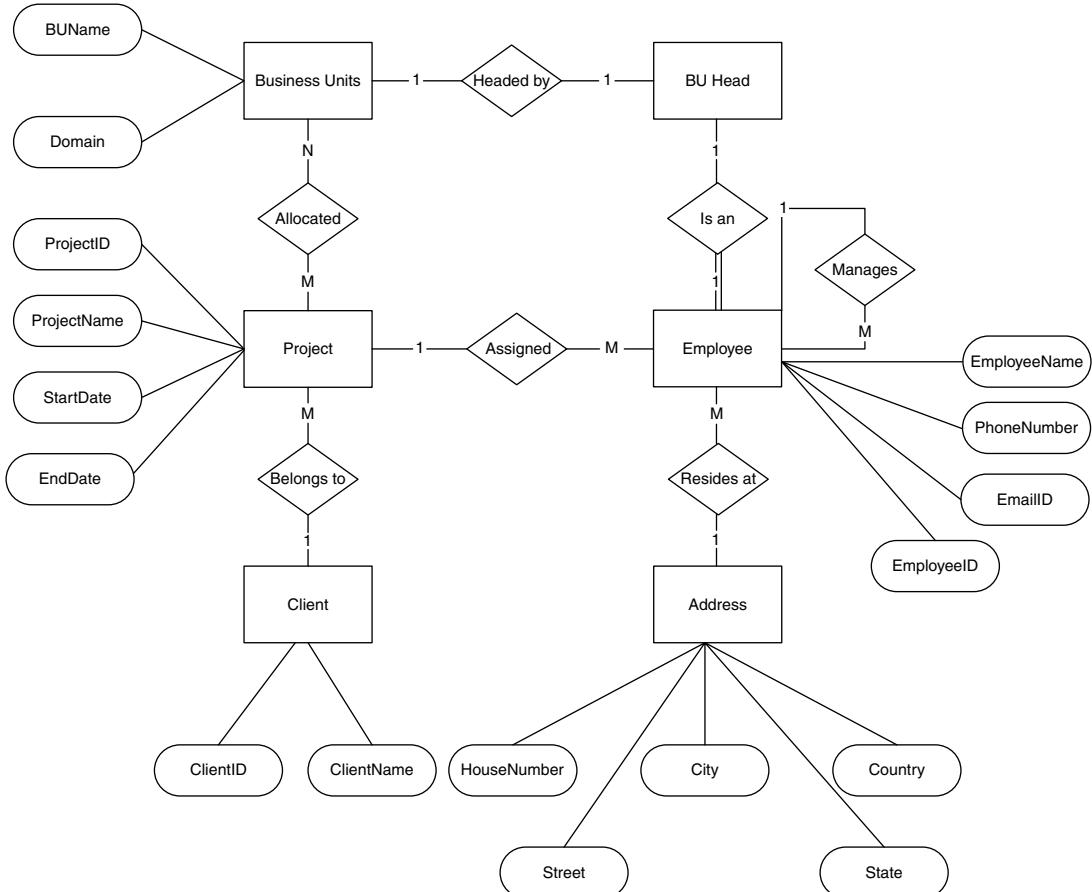
Let us take a quick look at the pros and cons of ER modeling.

#### Pros:

- The ER diagram is easy to understand and is represented in a language that the business users can understand.
- It can also be easily understood by a non-technical domain expert.
- It is intuitive and helps in the implementation on the chosen database platform.
- It helps in understanding the system at a higher level.

#### Cons:

- The physical designs derived using ER model may have some amount of redundancy.
- There is scope for misinterpretations because of the limited information available in the diagram.



**Figure 7.4** ER model for “InfoMechanists”.

### 7.4.2 Dimensional Modeling

Let us understand why dimensional modeling is required.

#### Picture this...

You have just reached the Bangalore International Airport. You are an Indian national due to fly to London, Heathrow International Airport. You have collected your boarding pass. You have two bags that you would like checked in. The person at the counter asks for your boarding pass, weighs the bags, pastes the label with details about your flight number, your name, your travel date, source airport code, and destination airport code, etc. He then pastes a similar label at the back of your boarding pass. This done, you proceed to the Immigration counter, passport, and boarding pass in hand. The person at the immigration counter verifies your identity, visa, boarding pass, etc. and then stamps the immigration seal with the current date on your passport. Your next stop is the security counter. The security personnel scrutinize your boarding pass, passport, etc. And you find yourself in the queue to board the aircraft. Again quick, careful rounds of verification by the aircraft crew before you find yourself ensconced in

your seat. You must be wondering what has all this got to do with multidimensional modeling. Well, we are trying to understand multidimensional perspectives of the same data. The data here is our “boarding pass”. Your boarding pass is looked at by different personnel for different reasons:

- The person at the check-in counter needs your boarding pass to book your check-in bags.
- The immigration personnel looked at your boarding pass to ascertain the source and destination of your itinerary.
- The security personnel scrutinized your boarding pass for security reasons to verify that you are an eligible traveller.
- The aircraft crew looked at your boarding pass to onboard you and guide you to your seat.

This is nothing but multidimensional perspectives of the same data. To put it simply, “Same Data, Multiple Perspectives”. To help with this multidimensional view of the data, we rely on dimensional modeling.

### **Consider another scenario...**

An electronic gadget distributor company “ElectronicsForAll” is based out of Delhi, India. The company sells its products in north, north-west, and western regions of India. They have sales units at Mumbai, Pune, Ahmedabad, Delhi, and Punjab. The president of the company wants the latest sales information to measure the sales performance and to take corrective actions if required. He has requested this information from his business analysts. He is presented with the report as below:

#### **Sales Report of “ElectronicsForAll”:**

##### **The number of units sold: 113**

Even though the data in the above sales report is correct, it is not able to convey any useful information to the president as he cannot view the data from any perspective.

Now to enhance the understanding of the data, the same data is presented by adding a perspective of time as shown below:

#### ***Sales Report of “ElectronicsForAll”:***

The number of units sold over time:

| January | February | March | April |
|---------|----------|-------|-------|
| 14      | 41       | 33    | 25    |

The above data conveys the information to a certain extent, but still it does not give a complete picture of the scenario. Now let us add yet another perspective, i.e. product, to the data, and the meaning of the data gets further enhanced.

#### **Sales Report of “ElectronicsForAll”:**

**The number of items sold for each product over time:**

| Product        | Jan | Feb | Mar | Apr |
|----------------|-----|-----|-----|-----|
| Digital Camera |     |     | 6   | 17  |
| Mobile Phones  | 6   | 16  | 6   | 8   |
| Pen Drives     | 8   | 25  | 21  |     |

Similar to the previous two cases, the meaning of the data can be further enriched by adding the region as another perspective as shown in the following table:

**Sales Report of “ElectronicsForAll”:**  
**The number of items sold in each region for each product over time:**

|        |                | Jan<br>(Units) | Feb<br>(Units) | Mar<br>(Units) | Apr<br>(Units) |
|--------|----------------|----------------|----------------|----------------|----------------|
| Mumbai | Digital Camera |                |                | 3              | 10             |
|        | Mobile Phones  | 3              | 16             | 6              |                |
|        | Pen Drives     | 4              | 16             | 6              |                |
| Pune   | Digital Camera |                |                | 3              | 7              |
|        | Mobile Phones  | 3              |                |                | 8              |
|        | Pen Drives     | 4              | 9              | 15             |                |

This method of analyzing a performance measure (in this case the number of units sold) by looking at it through various perspectives, or in other words the contextualized representation of a business performance measure, is known as dimensional modeling.

Dimensional modeling is a logical design technique for structuring data so that it is intuitive to business users and delivers fast query performance. Dimensional modeling is the first step towards building a dimensional database, i.e. a data warehouse. It allows the database to become more understandable and simpler. In fact, the dimensional database can be viewed as a cube having three or more dimensional/perspectives for analyzing the given data.

Dimensional modeling divides the database into two parts: (a) Measurement and (b) Context. Measurements are captured by the various business processes and other source systems. These measurements are usually numeric values called facts. Facts are enclosed by various contexts that are true at the moment the facts are recorded. These contexts are intuitively divided into independent logical clumps called *dimensions*. Dimensions describe the “*who, what, when, where, why, and how*” context of the measurements.

To better understand the fact (measurement)–dimension (context) link, let us take the example of booking an airlines ticket. In this case, the facts and dimensions are as given below:

**Facts** – Number of tickets booked, amount paid, etc.

**Dimensions** – Customer details, airlines, time of booking, time of travel, origin city, destination city, mode of payment, etc.

## ***Benefits of Dimensional Modeling***

- Comprehensibility:
  - Data presented is more subjective as compared to objective nature in a relational model.
  - Data is arranged in a coherent category or dimensions to enable better comprehension.
- Improved query performance.
- Trended for data analysis scenarios.

## **7.5 FACT TABLE**

A fact table consists of various measurements. It stores the measures of business processes and points to the lowest detail level of each dimension table. The measures are factual or quantitative in representation and are generally numeric in nature. They represent the *how much* or *how many* aspect of a question. For example, price, product sales, product inventory, etc.

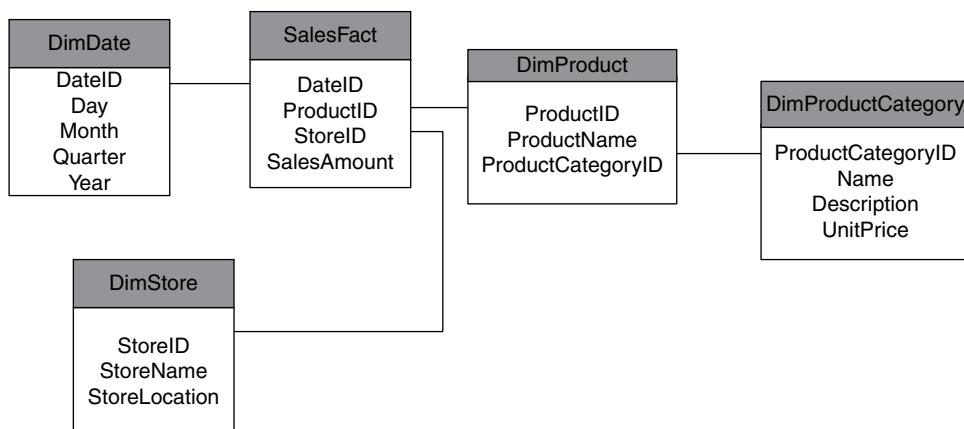
### **7.5.1 Types of Fact**

#### ***Additive Facts***

These are the facts that can be summed up/aggregated across all dimensions in a fact table. For example, discrete numerical measures of activity – quantity sold, dollars sold, etc.

Consider a scenario where a retail store “Northwind Traders” wants to analyze the revenue generated. The revenue generated can be in terms of product; it can be over a period of time; it can be across different regions; it can be by the employee who is selling the products; or it can be in terms of any combination of multiple dimensions. (Product, time, region, and employee are the dimensions in this case.) The revenue, which is a fact, can be aggregated along any of the above dimensions to give the total revenue along that dimension. Such scenarios where the fact can be aggregated along all the dimensions make the fact a fully additive or just an additive fact. Here revenue is the additive fact.

Figure 7.5 depicts the “SalesFact” fact table along with its corresponding dimension tables. This fact table has one measure, “SalesAmount”, and three dimension keys, “DateID”, “ProductID”, and “StoreID”. The purpose of the “SalesFact” table is to record the sales amount for each product in each store on a daily basis. In this table, “SalesAmount” is an additive fact because we can sum up this fact along any of the three



**Figure 7.5** An example of additive fact table.

dimensions present in the fact table, i.e. “DimDate”, “DimStore”, and “DimProduct”. For example, the sum of “SalesAmount” for all 7 days in a week represents the total sales amount for that week.

### **Semi-Additive Facts**

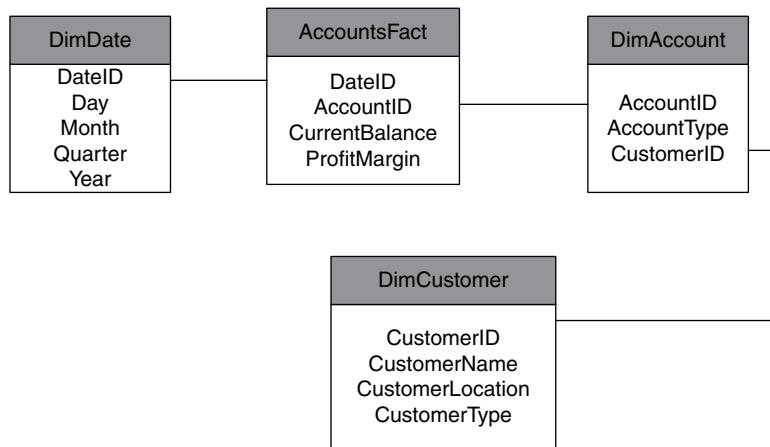
These are the facts that can be summed up for some dimensions in the fact table, but not all (e.g., account balances, inventory level, distinct counts, etc.).

Consider a scenario where the “Northwind Traders” warehouse manager needs to find the total number of products in the inventory. One inherent characteristic of any inventory is that there will be incoming products to the inventory from the manufacturing plants and outgoing products from the inventory to the distribution centers or retail outlets. So if the total products in the inventory need to be found out, say, at the end of a month, it cannot be a simple sum of the products in the inventory of individual days of that month. Actually, it is a combination of addition of incoming products and subtraction of outgoing ones. This means the inventory level **cannot** be aggregated along the “time” dimension. But if a company has warehouses in multiple regions and would like to find the total products in inventory across those warehouses, a meaningful number can be arrived at by aggregating inventory levels across those warehouses. This simply means inventory levels can be aggregated along the “region” dimension. Such scenarios where a fact can be aggregated along some dimensions but **not** along all dimensions give rise to semi-additive facts. In this case, the number of products in inventory or the inventory level is the semi-additive fact.

Let us discuss another example of semi-additive facts. Figure 7.6 depicts the “AccountsFact” fact table along with its corresponding dimension tables. The “AccountsFact” fact table has two measures: “CurrentBalance” and “ProfitMargin”. It has two dimension keys: “DateID” and “AccountID”. “CurrentBalance” is a semi-additive fact. It makes sense to add up current balances for all accounts to get the information on “what’s the total current balance for all accounts in the bank?” However, it does not make sense to add up current balances through time. It does not make sense to add up all current balances for a given account for each day of the month. Similarly, “ProfitMargin” is another non-additive fact, as it does not make sense to add profit margins at the account level or at the day level.

### **Non-Additive Facts**

These are the facts that cannot be summed up for any of the dimensions present in the fact table (e.g. measurement of room temperature, percentages, ratios, factless facts, etc.). Non-additive facts cannot be



**Figure 7.6** An example of semi-additive fact.

added meaningfully across any dimensions. In other words, non-additive facts are facts where the SUM operator cannot be used to produce any meaningful results. The following illustration will help you understand why room temperature is a non-additive fact.

| <i>DATE (Time)</i>          | <i>Temperature</i>           |
|-----------------------------|------------------------------|
| 5 <sup>th</sup> May (7 AM)  | 27°C                         |
| 5 <sup>th</sup> May (12 PM) | 33°C                         |
| 5 <sup>th</sup> May (5 PM)  | 10°C                         |
| Sum                         | 70°C (Non-meaningful result) |
| Average                     | 23.3°C (Meaningful result)   |

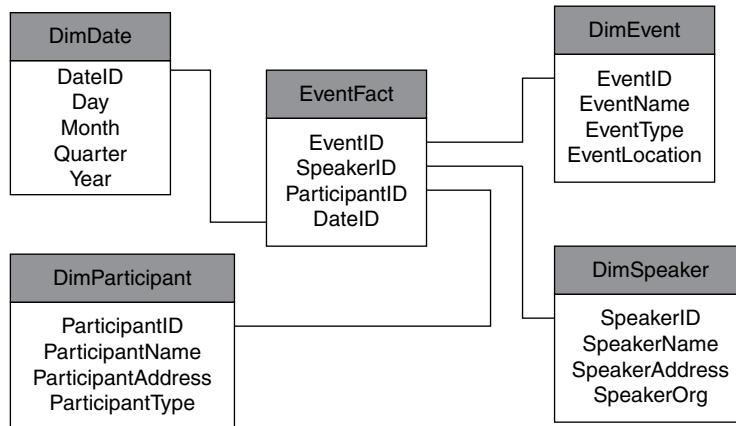
### Examples of non-additive facts are:

- **Textual facts:** Adding textual facts does not result in any number. However, counting textual facts may result in a sensible number.
- **Per-unit prices:** Adding unit prices does not produce any meaningful number. For example, the unit sales price or unit cost is strictly non-additive. But these prices can be multiplied with the number products sold and can be depicted as total sales amount or total product cost in the fact table.
- **Percentages and ratios:** A ratio, such as gross margin, is non-additive. Non-additive facts are usually the result of ratio or other calculations, such as percentages.
- **Measures of intensity:** Measures of intensity such as the room temperature are non-additive across all dimensions. Summing the room temperature across different times of the day produces a totally non-meaningful number.
- **Averages:** Facts based on averages are non-additive. For example, average sales price is non-additive. Adding all the average unit prices produces a meaningless number.
- **Factless facts (event-based fact tables):** Event fact tables are tables that record events. For example, event fact tables are used to record events such as Web page clicks and employee or student attendance. In an attendance recording scenario, attendance can be recorded in terms of “yes” or “no” OR with pseudo facts like “1” or “0”. In such scenarios, we can count the values but adding them will give invalid values. Factless facts are generally used to model the many-to-many relationships or to track events that did or did not happen.

Figure 7.7 depicts an example of a “factless fact table” – “EventFact”. This factless fact table has four dimension keys: “EventID”, “SpeakerID”, “ParticipantID”, and “DateID”. It does not have any measures or facts. This table can be queried to get details on the events that are the most popular. It can further be used to track events that did not happen. We can also use this table to elicit information about events that were the least popular or that were not attended.

### An Example of Multidimensional Modeling

Alex is excited. He will be travelling to the USA for business-related work. He has carefully planned his itinerary. Before embarking on the journey, he wants to check the weather in various US cities. He has

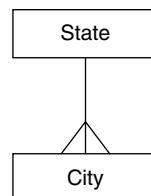


**Figure 7.7** An example of factless fact table.

searched the Internet to get the required information for the coming week. He has a table of data before him which looks like as shown below:

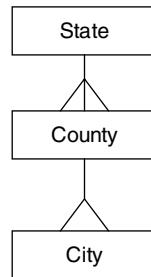
| Name of City  | DateDetails | MinTemp | MaxTemp |
|---------------|-------------|---------|---------|
| Los Angeles   | 22 May 2011 | 86      | 105     |
| San Francisco | 22 May 2011 | 78      | 107     |
| Phoenix       | 22 May 2011 | 88      | 98      |
| Los Angeles   | 23 May 2011 | 82      | 106     |
| San Francisco | 23 May 2011 | 76      | 104     |
| Phoenix       | 23 May 2011 | 86      | 96      |

In the above table, we have two dimensions, say, the “Geography” dimension and the “Time” dimension. “NameofCity” and “DateDetails” are attributes of the geography and time dimension, respectively. There are also two facts, “MinTemp” and “MaxTemp”. Using this table, it is possible to find out information about the maximum daily temperatures and the minimum daily temperatures for any group of cities or group of days. Now let us assume that we wish to view the maximum and minimum temperatures for states. A city belongs to a state. Let us add an attribute “State” to the “Geography” dimension. The relationship between the state and the city is as depicted in the following figure:



A state can have multiple cities. The relationship is one-to-many from the state to cities. Now assume that we wish to have a look at the minimum and maximum temperatures by counties. This can be

achieved by adding yet another attribute “County” to the geography dimension. The relationship between the state and county is as depicted in the following figure. The relationship is again one-to-many from the state to counties.



You already know that temperature is a non-additive fact. However, one can look at the average of temperatures for cities or states or for different time periods or for a combination of geography and time.

## 7.6 DIMENSION TABLE

Dimension tables consist of dimension attributes which describe the dimension elements to enhance comprehension. Dimension attributes (descriptive) are typically static values containing textual data or discrete numbers which behave as text values. Their main functionalities are: query filtering/constraining and query result set labeling. The dimension attribute must be

- **Verbose:** Labels must consist of full words.
- **Descriptive:** The dimension attribute names must be able to convey the purpose of the dimension element in as few and simple words as possible.
- **Complete:** Dimension attributes must not contain missing values.
- **Discrete values:** Dimension attributes must contain only one value per row in dimension table.
- **Quality assured:** Dimension attributes must not contain misspelt values or impossible values.

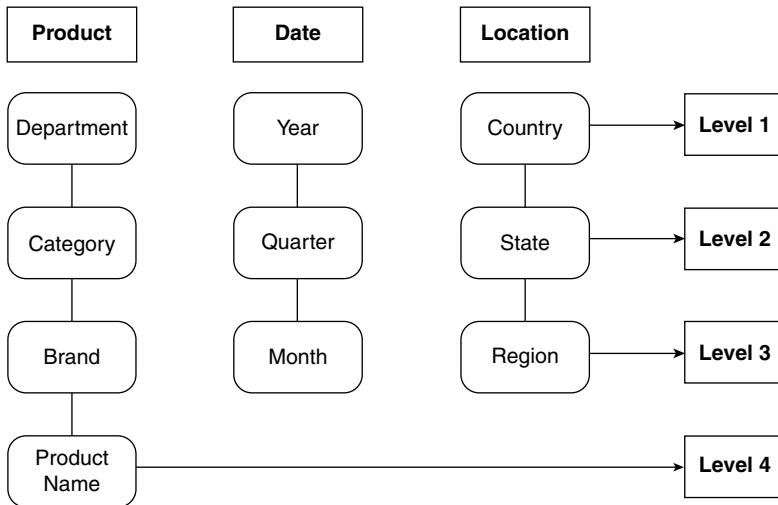
### 7.6.1 Dimension Hierarchies

A dimension hierarchy is a cascaded series of many-to-one relationships and consists of different levels. Each level in a hierarchy corresponds to a dimension attribute. Hierarchies document the relationships between different levels in a dimension.

A dimension hierarchy may also be described as a set of parent–child relationships between attributes present within a dimension. These hierarchy attributes, also known as levels, roll up from a child to parent. For example, Customer totals can roll up to Sub-region totals which can further roll up to Region totals. A better example would be – daily sales could roll up to weekly sales, which further roll up to month to quarter to yearly sales.

Let us understand the concept of hierarchy through the example depicted in Figure 7.8 In this example, the Product hierarchy is like this

Department → Category → Brand → Product Name



**Figure 7.8** An example for dimension hierarchies.

Similarly, the Date hierarchy is depicted as

Year → Quarter → Month

Example: 2011 → Q1 → April

For a better idea of dimension hierarchy, let us assume a product store, “ProductsForAll”. The store has several departments such as “Confectionary”, “Electronics”, “Travel Goods”, “Home Appliances”, “Dairy Products”, etc. Each department is further divided into categories. Example “Dairy Products” is further classified into “Milk”, “Butter”, “Cottage Cheese”, “Yogurt”, etc. Each product class offers several brands such as “Amul”, “Nestle”, etc. And, finally each brand has specific product names. For example, “Amul cheese” has names such as “Amul Slim Cheese”, “Amul EasySpread”, etc.

## 7.6.2 Types of Dimension Tables

- Degenerate Dimension
- Slowly Changing Dimension
- Rapidly Changing Dimension
- Role-playing Dimension
- Junk Dimension

### Degenerate Dimension

A degenerate dimension is a data that is dimension in temperament but is present in a fact table. It is a dimension without any attributes. Usually, a degenerate dimension is a transaction-based number. There can be more than one degenerate dimension in a fact table.

Degenerate dimensions often cause confusion as they don't feel or look like normal dimensions. They act as dimension keys in fact tables; however, they are not joined to corresponding dimensions in other dimension tables as all their attributes are already present in other dimension tables.

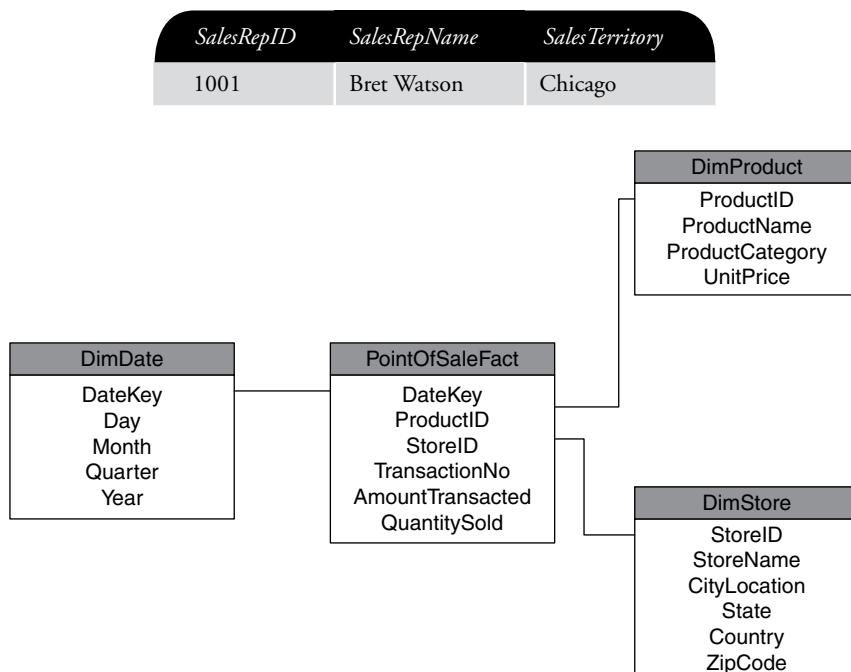
Degenerate dimensions can also be called textual facts, but they are not facts as the primary key for the fact table is often a combination of dimensional foreign keys and degenerate dimensions. As already stated, a fact table can have more than one degenerate dimension. For example, an insurance claim line fact table typically includes both claim and policy numbers as degenerate dimensions. A manufacturer can include degenerate dimensions for the quote, order, and bill of lading numbers in the shipments fact table.

Let us look at an example of degenerate dimension. Figure 7.9 depicts a PointOfSaleFact table along with other dimension tables. The “PointOfSaleFact” has two measures: AmountTransacted and QuantitySold. It has the following dimension keys: DateKey that links the “PointOfSaleFact” to “DimDate”, ProductID that links the “PointOfSaleFact” to “DimProduct”, and StoreID that links the “PointOfSaleFact” to “DimStore”. Here, TransactionNo is a degenerate dimension as it is a dimension key without a corresponding dimension table. All information/details pertaining to the transaction are extracted and stored in the “PointOfSaleFact” table itself; therefore, there is no need to have a separate dimension table to store the attributes of the transaction.

### ***Slowly Changing Dimension (SCD)***

In a dimension model, dimension attributes are not fixed as their values can change slowly over a period of time. Here comes the role of a slowly changing dimension. A slowly changing dimension is a dimension whose attribute/attributes for a record (row) change slowly over time, rather than change on a regular timely basis.

Let us assume a company sells car-related accessories. The company decides to assign a new sales territory, Los Angeles, to its sales representative, Bret Watson, who earlier operated from Chicago. How can you record the change without making it appear that Watson earlier held Chicago? Let us take a look at the original record of Bret Watson:



**Figure 7.9** An example of degenerate dimension.

Now the original record has to be changed as Bret Watson has been assigned “Los Angeles” as his sales territory, effective May 1, 2011. This would be done through a slowly changing dimension. Given below are the approaches for handling a slowly changing dimension:

### Type-I (Overwriting the History)

In this approach, the existing dimension attribute is overwritten with new data, and hence no history is preserved. This approach is used when correcting data errors present in a field, such as a word spelled incorrectly.

As per our example, the new record for Bret Watson after the change of his territory would look like:

| <i>SalesRepID</i> | <i>SalesRepName</i> | <i>SalesTerritory</i> |
|-------------------|---------------------|-----------------------|
| 1001              | Bret Watson         | Los Angeles           |

A disadvantage of managing SCDs by this method is that no historical records are kept in the data warehouse.

#### ***Advantages***

- It is the easiest and simplest approach to implement.
- It is very effective in those situations requiring the correction of bad data.
- No change is needed to the structure of the dimension table.

#### ***Disadvantages***

- All history may be lost in this approach if used inappropriately. It is typically not possible to trace history.
- All previously made aggregated tables need to be rebuilt.

### Type-II (Preserving the History)

A new row is added into the dimension table with a new primary key every time a change occurs to any of the attributes in the dimension table. Therefore, both the original values as well as the newly updated values are captured. In this method, the record for Bret Watson after the change of his territory would look like:

| <i>SalesRepID</i> | <i>SalesRepName</i> | <i>SalesTerritory</i> |
|-------------------|---------------------|-----------------------|
| 1001              | Bret Watson         | Chicago               |
| 1006              | Bret Watson         | Los Angeles           |

Type-II SCDs enable tracking of all the historical information accurately, hence these can have infinite number of entries due to the various types of changes.

#### ***Advantages***

This approach enables us to accurately keep track of all historical information.

### ***Disadvantages***

- This approach will cause the size of the table to grow fast.
- Storage and performance can become a serious concern, especially in cases where the number of rows for the table is very high to start with.
- It complicates the ETL process too.

### **Type-III (Preserving One or More Versions of History)**

This approach is used when it is compulsory for the data warehouse to track historical changes, and when these changes will happen only for a finite number of times. Type-III SCDs do not increase the size of the table as compared to the Type-II SCDs since old information is updated by adding new information. In this method, the record for Bret Watson after the change of his territory would look like:

| <i>SalesRepID</i> | <i>SalesRepName</i> | <i>OriginalSalesTerritory</i> | <i>CurrentSalesTerritory</i> | <i>EffectiveFrom</i> |
|-------------------|---------------------|-------------------------------|------------------------------|----------------------|
| 1001              | Bret Watson         | Chicago                       | Los Angeles                  | 01-May-2011          |

### ***Advantages***

- Since only old information is updated with new information, this does not increase the size of the table.
- It allows us to keep some part of history.

### ***Disadvantages***

Type-III SCDs will not be able to keep all history where an attribute is changed more than once. For example, if Bret Watson is later assigned “Washington” on December 1, 2012, the Los Angeles information will be lost.

Table 7.1 presents a comparison of the three types of handling of slowly changing dimensions.

### ***Rapidly Changing Dimension***

We have seen how to handle very slow changes in the dimension, but what would happen if the changes occur more frequently? A dimension is considered to be a fast changing dimension, also called a rapidly changing dimension, if its one or more attributes change frequently and also in several rows.

For example, consider a customer table having around 1,00,000 rows. Assuming that on an average 10 changes occur in a dimension every year, then in one year the number of rows will increase to  $1,00,000 \times 10 = 10,00,000$ .

To identify a fast changing dimension, look for attributes having continuously variable values. Some of the fast changing dimension attributes have been identified as

- Age
- Income
- Test score
- Rating
- Credit history score

**Table 7.1** Comparison of the three types of handling of slowly changing dimensions

|  | Type-I   | Type-II   | Type-III   |
|--|--|---|--|
| <b>When to use</b>                         | <ul style="list-style-type: none"> <li>When the attribute change is simple.</li> <li>Tracking of history not required.</li> </ul>  | <ul style="list-style-type: none"> <li>To keep a track of all the historical changes.</li> </ul>  | <ul style="list-style-type: none"> <li>To keep a track of a finite number of historical changes.</li> </ul>  |
| <b>Advantage</b>                           | <ul style="list-style-type: none"> <li>Easiest and simplest to implement.</li> <li>Effective in performing data correction.</li> <li>No change in data structure.</li> </ul> | <ul style="list-style-type: none"> <li>Enables tracking of historical changes accurately.</li> <li>Can track infinite number of changes.</li> </ul> | <ul style="list-style-type: none"> <li>Does not increase the size as Type-II, as old information is updated.</li> <li>Keeps a part of the history, equivalent to the number of changes predictable.</li> </ul> |
| <b>Disadvantage</b>                        | <ul style="list-style-type: none"> <li>All history is lost if used inappropriately.</li> <li>Previous aggregated tables have to be remade.</li> </ul>                        | <ul style="list-style-type: none"> <li>Dimension table grows fast.</li> <li>Complicated ETL process to load the dimension model.</li> </ul>         | <ul style="list-style-type: none"> <li>No complete history, especially when change occurs very often.</li> <li>Risk of losing history or changing design if more history has to be traced.</li> </ul>          |
| <b>Impact on existing dimension tables</b> | <ul style="list-style-type: none"> <li>No impact. No change in table structure.</li> </ul>   | <ul style="list-style-type: none"> <li>No impact. No change in table structure.</li> </ul>  | <ul style="list-style-type: none"> <li>Dimension table is modified to accommodate additional columns.</li> <li>Number of columns based on number of changes to track.</li> </ul>                               |
| <b>Impact on pre-aggregation</b>           | <ul style="list-style-type: none"> <li>All pre-existing aggregations have to be re-built.</li> </ul>   | <ul style="list-style-type: none"> <li>No impact. Aggregate tables need not be re-built.</li> </ul>   | <ul style="list-style-type: none"> <li>All pre-existing aggregations have to be re-built.</li> </ul>   |
| <b>Impact on database size</b>             | <ul style="list-style-type: none"> <li>No impact on the size of the database.</li> </ul>   | <ul style="list-style-type: none"> <li>Accelerated growth as a new row is added every time a change occurs.</li> </ul>                              | <ul style="list-style-type: none"> <li>No impact as the data is only updated.</li> </ul>   |

- Customer account status
- Weight

One method of handling fast changing dimensions is to break off a fast changing dimension into one or more separate dimensions known as mini-dimensions. The fact table would then have two separate foreign keys – one for the primary dimension table and another for the fast changing attributes.

### **Role-Playing Dimension**

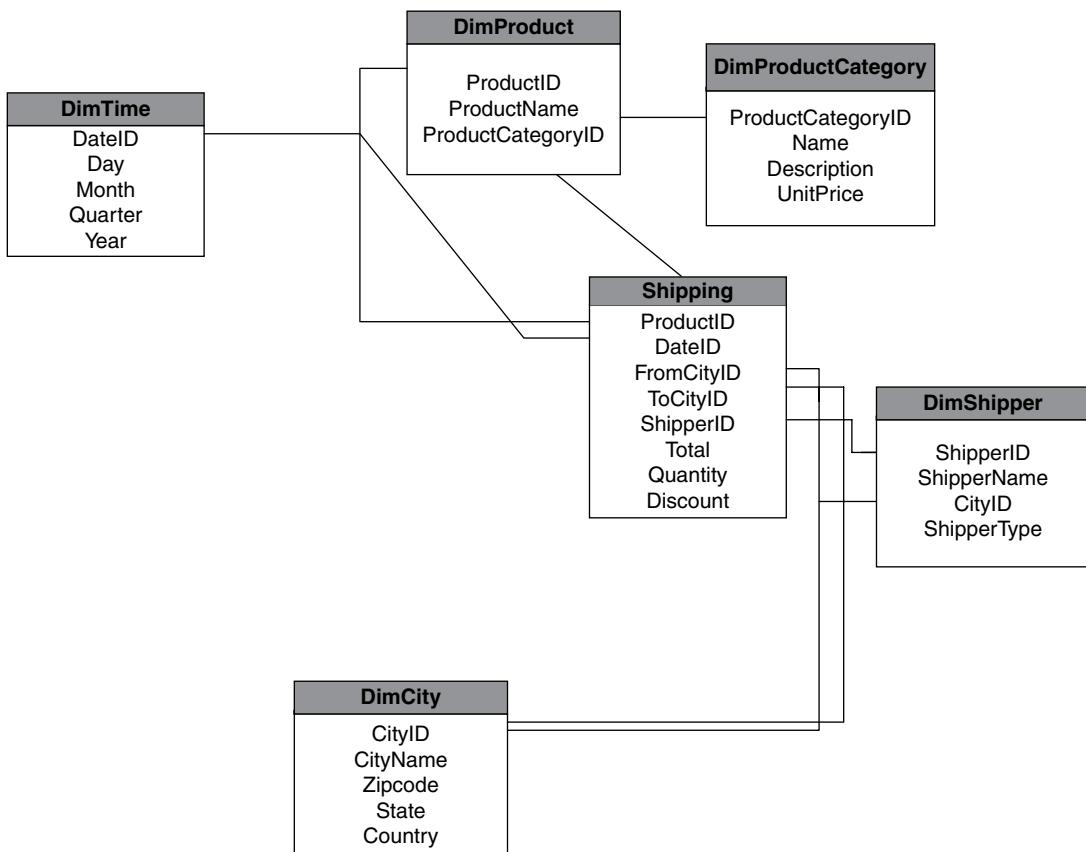
A single dimension that is expressed differently in a fact table with the usage of views is called a role-playing dimension. Consider an on-line transaction involving the purchase of a laptop. The moment an order is placed, an order date and a delivery date will be generated. It should be observed that both the dates are the attributes of the same time dimension. Whenever two separate analyses of the sales performance – one in terms of the order date and the other in terms of the delivery date – are required, two views of the same time dimension will be created to perform the analyses. In this scenario, the time dimension is called the role-playing dimension as it is playing the role of both the order and delivery dates.

Another example of the role-playing dimension is the broker dimension. The broker can play the role of both sell broker and buy broker in a share trading scenario. Figure 7.10 will help you have a better understanding of the role-playing dimension.

In Figure 7.10, “Shipping” is a fact table with three measures – “Total”, “Quantity”, and “Discount”. It has five dimension keys – “ProductID” that links the fact table “Shipping” with the “DimProduct” dimension table; “DateID” that links “Shipping” with the “DimTime” dimension table; “ShipperID” that links “Shipping” with the “DimShipper” dimension table; and the remaining two dimension keys, “ToCityID” and “FromCityID”, link the “Shipping” fact table with the same dimension table, i.e. “DimCity”. The two cities, as identified by the respective CityIDs, would have the same structure (DimCity) but would mean two completely different cities when used to signify FromCity and ToCity. This is a case of role-playing dimension.

### **Junk Garbage Dimension**

The garbage dimension is a dimension that contains low-cardinality columns/attributes such as indicators, codes, and status flags. The garbage dimension is also known as junk dimension. The attributes in a garbage dimension are not associated with any hierarchy.



**Figure 7.10** An example of role-playing dimension shown in represented the “Shipping” fact table.

We recommend going for junk/garbage dimension only if the cardinality of each attribute is relatively low, there are only a few attributes, and the cross-join of the source tables is too big. The option here will be to create a junk dimension based on the actual attribute combinations found in the source data for the fact table. This resulting junk dimension will include only combinations that actually occur, thereby keeping the size significantly smaller.

Let us look at an example from the healthcare domain. Shown below are two source tables and a fact table.

**CaseType (Source Table)**

| <i>CaseTypeID</i> | <i>CaseTypeDescription</i>                   |
|-------------------|--|
| 1                 | Referred by another hospital                 |
| 2                 | Walkin                                       |
| 3                 | Consultation                                 |
| 4                 | Transferred by a branch of the same hospital |

**TreatmentLevel (Source Table)**

| <i>TreatmentTypeID</i> | <i>TreatmentTypeDescription</i> |
|------------------------|---------------------------------|
| 1                      | ICU                             |
| 2                      | Pediatrics                      |
| 3                      | Orthopedic                      |
| 4                      | Ophthalmology                   |
| 5                      | Oncology                        |
| 6                      | Physiotherapy                   |

**CaseTreatmentFact (Fact Table)**

| <i>CaseTypeID</i> | <i>TreatmentTypeID</i> | <i>CountofPatients</i> |
|-------------------|------------------------|------------------------|
| 4                 | 1                      | 2                      |
| 1                 | 3                      | 3                      |
| 3                 | 4                      | 5                      |

A junk dimension will combine several low cardinality flags and attributes into a single table rather than modeling them as separate dimensions. This will help reduce the size of the fact table and make dimensional modeling easier to work with. In our example, each of the source tables [CaseType (CaseTypeID, CaseTypeDescription) and TreatmentLevel (TreatmentTypeID, TreatmentTypeDescription)] has only two attributes each. The cardinality of each attribute is also low.

One way to build the junk dimension will be to perform a cross-join of the source tables. This will create all possible combinations of attributes, even if they do not or might never exist in the real world.

The other way is to build the junk dimension based on the actual attribute combinations found in the source tables for the fact table. This will most definitely keep the junk dimension table significantly smaller since it will include only those combinations that actually occur. Based on this explanation, we redesign the fact table along with the junk dimension table as shown below:

**CaseTreatmentFact**

| <i>SurrogateKeyID</i> | <i>CountofPatients</i> |
|-----------------------|------------------------|
| 1                     | 2                      |
| 2                     | 3                      |
| 3                     | 5                      |

**CaseTreatmentJunk**

| <i>SurrogateKeyID</i> | <i>CaseTypeID</i> | <i>CaseTypeDescription</i>                   | <i>TreatmentTypeID</i> | <i>TreatmentTypeDescription</i> |
|-----------------------|-------------------|--|------------------------|---------------------------------|
| 1                     | 4                 | Transferred by a branch of the same hospital | 1                      | ICU                             |
| 2                     | 1                 | Referred by another hospital                 | 3                      | Orthopedic                      |
| 3                     | 3                 | Consultation                                 | 4                      | Ophthalmology                   |

## 7.7 TYPICAL DIMENSIONAL MODELS

As stated earlier, the Entity Relationship (ER) data model is a commonly used data model for relational databases. Here, the database schema is represented by a set of entities and the relationship between them. It is an ideal data model for On-Line Transaction Processing (OLTP).

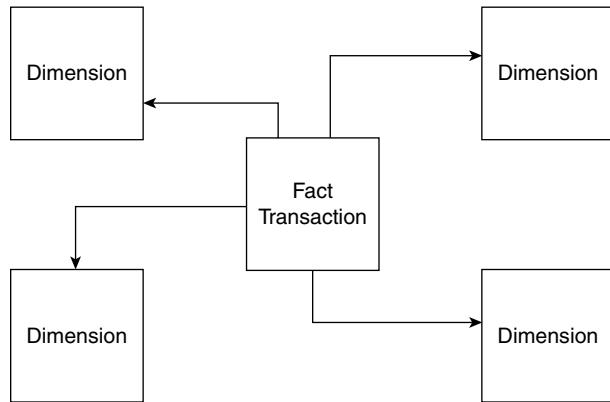
Let us look at a data model that is considered apt for On-Line Data Analysis. Multidimensional data modeling is the most popular data model when it comes to designing a data warehouse. Dimensional modeling is generally represented by either of the following schemas:

- Star Schema.
- Snowflake Schema.
- Fact Constellation Schema.

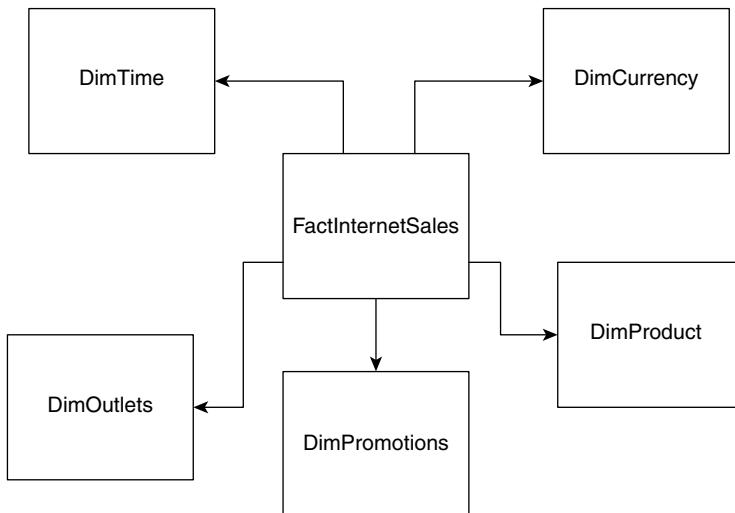
### 7.7.1 Star Schema

It is the simplest of data warehousing schema. It consists of a large central table (called the fact table) with no redundancy. The central table is being referred by a number of dimension tables. The schema graph looks like a starburst (Figure 7.11). The dimension tables form a radial pattern around the large central fact table. The star schema is always very effective for handling queries.

In the star schema, the fact table is usually in 3NF or higher form of normalization. All the dimension tables are usually in a de-normalized manner, and the highest form of normalization they are usually present in is 2NF. The dimension tables are also known as look up or reference tables.



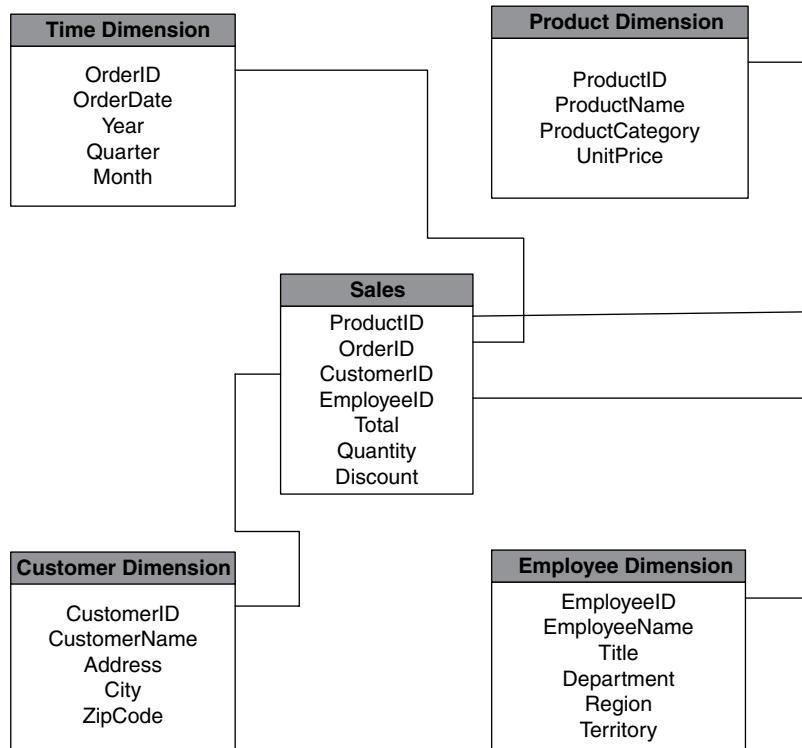
**Figure 7.11** The data model for star schema.



**Figure 7.12** The data model for “TenToTen” Stores in Star Schema.

Figure 7.12 shows the Star schema for “TenToTen” Stores. A few tables from the logical model of “TenToTen” Stores depicted in Figure 7.12 are selected in the construction of the star model for TenToTen Stores. The tables, “Product”, “SubCategory”, and “Category” are consolidated into the “DimProduct” table. The tables “ProductOffer” and “PromotionOffers” are consolidated into the “DimPromotions” table. Table “Date” is mapped as the “DimDate” dimension. The tables “Outlets”, “MarketType”, “Operation\_Type”, and “Territory” are modeled into the “DimOutlets” table. Finally, as payment can be made via several currencies, a new dimension “DimCurrency” is modeled to enable standardization.

Figure 7.13 shows the Star schema for “ElectronicsForAll”. Here, the sales are considered along four dimensions, i.e. Time, Product, Employee, and Customer. The schema diagram shows a central fact table for “Sales”. The “Sales” central fact table has the keys to each of the four dimensions along with three measures – Total, Quantity, and Discount. Each dimension is represented by only one table. Each



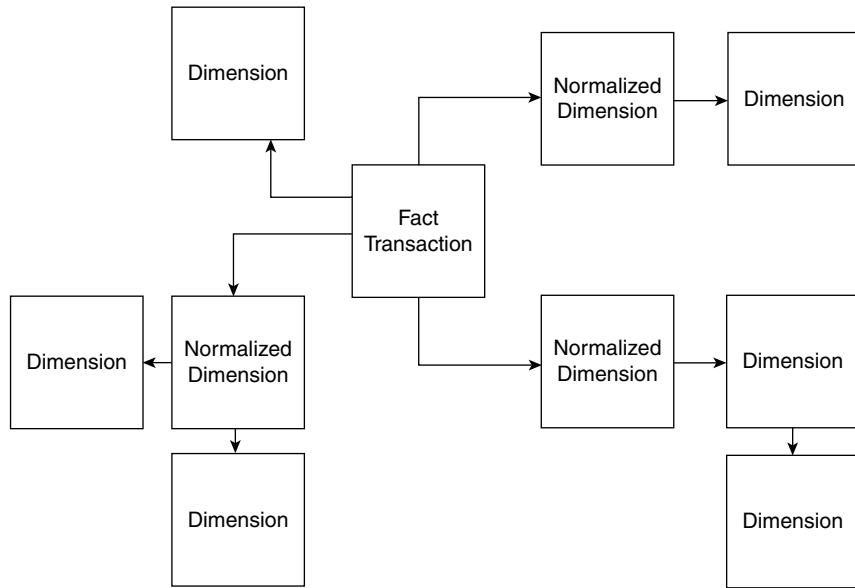
**Figure 7.13** Star schema for sales of “ElectronicsForAll”.

table further has a set of attributes. For example, the Product dimension table has these attributes – ProductID, ProductName, ProductCategory, and UnitPrice. Similarly, the Customer dimension table has the attributes – CustomerID, CustomerName, Address, City, and ZipCode.

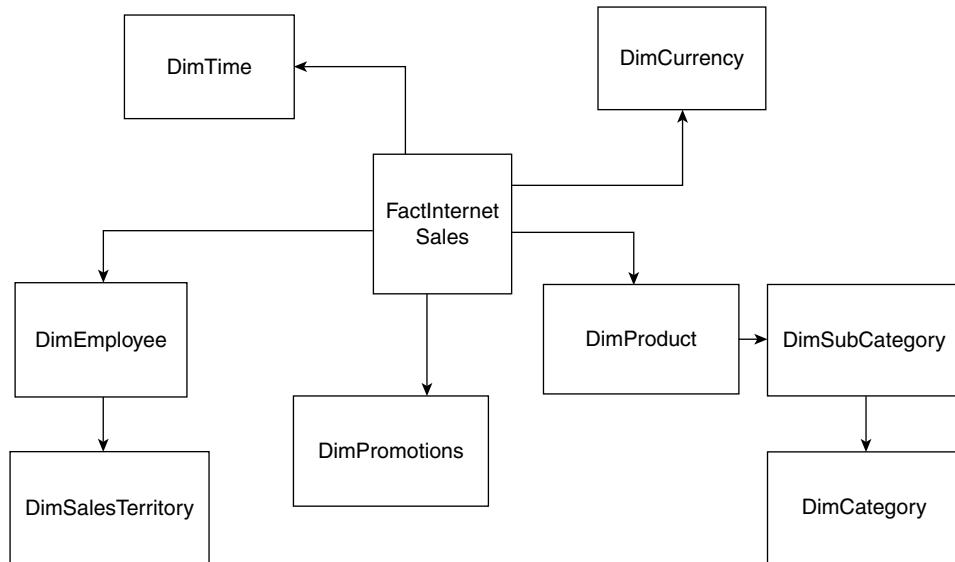
### 7.7.2 Snowflake Schema

The Snowflake schema is a variant of the Star schema. Here, the centralized fact table is connected to multiple dimensions. In the Snowflake schema, dimensions are present in a normalized form in multiple related tables (Figure 7.14). A snowflake structure materializes when the dimensions of a star schema are detailed and highly structured, having several levels of relationship, and the child tables have multiple parent tables. This “snowflaking” effect affects only the dimension tables and does not affect the fact table.

Figure 7.15 shows the **data model** for “TenToTen” Stores in Snowflake schema. Figure 7.16 depicts the snowflake schema for “ElectronicsForAll”. If you look carefully at Figures 7.13 and 7.16, there is no change in the “Sales” fact table. The difference lies in the definition of dimension tables. The single dimension table for “Employee” in the Star schema is normalized in the Snowflake schema, resulting in a new “Department” table. The “Employee” dimension table now contains the attributes – EmployeeID, EmployeeName, Title, DepartmentID, Region, Territory. The “DepartmentID” attribute links the “Employee” dimension table with the “Department” dimension table. The “Department” dimension

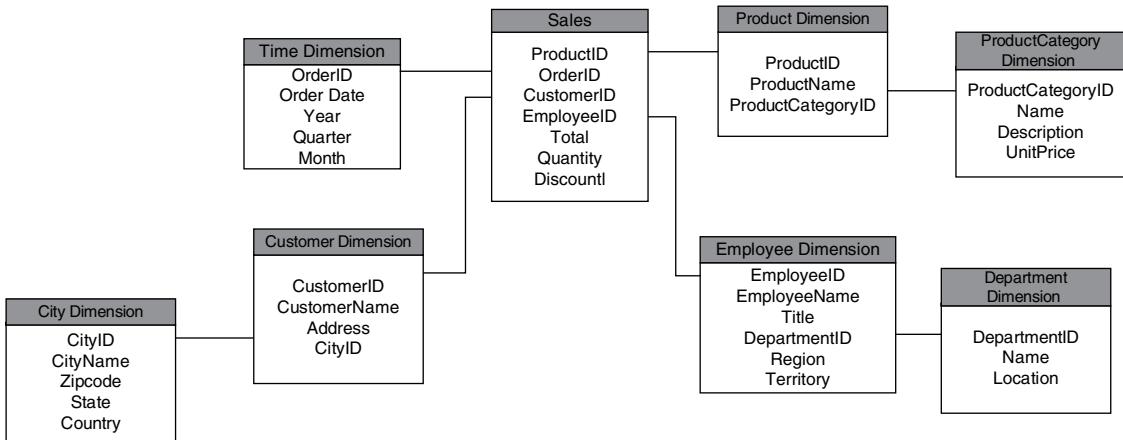


**Figure 7.14** The data model for Snowflake schema.



**Figure 7.15** The data model for “TenToTen” Stores in Snowflake schema.

table has details about each department, such as “Name” and “Location” of the department. Similarly, the single dimension table for “Customer” in the Star schema is normalized in the Snowflake schema, resulting in a new “City” table. The “Customer” dimension table now contains the attributes: CustomerID, CustomerName, Address, CityID. The “CityID” attribute links the “Customer” dimension



**Figure 7.16** Snowflake schema for sales of “ElectronicsForAll”.

table with the “City” dimension table. The “City” dimension table has details about each city such as “CityName”, “Zipcode”, “State”, and “Country”.

As we have in the example of “ElectronicsForAll”, the main difference between the Star and Snowflake schema is that the dimension tables of the Snowflake schema are maintained in normalized form to reduce redundancy. The advantage here is that such tables (normalized) are easy to maintain and save storage space. However, it also means that more joins will be needed to execute a query. This will adversely impact system performance.

### ***Identifying Dimensions to be Snowflaked***

In this section, we will observe the practical implementation of the dimensional design.

#### **What is snowflaking?**

The snowflake design is the result of further expansion and normalization of the dimension table. In other words, a dimension table is said to be snowflaked if the low-cardinality attributes of the dimensions have been divided into separate normalized tables. These tables are then joined to the original dimension table with referential constraints (foreign key constraints).

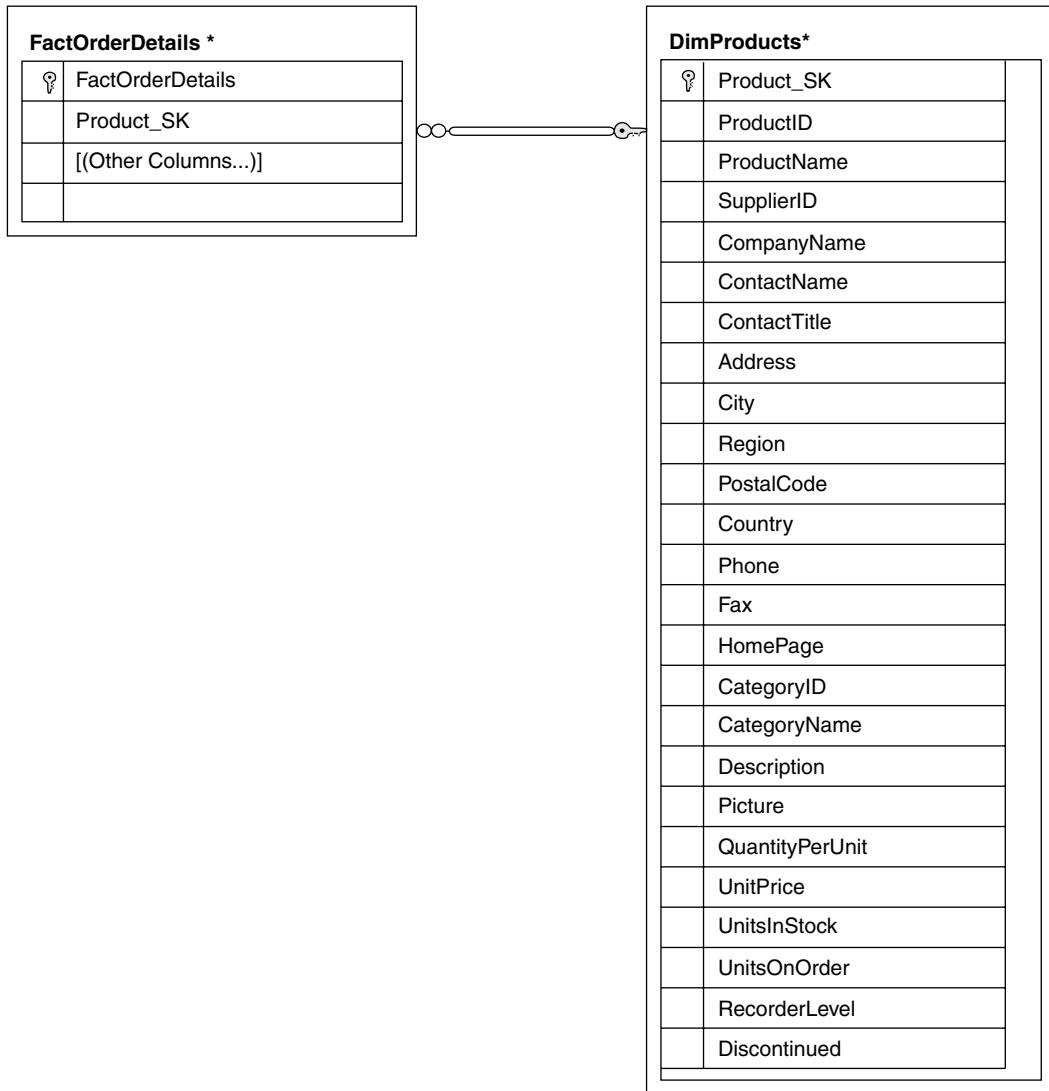
Generally, snowflaking is not recommended in the dimension table, as it hampers the understandability and performance of the dimensional model as more tables would be required to be joined to satisfy the queries.

#### **When do we snowflake?**

The dimensional model is snowflaked under the following two conditions:

- The dimension table consists of two or more sets of attributes which define information at different grains.
- The sets of attributes of the same dimension table are being populated by different source systems.

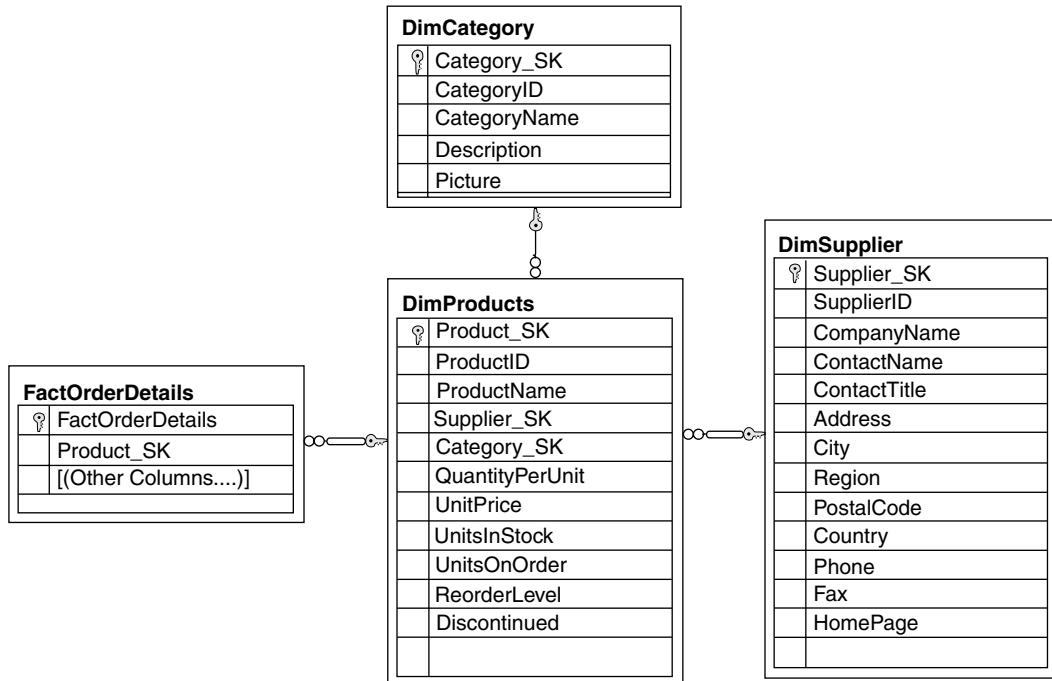
For understanding why and when we snowflake, consider the “Product” dimension table shown in Figure 7.17.



**Figure 7.17** The “Product” dimension table.

The “Product” table (Figure 7.17) has three sets of attributes related to the “Product”, “Category”, and “Suppliers”. These sets of attributes have different levels of grains (detail level) and are also populated by different source systems. The “Product” dimension table is a perfect example for snowflaking for the following two reasons:

- The product table represents three different sets of attributes. One set shows “Product” attributes, the second set contains the “Category” attributes, and the third set shows the “Suppliers” attributes. The level of detail (granularity) for the three sets is very different from each other.
- On a detailed analysis, we observe that the “Products” attributes are populated by the OLTP (On-Line Transaction Processing) database while the other two attributes are populated from



**Figure 7.18** The data model for the “Product” table in Snowflaked schema.

the external consultancy firms. It may be possible that another source system in the enterprise is responsible for supplying some of the other attributes.

The “Product” table after it has been snowflaked is shown in Figure 7.18. Due to Snowflaking, dimension table low-cardinality attributes have been divided into separate normalized tables called “DimCategory” and “DimSupplier”.

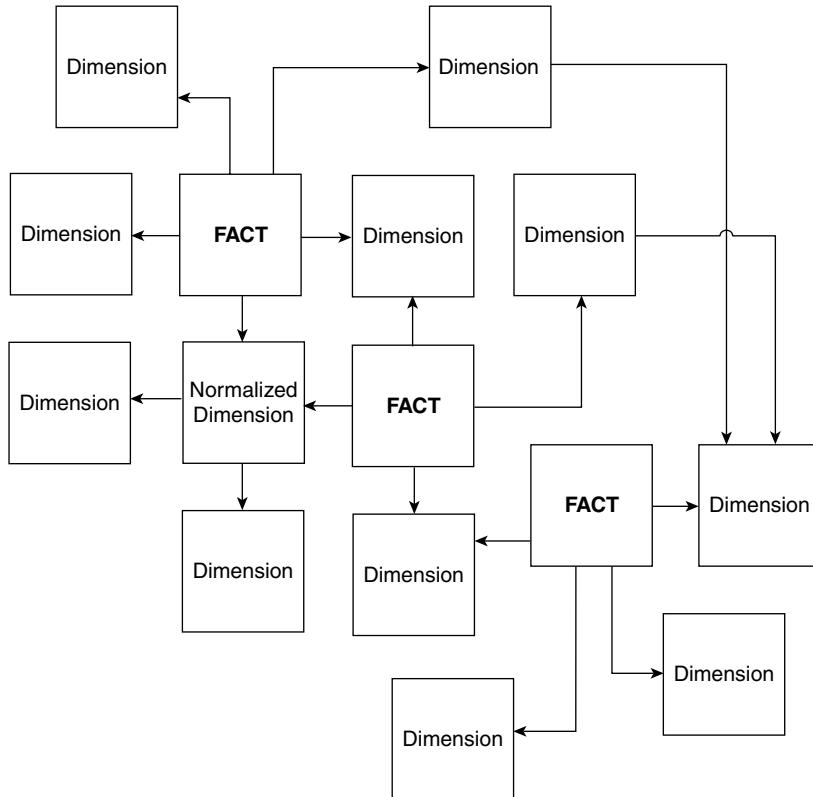
### When NOT to snowflake?

Normally, you should avoid snowflaking or normalization of a dimension table, unless required and appropriate. Snowflaking reduces space consumed by dimension tables, but compared with the entire data warehouse the saving is usually insignificant.

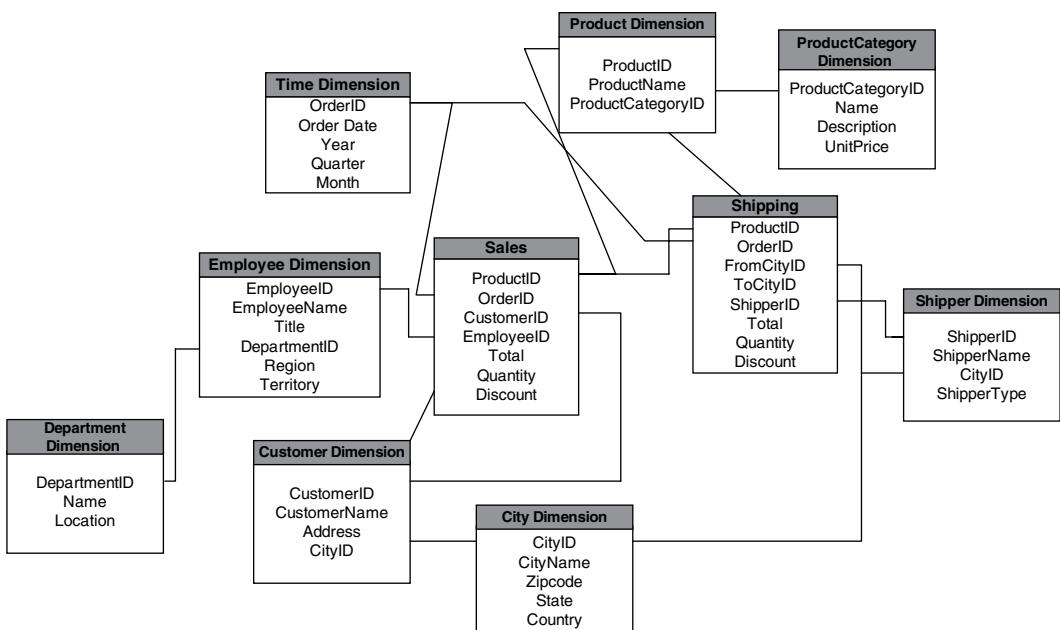
Do not snowflake hierarchies of one dimension table into separate tables. Hierarchies should belong to the dimension table only and should never be snowflaked. Multiple hierarchies can belong to the same dimension if the dimension has been designed at the lowest possible detail.

### 7.7.3 Fact Constellation Schema

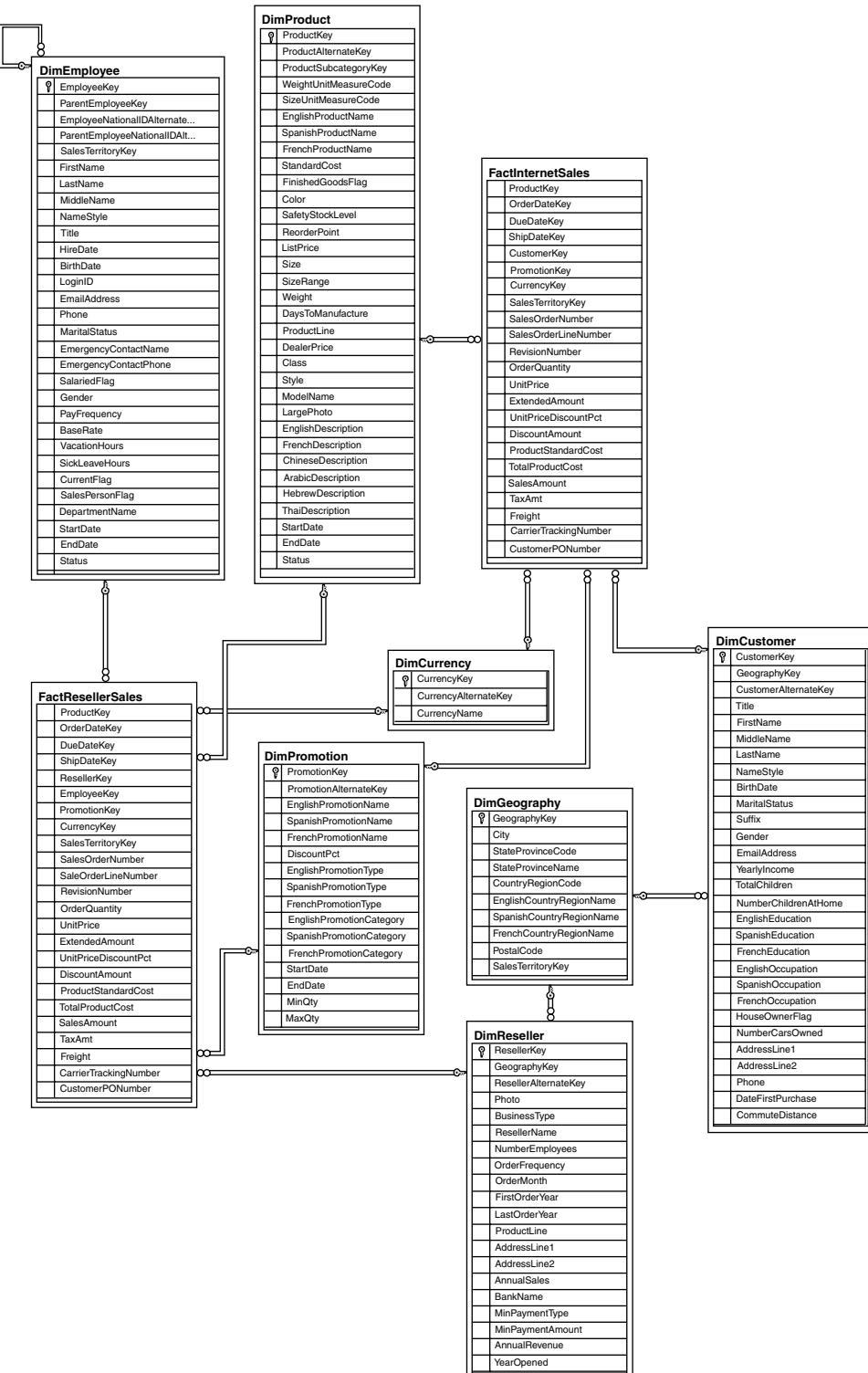
The constellation schema is shaped like a constellation of stars (i.e. Star schemas). This is more complex than Star or Snowflake schema variations, as it contains multiple fact tables. This allows the dimension tables to be shared among the various fact tables. It is also called “Galaxy schema”. The main disadvantage of the fact constellation is more complicated design because multiple aggregations must be taken into consideration (Figure 7.19).



**Figure 7.19** The data model for Fact Constellation schema.



**Figure 7.20** Fact constellation schema of a data warehouse for “Sales” and “Shipping”.



**Figure 7.21** The data model for Fact Constellation schema of “TenToTen” Stores.

Figure 7.20 depicts the Fact Constellation schema for two fact tables “Sales” and “Shipping” of a data warehouse. The “Sales” table is identical to the “Sales” table shown in Figures 7.13 and 7.16. Let us take a look at the “Shipping” fact table. The “Shipping” table has five dimensions or keys – “ProductID”, “OrderID”, “FromCityID”, “ToCityID”, and “ShipperID”. Further, it has three measures – “Total”, “Quantity”, and “Discount”. The dimension tables “Time”, “Product”, and “City” are shared between the two fact tables “Sales” and “Shipping”. As stated earlier, the fact constellation allows the dimension tables to be shared between fact tables.

One question we still have to answer is which model (Star, Snowflake, or Fact Constellation) is preferable for an enterprise-wide data warehouse and for data marts. And the answer is...

An enterprise-wide data warehouse collects information about subjects that span the entire organization. Therefore, the Fact Constellation schema is the preferred schema for enterprise-wide data warehouse. The Fact Constellation schema can model multiple, inter-related subjects.

Data marts, on the other hand, focus on selected subjects, and therefore their scope is department-wide. Both the Star and Snowflake schemas are commonly used as data models for data marts. Figure 7.21 depicts the data model for Fact Constellation schema of “TenToTen” Stores.

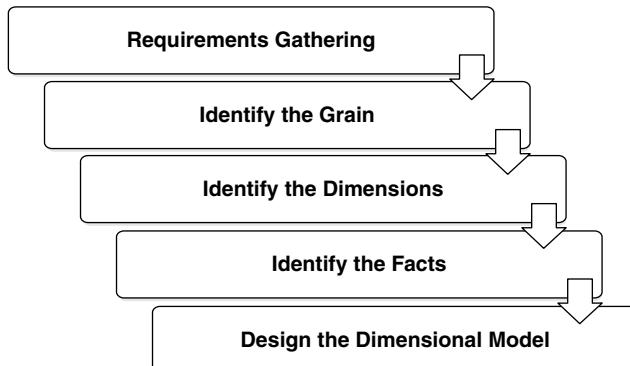
Now that we are familiar with dimensions, facts and dimensional models, let us see the evolution of dimensional model right from identifying the requirements to analyzing the performance of a business process to designing the model to actually analyzing the performance.

## 7.8 DIMENSIONAL MODELING LIFE CYCLE

In this section, we will discuss the process followed while designing a dimensional model. Designing a suitable dimensional model can be a difficult task as requirements are typically difficult to define. Many a time only after seeing the result of a data model we are able to decide whether it satisfies our requirement or not. Also, the organizational requirements may change over time.

But, where should we start designing the model? What should be the first step? To help in this decision process, we have a dimensional modeling life cycle which consists of the following phases which are also depicted in Figure 7.22:

- Requirements Gathering.
- Identifying the Grain.



**Figure 7.22** Dimensional modeling life cycle.

- Identifying Dimensions.
- Identifying Facts.
- Designing the Dimensional Model.

### 7.8.1 Requirements Gathering

Requirements gathering is a process of selecting the business processes for which the dimensional model will be designed. Based on this selection, the requirements for the business process are gathered and documented. Hence it can be said that requirements gathering focuses on the study of business processes and information analysis actions in which users are involved.

While doing requirements gathering, it is important to focus on two key elements of analysis:

- What is to be analyzed?
- The evaluation criteria.

A requirements gathering process, thus, is extremely oriented toward understanding the domain for which the modeling is to be done. Typically, requirements at this stage are documented rather informally or, at least, they are not represented in detailed schemas. There are two methods for deriving business requirements:

- Source-driven requirements gathering.
- User-driven requirements gathering.

#### ***Source-Driven Requirements Gathering***

Source-driven requirements gathering is a method in which requirements are determined by using the source data in production operational systems. This is done by analyzing the ER model of source data if it is available or the physical record layouts, and selecting data elements that seem to be of interest.

**Advantages:** The major advantages of this approach are:

- From the beginning you know what data you can supply because you limit yourself to what is available.
- The time spent on gathering requirements from the users in the early stages of the project is saved. However, there is no alternative to the importance and value you get when you involve the users in the requirements gathering phase.

**Disadvantages:** Listed below are a few disadvantages of this approach.

- Minimal user involvement increases the risk of producing an incorrect set of requirements.
- Depending on the volume of source data you have, and the availability of ER models for it, this can also be a very time-consuming approach.
- Perhaps some of the user's key requirements may need the data that is currently unavailable. Without identifying these requirements, there is no chance to investigate what external data is required. (External data is data that exists outside the enterprise. External data can be of significant value to the business users.)

Simply put, the source-driven approach provides only what you have. This approach can be appropriate in two cases. First, relative to dimensional modeling, it can be used to develop a fairly comprehensive list of the major dimensions that are of interest to the enterprise. So if you plan to have an enterprise-wide

data warehouse, this approach could minimize duplicate dimensions across separately developed data marts. Second, analyzing relationships in the source data helps you identify areas on which your data warehouse development efforts should be focused.

### **User-Driven Requirements Gathering**

The user-driven requirements gathering method is based on identifying the requirements by analyzing the functions that the users perform. This is generally done by conducting a series of meetings and/or interviews with users.

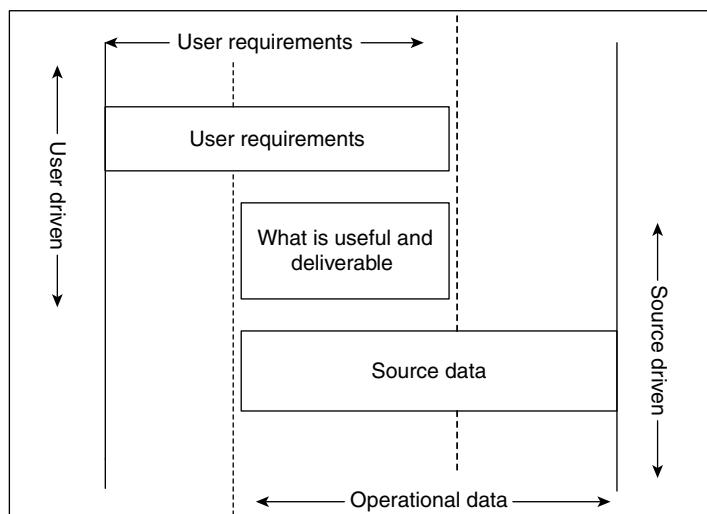
**Advantages:** The main benefits of this approach are:

- It focuses on providing what is really needed rather than what is available.
- This approach has a smaller scope than the source-driven approach; therefore, it generally produces a useful data in a shorter span of time.

**Disadvantages:**

- On the negative side, the users might have unrealistic expectations. They must clearly understand that some of the data they need might not be available due to various reasons.
- It is a time-consuming process as the user requirements may change over time.
- If a user is too focused, it is possible to miss useful data that is available in the production systems.

While using user-driven requirement gathering approach, we must try not to limit ourselves to the things asked by the user. Out-of-the-box thinking should be promoted while defining the requirements for a data warehouse. We believe that a user-driven requirement gathering is the approach of choice, especially while developing dependent data marts or populating data marts from a business-wide enterprise warehouse.



**Figure 7.23** Source-driven vs. user-driven approach.

Refer to Figure 7.23. The ideal approach is to design the data warehouse keeping in mind both the user requirements and the source data available such that the solution developed is deliverable as well as useful. While gathering the requirements, the questions should be asked so as to gather “who”, “when”, “where”, “what”, and “how” of the business model. The questions for this purpose could be as follows:

- Who are the organizations, groups, and people of interest?
- What functions are required to be analyzed?
- Why do we require data?
- When should the data be recorded?
- Where do relevant processes occur, geographically as well as organizationally?
- How is the performance of the processes being analyzed?
- How is the performance of the various business modules measured?
- What factors decide the success or failure?
- What is the method used for distribution of information? Is it SMS or email (examples)?
- What steps are presently taken to fulfil the information?
- What level of detailing would enable data analysis?

### 7.8.2 Identify the Grain

The second phase of the dimensional modeling life cycle is to identify the grain. To understand how to identify a grain first we need to understand what is meant by the term “grain”?

#### **What is a Grain?**

The “grain” refers to the level of detail or fineness to which the data can be analyzed. The grain in a dimensional model is the finest level of detail implied by the joining of the fact and dimension tables. To have a better idea of grain, let us consider the following tables with their attributes:

- **Date** (year, quarter, month, and day)
- **Store** (region, district, and store)
- **Product** (category name, price, and product)

For this group of tables, the grain is the product sold in a store in a day. Some other examples of grain are:

- A daily snapshot of the inventory levels for each product in a warehouse.
- A monthly snapshot for balance of each bank account.
- A bus ticket purchased on a day.
- A single item on an invoice.
- A single item on a restaurant bill.

**Granularity:** Granularity is defined as the detailed level of information stored in a table.

- The more the detail, the lower is the level of granularity.
- The lesser the detail, the higher is the level of granularity.

Choosing the right granularity is a very important step while designing a dimensional model for a data warehouse. Let us understand this through the following examples:

- If we consider a table **LocationA** (Country, State), it will have a granularity at the state level, i.e. it will have information at state level but will not contain information at the city level. But if we modify the same table and change it to **LocationB** (Country, State, City), it will now have a granularity at the city level, i.e. we can retrieve information at the city level using this table.
- Consider a table **dateA** (year, quarter). It has a granularity at the quarter level, i.e. it will contain information at the quarter level but will not have information for a particular day or month. Now, if we modify the table to **dateB** (year, quarter, month), it now has a granularity at the month level, but does not contain information at the day level.

So while designing a dimensional model, we should make sure that the granularity of all the tables (i.e. fact and dimension tables) should be same.

### 7.8.3 Identify the Dimensions

Once the grain has been identified, we shall proceed towards determining the dimensions for the data model. The key features of a dimension table are

- Dimension tables contain attributes that describe facts in the fact table.
- Each dimension table has only one lowest level of detail called the *dimension grain*, also referred to as the *granularity of the dimension*.
- Each non-key element (other than the surrogate key) should appear in a single dimension table.

### 7.8.4 Identify the Facts

After identifying the dimensions in the model, our next step is to identify the fact table and the relevant facts/measures in that table. Before identifying the relevant facts/measures we should identify the fact tables in the database design. The following features of fact table will help you identify the fact tables:

- The fact table will mostly contain numeric and additive values.
- It contains at least two foreign keys.
- It usually comprises vast number of records.
- Tables having many-to-many relationship in an ER model can be used as fact tables in a dimensional model.

Having identified the fact table, our next step is to identify the facts/measures for the same. While identifying the facts the following points should be kept in mind:

- A fact should generally be numeric in nature.
- It should be of importance to the business analysis.
- It should confirm the grain level identified in the step “Identify the grain”.

The essence of multidimensional data modeling is to prepare the tables and hence the associated data in a format suitable for analysis. Once the database is modeled multidimensionally, we can start analyzing the data from various perspectives.

## DESIGNING THE DIMENSIONAL MODEL

Draw a Star/Snowflake/Fact Constellation model.



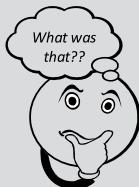
### Point Me

- *Dimensional Modeling: In a Business Intelligence Environment*, IBM.
- *Excel 2010: Data Analysis and Business Modeling*, Wayne L. Winston.



### Connect Me

- <http://www.stanford.edu/dept/itss/docs/oracle/10g/olap.101/b10333/multimodel.htm>
- <http://office.microsoft.com/en-us/excel-help/overview-of-pivottable-and-pivotchart-reports-HP010342752.aspx?CTT=1>



### Remind Me

#### Normalization (Entity Relationship) Model:

- Logical model that seeks to eliminate data redundancy.
- It depicts data relationships.
- Difficult to master and understand.
- Oriented towards storing and saving data rather than business user understanding.

#### Dimensional Model:

- Logical design that is used in data warehousing.
- Composed of a set of fact and dimension tables.
- Models data as a hypercube.
- Oriented towards faster retrieval of data, together with better understanding of data by business users.



## *Test Me Exercises*

Below are a set of data modeling terminologies which are jumbled. The activity is to find the hidden terms.

- |                       |   |       |
|-----------------------|---|-------|
| Resume as             | : | _____ |
| The sarcasm           | : | _____ |
| On and it moralize    | : | _____ |
| Self facts cast       | : | _____ |
| Grainy ultra          | : | _____ |
| Tube artist           | : | _____ |
| To concealed lump     | : | _____ |
| Is candid, clinging   | : | _____ |
| Compile dashingly     | : | _____ |
| Neediest or demeaning | : | _____ |

### *Answers for anagrams*

- |                  |                          |
|------------------|--------------------------|
| 1. Measures      | 6. Attributes            |
| 2. Star Schema   | 7. Conceptual Model      |
| 3. Normalization | 8. Slicing and dicing    |
| 4. Factless fact | 9. Physical modeling     |
| 5. Granularity   | 10. Degenerate dimension |



## *Challenge Me*

In the TenToTen Stores scenario, try to discover all the possible entities that can be identified towards helping the group in taking a good deci-

sion which would help the group in deciding on various aspects of business expansion, product promotion, and consumer preferences.



## *Let Me*

Explore the “What-If” analyses features in Excel 2010 and check if the future of business in the

Northwind Traders scenario can be predicted with the help of the entities present in their business.

## SOLVED EXERCISES

---

1. Is the OLTP database design optimal for a data warehouse?

**Solution:** No. The tables in an OLTP database are in a normalized state and therefore will incur additional time to execute queries and finally to return with results. Additionally, an OLTP database is smaller in size and does not contain historical/longer period (yesteryears) data, which needs to be analyzed. An OLTP system basically is based on the ER model and not on Dimensional Model. If a complex query is executed on an OLTP system, it may cause a heavy overhead on the OLTP server that might affect normal business processes.

2. If de-normalization improves data warehouse processes, why is the fact table in normal form?

**Solution:** Foreign keys of fact tables are primary keys of dimension tables. It is clear that a fact table contains columns which are primary keys to other tables that decide the normal form for the fact table.

3. What is real time data warehousing?

**Solution:** Data warehousing can also capture business activity data. Real time data warehousing captures business activity data as it occurs. As soon as the business activity is complete and there is data about it, the completed activity data flows into the data warehouse and becomes available instantly.

4. What is ODS?

**Solution:** The expansion of ODS reads Operational Data Store. An ODS has a database structure that serves it well as a repository for near real time operational data rather than long-term trend data. The ODS has the potential to become the enterprise-shared operational database, allowing operational systems that are being re-engineered to use the ODS as their operation databases.

5. What is “factless fact table”?

**Solution:** A fact table which does not contain numeric fact columns is called “factless fact table”.

6. What is a level of granularity of a fact table?

**Solution:** The level of granularity means the level of detail that you place into the fact table in a data warehouse. In other words, it implies the level of detail that one is willing to put for each transactional fact.

7. What are the different methods of loading dimension tables?

**Solution:** There are two ways to load data into the dimension tables. They are:

- **Conventional (slow) method:** All the constraints and keys are validated against the data before it is loaded. This way data integrity is maintained.
- **Direct (Fast) Method:** All the constraints and keys are disabled before data is loaded. Once data is loaded, it is validated against all the constraints and keys. If data is found invalid or dirty, it is not included in index and all future processes are skipped on this data.

8. What is surrogate key?

**Solution:** Surrogate key is a substitution for the natural primary key. It is simply a unique identifier or number for each row that can be used for the primary key to the table. The only requirement for a surrogate primary key is that it is unique for each row in the table. Surrogate keys are always integer or numeric.

**9.** What is data mining?

**Solution:** Data mining is the process of

- Analyzing data from different perspectives and summarizing it into useful information.
- Identifying patterns and trends and being able to predict the future.

**10.** How is an ODS different from an enterprise data warehouse?

**Solution:**

| <i>Operational Data Store (ODS)</i>                             | <i>Enterprise Data Warehouse</i>                                 |
|---|--|
| Purpose is operational monitoring.                              | Purpose is strategic decision support.                           |
| Contains current data only.                                     | Contains both current as well historical data.                   |
| Volatile data   | Static data  |
| Contains only detailed low-level or atomic or indivisible data. | Contains both detailed and summary data.                         |
| Satisfies the organization's need for up-to-the-second data.    | Satisfies the organization's need for archived, historical data. |

**11.** Why data marts are required?

**Solution:** Data marts facilitate querying and maintenance because the data mart usually contains data pertaining to a business process or a group of related business processes.

**12.** What are conformed dimensions?

**Solution:** A dimension that is shared between more than one fact table is called as conformed dimension.

**13.** What are the various data modeling tools available in the market?

**Solution:**

| <i>Vendor</i>       | <i>Product</i>  |
|---------------------|-----------------|
| Computer Associates | Erwin           |
| IBM Corporation     | Rational Rose   |
| Microsoft           | Visio           |
| Oracle Corporation  | Oracle Designer |

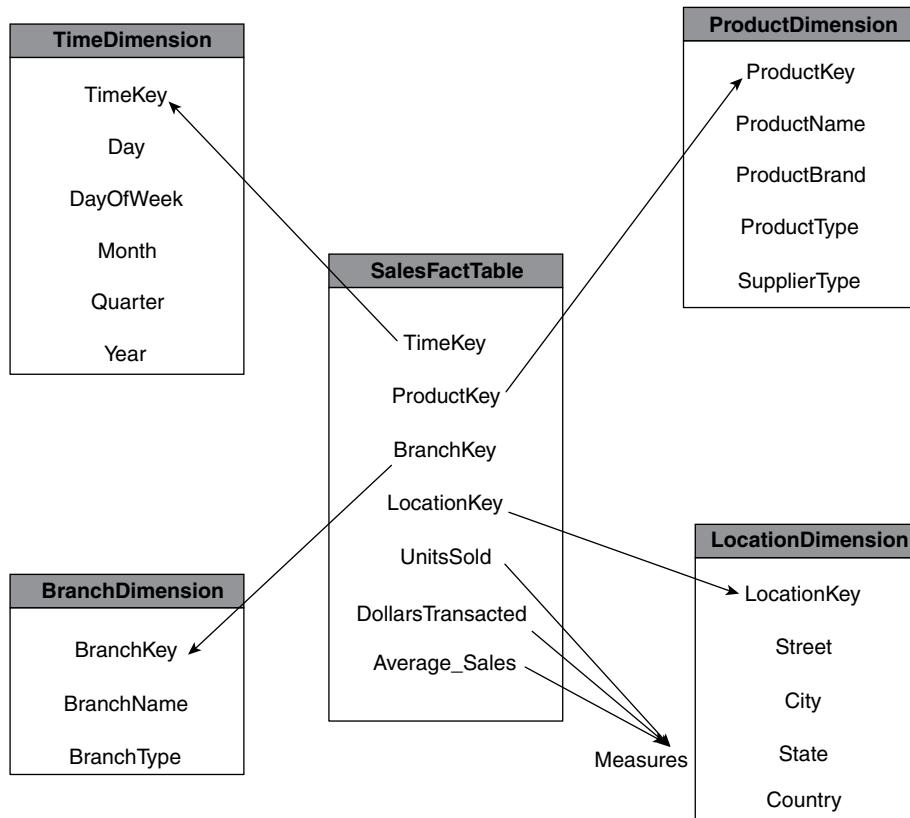
**14.** What are the differences between Star schema and Snowflake schema?

**Solution:**

| <i>Star Schema</i>   | <i>Snowflake Schema</i>  |
|--|--|
| Requires relatively more space on the database.  | Requires relatively less space on the database.                                    |
| Has relatively lesser number of dimension tables and therefore lesser number of joins. | Has relatively more number of dimension tables and therefore more number of joins. |
| Increases query performance  | Reduces query performance.   |
| Suitable for heavy end-user query workloads.   | Not suitable for heavy end-user query workloads.                                   |

15. Draw the star model for “HealthyYou”, a home healthcare product and services branch of “Good-Life HealthCare Group”.

**Solution:**



## UNSOLVED EXERCISES

1. What is a data model? Why is there a need for a data model?
2. What are the salient features of the Conceptual Model? Explain.
3. How is the Conceptual Model different from a Logical Model? Explain.
4. How is the Logical Model different from a Physical Model? Explain.
5. Why should you normalize your database? Explain giving example.
6. Assume you are working on an on-line inventory application. Would you like to go with normalizing your database or de-normalizing your database?
7. Assume you are the owner of a reporting application. Your application pulls data from an underlying database and generates a report. Would you recommend going with de-normalization to your database team?
8. Which situation will demand to go with dimensional modeling? Explain your answer with an example.

9. What constitutes a fact table? What are the various types of facts? Explain using examples.
10. What constitutes a dimension table? Explain.
11. What is your understanding of slowly changing dimension? Explain.
12. What is your understanding of the rapidly changing dimension? Explain.
13. Explain the role-changing dimension with examples.
14. Explain the junk/garbage dimension with examples.
15. Explain the various types of dimensional models: star, snowflake and fact constellation.
16. Explain the term “grain” with an example.
17. Explain the term “hierarchy” with an example.
18. What scenario will prompt you to use Star schema? Explain.
19. What scenario will prompt you to use Snowflake schema? Explain.
20. What scenario will prompt you to use Fact Constellation schema? Explain.
21. What is a “factless fact”? Explain with an example.
22. Consider this scenario. An employee of an enterprise has completed work on the project that he was working on. He has been assigned to a new project. Each project is identified using a unique project code and project name. At the completion of this new project, he will move into a new project. What we are trying to say here is that an employee does not always remain on the same project code and project name. How will you maintain his data? (Hint: Rapidly changing dimension.)
23. How does maintaining the data in a multidimensional format help with data analysis? Comment.
24. Compare and contrast the various types of slow changing dimensions. Use an example to better explain your answer.
25. What is the difference between data warehousing and business intelligence? Explain.
26. Draw the star model for GoodFood Restaurants Inc.
27. Draw the snowflake model for “HealthyYou”.

# 8



# Measures, Metrics, KPIs, and Performance Management

## BRIEF CONTENTS

|   |  |
|---|--|
| What's in Store                           | Where do Business Metrics and KPIs Come From?                  |
| Understanding Measures and Performance    | Connecting the Dots: Measures to Business Decisions and Beyond |
| Measurement System Terminology            | Summary  |
| Navigating a Business Enterprise,         | Unsolved Exercises   |
| Role of Metrics, and Metrics Supply Chain |  |
| "Fact-Based Decision Making" and KPIs     |  |
| KPI Usage in Companies                    |  |

## WHAT'S IN STORE

You are already familiar about the core purpose of Business Analytics, viz. supporting decision making with facts. These facts primarily are numeric data, measured and made available at the right time, in the right format to facilitate decision making. In this chapter you will learn all about these “numeric facts” that are like the “seed” which could grow into a big “tree”. We would like to reaffirm that the information presented here is very basic in nature and forms the foundation for deeper understanding of the subject.

In this chapter we will familiarize you with the terminology associated with measurement, need for a system of measurement, characteristics of measures, process used for defining good measures, relationship of these measures with individuals/teams/departments and the entire company. We will share several examples in the industry we have chosen, i.e. retail, hotel, and domains that are very familiar to all of us.

We suggest you refer to some of the learning resources suggested at the end of this chapter and also complete the “Test Me” exercises. We suggest you conduct self-research in your area of interest leveraging the information available on the Internet.

## 8.1 UNDERSTANDING MEASURES AND PERFORMANCE

---

### Picture this familiar scenario...

When you joined your high school, you were familiarized about the subjects you are required to study, lab experiments you need to conduct, tests you must write, assignments/projects you need to submit, how your final marks will be computed, how rank students will be rewarded, minimum score needed to move to next class, and so on. Why all these? Would you join a high school where there was no duration of the high school programs, no definition of the course curriculum, no idea of subjects to study, no pass criteria, or no clarity about when and how you would move to the college level?

Virtually everything around us has a system of measurement. You have already studied about basic measures such as length, mass, and time. You know how speed, velocity, and acceleration are measured. You know when you are running a high temperature (fever) or have gained weight; you know the purpose of the speedometer in a scooter/car. So, now we will attempt to help you understand “measures” relating to businesses.

In the example of your high school system, think about the following questions:

- How your report card is defined – Subjects, Minimum marks, Maximum marks, Your score in each subject?
- How top 10 students are declared in every grade?
- How your school compares last year’s results with current year to show progress to your parents? How does your school fare in comparison to other high schools in the neighborhood?
- How the high school “revises the curriculum” or introduces a new subject to improve the relevance of the curriculum with changing times?
- How does your high school plan for increasing the intake capacity by adding more sections and new teachers?

You will find a common thread running in the answers of the above questions. And, that is, “*performance*” which is all about achieving goals/results. To analyze performance and to declare achievements we need “measures” and systematic data/facts collection. This is the heart of this chapter.

## 8.2 MEASUREMENT SYSTEM TERMINOLOGY

---

There are some terms associated with the system of measurement that need to be defined before we look at how businesses use the system of measurement for analyzing business performance.

- **Data:** It is a collection of facts which have similar attributes or characteristics.  
“Phone number” is a named collection of, say, mobile phone numbers of your friends.  
“Email IDs” is an example of collection of email IDs of your classmates.  
Notice that all mobile phone numbers are of numeric type (10-digit long) and all email IDs will have the format of friend@mycollege.com.
- **Measure:** Data with associated unit of measure (UOM) is typically termed as measure.  
“Lab hours per month” has a numeric data associated with “time duration”.  
“Average wait time for bill payment” is a measure derived out of multiple data points.

- **Metric:** It is a system of measures based on standard UOM with a business context. The term business metric also refers to the same.  
“Product defect rate” by city is an example of measuring “what percentage of goods was returned by customers in different cities”.  
“Employee attrition rate” by quarter measures the percentage of employees leaving the company within each three-month period.
- **Indicator:** It is a business metric used to track business results or success/performance.  
“Call drop frequency” for mobile phone users is an indicator of user dissatisfaction.  
“Earnings per share” may indicate increased investor interest.  
“Cost per unit shipped” may indicate the rising cost of delivery of products.
- **Index:** It consists of a composite set of indicators used to address the overall health of business operations.  
“Customer satisfaction index” measured on a scale of 1 to 5.  
“Market performance index” measured using standard stock market scale.

A metric data when properly defined includes four components. Two of these are descriptive in nature, that is, they are qualitative. These components are: Subject and Stratum. The other two are quantitative components: Quantum and Application. Let's know more about these four components.

- **Subject:** This measure is about a customer, a product, a supplier, an employee, etc.
- **Quantum:** It is the value of the measure, such as cost, frequency, duration, amount, etc.
- **Stratum:** It is the grouping consideration expressed like By Location, By Quarter, By Customer, etc.
- **Application:** Value compared with similar measurements like previous month, forecast, target, etc.

Let us identify the above components in the following piece of information.

***“Cell phone” Cost in “Asia Pacific Region” is USD 100 against Target of USD 75.***

In the above information, Subject is product (cell phone), Quantum is actual cost (USD 100), Stratum is region (in our case Asia Pacific), and Application is comparison with target (USD 75). Notice that cost and target must have the same unit of measure and scale (not hundreds or thousands of USD) for comparison. You will come across several examples throughout this chapter that will help you understand these terms.

### **8.3 NAVIGATING A BUSINESS ENTERPRISE, ROLE OF METRICS, AND METRICS SUPPLY CHAIN**

#### **Picture this familiar scenario...**

You are a passenger on a flight from Mumbai to New York. The flight is about to depart and the in-flight announcements have just been completed. Think about the role of the flight captain and his team. Your thoughts may include:

- The goal of the flight captain is to reach the destination safely, on-time.
- He has carefully planned the route for the day, altitude, average speed, and traffic based on the weather, system maps, recommended critical changes, if any, etc.

- He has checked fuel sufficiency, technical report, food and beverages status, flight attendants, etc. to ensure that the aircraft is fully ready for the flight.
- He is aware of the cockpit instrumentation readings, target values, allowed deviation range, and emergency procedures if the need arises.
- He is aware of the impact of the decisions he might make during flight. He has alternative strategies in mind.
- He knows his team members' competencies, knows resources available to make every customer happy and uphold the values of the airline he represents.
- Many more...

Today, running a large business enterprise is like navigating an airliner or a ship safely to the pre-determined destination. Just like the captain relies on critical flight measurements like altitude, speed, cabin pressure, aircraft functions, etc. similarly businesses depend on measurements that help business leaders navigate the enterprise. You will recognize that all the work is not carried out just by the flight captain; it's the responsibility of a large team and the systems starting with ground staff to a series of control towers that guide each aircraft to its destination. Drawing a parallel with the business, business leaders too depend on key functions such as human resource, finance, sales, marketing, research, production, procurement, facilities, distribution teams, etc. to reach business goals. Each business function will need its own goals and measurements to ensure that the entire company achieves business success together. Every team within each function needs to achieve its goals, and finally every individual also needs to contribute his/her best for the business to reach its goals. This in business performance terms is called **alignment**.

Let us see what happens in a retail store in the following scenario. The marketing department has its independent goal to maximize sales volume. They launch a new "Discount sales campaign". At the same time the store operations department sets a goal to reduce sales support staff in each store. The purchase department has its own independent goal to reduce inventory cost, and hence reduces stock. What happens if the above-stated goal remains independent department-level goals? You are right! The outcome would be – When more customers come to store they don't find items that were on "Discount sale". There is no one to help on the shop floor! There is no stock in the store to service the customer demand! This happened because there was no alignment of goals of different departments.

A good measurement system, together with the system to plan and collaborate will make the scene very different. Purchase and operations are aware of the marketing campaign and projected sales and they have prepared for the "Discount sales campaign". Operations defer the staff reduction move. Rather, it increases the help available on the floor. Purchase ensures processes to fulfil customer orders smoothly. This is what is termed as **being on-the-same-page**.

The measurement system has a huge role in guiding the execution of business strategy. Exactly as the flight captain slightly alters the flight path to take care of local turbulence and yet he is aware of the destination/goal, business decision makers too test new approaches and reach business goals leveraging business metrics.

Let's now focus our attention to the question – What are the attributes of a good metric? The description given in Table 8.1 will enable you to answer this question.

**Table 8.1** Salient attributes of a good metric

| Metric Attribute  | Remarks   | Example   |
|---|---|---|
| <b>Name</b>   | Metric should be assigned a simple, easy-to-remember name. Do not include codes, long words, and unit of measure.   | 1. eLearning Training Days<br>2. Average Lines of Code  |
| <b>Abbreviation</b>   | Short form used inside the organization.  | eTD/ ALOC in above cases.   |
| <b>Description</b>  | Provide explanation to help users understand more contexts and comprehend the metric unambiguously.   | eLearning Days – Total number of full-time days equivalent spent in training using online course delivery system. Users may log-in any number of times and duration of each session is captured in minutes.       |
| <b>Unit of Measure (for data capture)</b>   | The commonly measured base unit needs to be included.   | In the eLearning example, the unit is “Minutes”.  |
| <b>Scale</b>  | Commonly reported granularity of unit of measure. We need to capture the conversion formula. Simple multiples like 1000 (K) or M (Million) are commonly used.   | In the eLearning example, as the data storage granularity is “Days”, the scale is “minutes/(60 * 8)” assuming 8 hours is a standard training day.   |
| <b>Metric Owner</b>   | Position/department responsible and accountable for the metric.   | The training support manager in the training department could be an owner.  |
| <b>Frequency</b>  | Indicates how often this metric will be measured.   | In the eLearning example, it could be “Month”, i.e. every month the metric is consolidated. It’s useful to indicate when the metric will be available for use, e.g. available after 2 working days in each month. |
| <b>Priority</b>   | Department/organization priority in terms of value of this metric in the current organizational context.  | High/medium/low could be used.  |
| <b>Data Values</b><br><b>Target</b><br><b>Actual</b><br><b>(Computed will include Minimum, Maximum, and Average, Valid range)</b> | These are some very important attributes to be defined. <i>Actual</i> represents current period/cycle captured value. <i>Target</i> represents the intended/planned value. The lowest and highest possible values indicate the range. | In the eLearning example, it may be the mandate of the organization to ensure 2 days equivalent for each employee as <i>target</i> . <i>Actual</i> is the value captured for each employee.                       |
| <b>Target Description</b>   | What is the definition of normal achievement? What constitutes high performance? What is the escalation threshold? What is the overall rationale for arriving at average/normal performance?  | In our example, it’s good to encourage employees to keep themselves current and hence continuous learning is critical. 2 days out of 22 working days (10%) could be a good average.                               |

(Continued)

**Table 8.1** (Continued)

| <i>Metric Attribute</i>                 | <i>Remarks</i>  | <i>Example</i>   |
|---|---|--|
| <b>Formula</b>                          | How to compute the metric and exceptions?<br>This again makes the definition unambiguous for ALL stakeholders.  | In this example, organization may indicate terms like: <ul style="list-style-type: none"><li>• For permanent staff</li><li>• Month = 22 working day</li><li>• Deduct leave days</li><li>• Include holidays</li></ul> |
| <b>Weight</b>                           | When we define a group of 6–8 strategic metrics, sometimes it may be useful to declare the relative importance of metrics in that group. Weight of the metrics need to add up to 100.   | Consider training weight:<br>Instructor-led Training – 30, eLearning – 40, Team project – 30<br><b>Total Learning weight 100</b>   |
| <b>Measurement Perspective</b>          | It may be useful to indicate the organization's strategic perspective that is impacted by this metric.  | In this case, customer and finance perspectives of the balanced scorecard are impacted by talent development.<br>Balanced scorecard has been discussed in detail in Chapter 9.                                       |
| <b>Trusted Data Source</b>              | The IT application that has the trusted actual values.  | In our case it could be LMS (Learning Management System).  |
| <b>Revision History</b>                 | List of all changes made to metric definition with date/person ID stamps from the data of approval for implementation.  |  |
| <b>Date of Approval</b>                 | First implementation roll-out date.   |  |
| <b>Organization-specific indicators</b> | Organizations attempt to record metric type (lead or lag) data quality indicator, internal metrics classification category, typical applications that use the metric, initiatives that currently depend on this metric, expected direction of trend, etc. |  |

Finally, let's look at how can we say that the defined metric is a good metric? This is really a complex question. While we can suggest test for structural correctness, it may not turn out to be a metric of business importance or vice versa.

Experts suggest the following **SMART** test for ensuring metric relevance to business. Refer to Table 8.2. Some of the readers may be familiar with this SMART test that is as applied for goals/objectives as well.

Refer to Table 8.3. We are now ready to understand the “supply chain” associated with the metric, i.e. where does measurement start, how the measures get transformed into metric, how do they get distributed, how users leverage them, how users achieve business results? This entire process could be compared to how a product (say, smartphone) you purchased creates personal productivity value starting with raw materials used to manufacture phone, its assemblies, product, delivery channels, use of phone features to enhance personal productivity and measuring the product gain (value).

**Table 8.2** SMART test for ensuring metric relevance to business

| <i>Test</i>     | <i>Test Focus</i>  |
|-----------------|--|
| Specific        | Metric is clearly defined, articulated, and understood by all stakeholders, and is triggering action.  |
| Measurable      | Someone in the organization must have the ability/instrumentation to accurately, easily, and regularly measure the actual value at reasonable cost and technology. Think if a clinical thermometer would cost USD 1000!!                                 |
| Attainable      | There will be no metric without target. This target may be stretched but must be attainable with the current level of people efforts and processes. Speed by cycle can't be enhanced to 300 kmph no matter whatever be the technology used!              |
| Result-oriented | The metric must motivate team members performing the work. In businesses, results are crucial.   |
| Time-bound      | All actual values of metrics should be traceable to the date/time when the actual value measurement was taken. The instrument used for measurement also has a key role in sampling, accuracy, speed, and correctness that can be verified in other ways. |

**Table 8.3** Supply chain associated with the metric

| <i>Component of Measurement</i>                    | <i>Supply Chain Contribution</i>  |
|--|---|
| <b>Entities to be measured</b>                     | Includes employee, vendor, product, customer, asset, expense category, sales promotion, service feedback, ... |
| <b>Instrumentation</b>                             | Measurement data, data capture, and storage in raw form   |
| <b>Raw material</b>                                | Reference data, definitions, benchmarks, limits, ...  |
| <b>Sub-assemblies</b>                              | Measures with unit, format, storage structure, archives, ...  |
| <b>Product</b>                                     | Business metrics approved, communicated and measured, verified and analyzed with rigor                        |
| <b>Metrics Delivery</b>                            | Reports, dashboards, scoreboards, alerts, Flash updates   |
| <b>Business Activity Areas (Decisions/Actions)</b> | Plan review, tracking project progress, sales campaign analysis, profit forecast                              |
| <b>Business Application</b>                        | Budget control, quality improvement, innovation projects  |
| <b>Business Value</b>                              | Business results meeting and exceeding plan   |

As an example to illustrate the above supply chain concept, think of the electrical energy meter or water meter installed in your home. The entity to be measured is energy consumption or water consumption and the associated instrument is energy/water meter. In this system of measurement, the utility company keeps record of meter reading and takes meter reading every month. The sub-assembly is the amount of energy consumed, added tax, adjustments of advances, calculation of late fee if any, and arriving at the net bill amount. The delivery to consumer is total energy/water consumed and net payable utility bill amount. The same data for energy/water consumption analysis will reveal different patterns of energy/water consumption in locations/state/country. They will further look at energy/water requirements by household, industry, agriculture, and environment. The same data when goes to marketing function helps identify consumers for new projects like solar water heater/rain water harvesting to support energy/water conservation. These are applications of metrics. They help business managers make decisions about target conservation they must achieve to escape increased taxation for industries consuming very large amounts of resources or promote green innovations. The utility company achieves goals of meeting energy/water requirements, promote conservation of resources, and serve consumers with accurate/timely billing and collection channels. That is the real power of each metric!

Now that you have some idea about the “big picture” of business enterprise operations, we will further explore how metrics helps managers make business decisions.

## 8.4 “FACT-BASED DECISION MAKING” AND KPIs

---

We all decide all the time about the actions we want to take. Some of our actions are very intuitive, and we don't analyze thoroughly before acting. Some decisions raise questions in our minds – Think about the new vehicle you want to buy or a new house you are considering to buy! This is the place where measures and quantitative values will help us compare alternatives. When we use “facts” we are not driven by emotions or gut feel. The true picture emerges when we start comparing similar quantities or qualitative ratings. It's now very evident that metrics help us evaluate alternatives more objectively and help us in decision making. Just think of the analogy of the flight captain looking at the cockpit dials and weighing options in real time to choose the best alternative procedure to address a particular challenge. Similarly, team members engaged in business activities use the “facts” to objectively compare and choose the “best option” for a given business situation. This is the heart of decision making at individual levels and is termed as “fact-based decision making”.

Now let's understand the need for “fact-based systems” for decision making in businesses. All businesses develop a high-level vision, objectives for each function/initiative, key performance areas/goals, key performance indicators, targets, and allowed variance for targets to steer the business towards its goals. This is the performance management process of enterprises.

Essentially, **performance measurement** is about analyzing the success of a project team, department, or country-wide business or global business's efforts by comparing data on what actually happened to what was planned or intended. Performance measurement is about asking: Is progress being made towards the desired goals? Are appropriate decisions and actions being undertaken to ensure achieving those goals? Are there any challenges that need attention?

Performance measurement is the selection and use of quantitative measures specific to team initiatives, department, group, or entire company to develop information about critical aspects of business activities, including their results/outcomes. Performance measurement involves regular collection, tracking, and reporting of data relating to effort invested as well as work produced and results achieved.

If the measures have such a huge impact on the business results, it's needless to say that the measures are required to be chosen very carefully. Let's look at some characteristics of good business metrics. We will identify those metrics that are important at the entire company level and call them as Key Performance Indicators (KPIs). Remember that "*not everything that can be counted counts and not everything that counts can be counted*".

- **Relevance and functionality:** The KPIs chosen should be directly related to business results that the company is trying to produce in the specific business function. Like, your body temperature measurement can only indicate whether you have fever or not, but can say nothing about your blood pressure! You may even remember popular Murphy's Law – "You cannot determine the direction in which the train went by looking at the tracks".
- **Understandable:** Chosen KPIs must be defined unambiguously. A KPI needs to be understood in one and only one way by all stakeholders. It must be documented, and its definition must be easily accessible to all users.
- **Reliability and Credibility:** The value of KPIs needs to be authentic and should be validated as "trusted" or "dependable". Someone is going to base an important decision on the chosen metric. Adequate checks are needed to declare data as trustworthy. This also means that the data must represent the "single version of truth".
- **Abuse-proof:** An abuse-proof measure is unlikely to be used against intended purpose or individual(s) involved in the measurement process.

Performance management could be defined as the use of performance measurement facts to help set agreed-upon business performance goals, define objectives for growth and innovation, or excel, allocate, and prioritize resources, inform managers of their targets, and report on the success in meeting those goals.

In other words, we may say that KPIs are objective, measurable attributes of business performance, which assist in informed decision making. They are a means of assessing the business functions' health and a means of assisting in the prediction of business success and potential failure. They can also be a means of capturing best practices and lessons learned at individual, team, department, and company levels.

Key performance indicators are quantitative or qualitative measures which reflect the business performance of a company in achieving its goals and strategies. KPIs reflect strategic value drivers rather than just measuring non-critical business activities and processes. They align all levels of an organization clearly defined and cascaded down to individual level targets to create accountability and track business progress. KPIs are designed specifically for each organization and may use industry standard values for benchmarking (comparing with world-class standards).

Let us look at a few sample KPIs used by the Human Capital and Training Management division of "GoodFood Restaurants Inc.":

- Average time to recruit.
- Average open time of job positions.
- # of responses to open job positions.
- # of interviews to fill up open job positions.
- # of offers that were made.
- # of responses to the offers made.
- % of vacancies that were filled within  $x$  time.
- % of new employees that remained after  $x$  time.
- % of new employee satisfaction rate.

Few sample KPIs employed by the Employee Training Management Division of “GoodFood Restaurant” are as follows:

- % of employees who underwent training.
- Average training cost per employee.
- % of employees satisfied with training.
- Average training hours per employee.
- Ratio of internal vs. external training.
- % of budget spent on employees training.
- ROI of training.

Likewise, let us look at a few sample KPIs likely in use by the Help Desk of “GoodFood Restaurant”:

- Average no. of calls by customers in a day.
- Average time spent by a help desk employee in attending to calls.
- % of complaints serviced in a day.
- % of customers satisfied by the services offered.
- % of complaints serviced well before the SLA (service-level agreement) time.

KPIs help change the way business team members do their jobs, approach their daily planning/schedule, and deal with daily urgencies/escalations. KPIs help people focus on the “big picture”. They help people distinguish the important from the trivial, the “must be done” from the “could be done”, and allow employees to set their own priorities. KPIs are best employed at the lowest level for an individual.

KPIs differ from industry to industry as indicated below:

- For the retail industry, the average dollars per sale is a good KPI.
- For project managers, employee attrition is an important KPI.
- Inventory accuracy is an important KPI for distribution companies.
- Inventory turns is a very important KPI for manufacturing and distribution companies.
- For telemarketers, the number of phone calls made is an important KPI.
- For accounts payable departments, the number of AP Days outstanding is important.

## 8.5 KPI USAGE IN COMPANIES

---

KPIs could be used in the company at strategic, tactical, and operational levels. We will not be detailing the techniques used for defining KPIs as this is clearly an advanced topic. At this stage it is sufficient to know that there are techniques like Cause and Effect Modeling, Goal Question Metric and Measure (GQMM), Balanced Scorecard (BSC), Six Sigma, MBNQA, TQM, Economic Value Add (EVA) Management Frameworks to represent KPIs to define, align, track, and communicate strategy. The following points highlight how KPIs are used by companies.

- KPIs often times are built as KPI tree. A lower order metric aggregates actual value into a higher level and the structure looks like a tree structure.
- KPIs are typically classified into lag indicators that reflect the result achieved (net profit) or lead indicators that project the possibility of achieving the target (large projects using patented ideas). While companies look at both the effort and results picture, the result/outcome KPIs are tracked with rigor.

- KPIs reflect the value drivers or the control levers that directly and significantly influence the outcome that is being tracked.
- KPIs are cascaded down from corporate level to regional, functional, team, and individual levels. This helps in aligning the efforts of the company towards achievement of strategic goals.
- KPIs are also targeted for specific roles. This technique avoids information overload and helps team members focus on what they need to achieve to meet the overall goals of the company.
- KPIs are employed to track several non-financial goals of large global companies in areas including strategy quality of the company, execution capability, research and thought leadership, talent management ability, innovation, quality of processes, top management effectiveness, and so on.
- Standard KPIs may be purchased from third-party consulting firms, and it's even possible to buy industry benchmarks that help business leaders compare their competitive position in the global marketplace.

## 8.6 WHERE DO BUSINESS METRICS AND KPIs COME FROM?

It's important to know how companies evolve business metrics and KPIs and the IT technology that is critical to deliver the agreed metrics to the right user, at the right time, and in the right format on the preferred device of the user.

There are several sources of inputs that companies use to evolve KPIs at corporate, regional, functional/department, strategic, team, and individual levels. Some of the inputs for metrics are as depicted in Table 8.4.

**Table 8.4** Potential sources for metrics

| Source   | Inputs for Metrics  |
|--|---|
| <b>Corporate Vision, Mission, Values</b>                     | External stakeholder KPIs, metrics that are tracked by business analysts/market research firms, etc.  |
| <b>Business Performance Projections/ Forecasts/ Guidance</b> | Typically financial performance, profit, industry-specific benchmark metrics, and so on.  |
| <b>Business Plan (Sales)</b>                                 | Revenue – Customer, product, service line revenues<br>Competitive position and advantages<br>Talent and partner plans<br>Regional and channel plans<br>Strategic investments, innovations, research |
| <b>Production/Manufacturing</b>                              | Inventory, purchase, strategic sourcing projects<br>Technology adoption, mass production, quality improvement-based cost savings  |
| <b>Operations &amp; Service</b>                              | Customer satisfaction, field service expenses<br>Service quality inputs/benchmarks<br>Support costs, replacements   |
| <b>Society and CSR</b>                                       | Care for environment  |

(Continued)

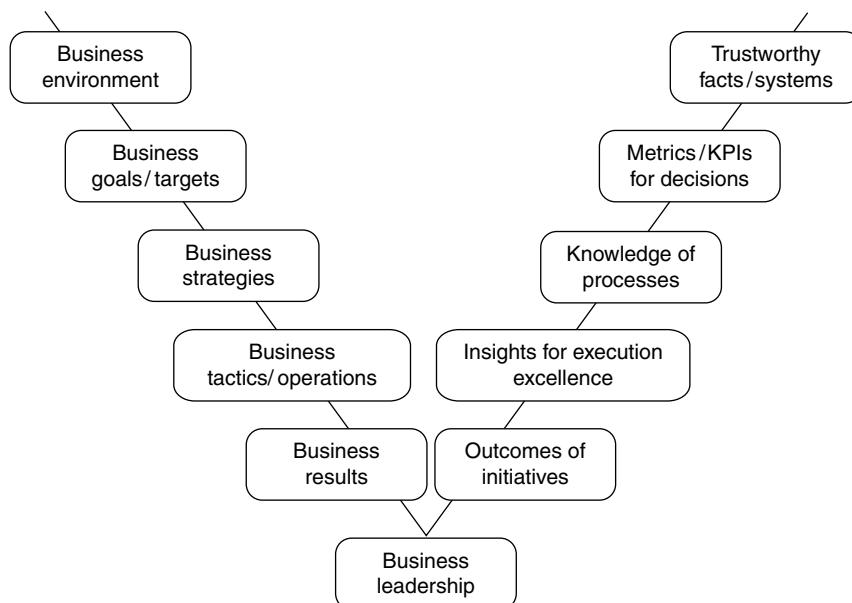
**Table 8.4** (Continued)

| Source   | Inputs for Metrics                   |
|--|--------------------------------------|
| Function-specific such as Human capital & training, Finance, Procurement, Technology Infrastructure, Facilities & Maintenance, Marketing, Quality, R & D | Function-specific                    |
| Talent Performance Management  | Strategic talent development metrics |

Companies look at past performance, new requirements, and competitive data to track KPIs in a scorecard just like the student report card. When it comes to senior leadership levels, the number of KPIs will be just around 20. It cascades down to various functions, business lines, and support organizations. At individual levels, the KPIs will be a combination of applicable corporate KPIs, function-level KPIs, strategic innovation KPIs, and individual contribution KPIs.

## 8.7 CONNECTING THE DOTS: MEASURES TO BUSINESS DECISIONS AND BEYOND

You are now ready to get an integrated picture of different concepts we have outlined about metrics. Measure is a part of our life. It has a powerful influence on the decisions we make, if used properly. There exist scientific methods and frameworks that have been researched for years for harnessing the power of measures. There is a need to transform simple facts into key performance indicators by adding more “contexts” to basic facts. Look at Figure 8.1 to understand the relationship that exists between the business world and the technical world centered on facts.

**Figure 8.1** Mapping metrics to business phases.

Global businesses rely on technology-enabled initiatives to achieve market leadership position. Such strategic initiatives may be focused on innovation, targeted customer marketing, process optimization, risk management, global resource planning, global procurement, productivity enhancement, new opportunities identification, corporate performance management, and so on. These strategies need to be executed flawlessly or with “zero defect”. This only means that every decision has to be accurate and must contribute to reaching/exceeding the pre-set goals/targets. There is no room for mistakes. Such accurate decisions only need to be objective, i.e. without emotions, and collaboratively supported by leadership team. Only process metrics/KPIs can provide the reality status on the ground accurately in a consistent and repeatable fashion. This data will be used to alter/re-align or fine tune processes. Hence, decision makers need trustworthy KPIs/metrics. This can only come from carefully thought-out, pre-designed, and technology-enabled systems of facts. This is the true power of metrics.

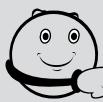
## SUMMARY

- Businesses depend on metrics for objective decision making.
- Business metrics are goal/result oriented.
- Metrics are not isolated but typically connect as KPI tree.
- Metrics motivate performance at individual, team, function, and company levels.
- There exists a scientific process to define, transform, communicate, and apply metrics for business value.
- Metrics have four distinct components, viz., Subject, Stratum, Quantum, and Application.
- Decision quality depends on the quality of actual metric values. Hence IT is leveraged to support consistent, accurate, speedy collection, tracking, consolidation, and reporting of metrics.
- Not everything that can be counted counts, and not everything that counts can be counted.
- Organization culture, standardization, education, and leadership are needed to implement effective fact-based decision culture.



### Remind Me

- There is nothing special about business metrics. They are just as natural as any other entity we measure.
- Navigating business is similar to navigating airplane. Both the aircraft captain and the business leader need accurate data at the right time in the right form.
- Every metric needs to be designed carefully starting with instruments to value it can provide. (Energy meter to discount plan for solar water heater users.)
- Many times a single metric will not tell the complete story. The airplane cockpit doesn't have just one meter but several for different measurements. In business, this takes the form of KPI tree.
- KPIs reflect the entire performance of the company and will be cascaded down to the level of individuals. This is called alignment.
- In an ideal situation every individual needs to know only the KPIs that he/she is responsible for and some KPIs that represent collaborative outcomes.



### Point Me (Books)

- *Corporate Performance Management*, Wade & Recardo, Butterworth-Heinemann.
  - *Six Sigma Business Scorecard*, Gupta, McGraw-Hill.
  - *Measuring Business Performance*, Neely, The Economist Books.
  - *Decision Making*, Rowe, Harvard Business School Press.
  - *Measure Up! How to Measure Corporate Performance*, Lynch & Cross Blackwell Business.
  - *The Balanced Scorecard*, Kaplan & Norton, Harvard Business School Press.
- (Please note that the content in the books are for advanced learning.)



### Connect Me (Internet Resources)

- [www.bettermanagement.com](http://www.bettermanagement.com)
- Google for “Performance measures” Retail/Manufacturing/Supply Chain.
- Download Microsoft Dynamics KPI Guide.

## UNSOLVED EXERCISES

1. Define KPI. Comment on the need for KPIs.
2. How is KPI different from Critical Success Factors? Explain.
3. A company is looking for recruiting graduates for their new business unit. They have asked the HR (Human Resource) team to go ahead with recruitment. What KPIs will be most relevant for measuring the effectiveness of the recruiting process?
4. Define KPIs to measure the effectiveness of a learning program.
5. “You cannot manage what you cannot measure, and you cannot measure what you cannot define”. Comment giving an example.
6. Search and prepare KPIs for measuring the effectiveness of a TV channel.
7. Determine 10 KPIs that will help decide the candidate for “Best Driver” award in the state transport company.
8. What KPIs are used in a cricket academy to select batsmen, bowlers, and fielders?
9. How would you aggregate KPIs for selecting sports teams for tournaments at college, state, regional, national, and international levels?
10. What KPIs will need to cascade from the national level to the individual outlet level for a country-wide retail chain such as Cafe Coffee Day to ensure same consumer experience?

11. What is “fact-based decision making”? Explain your answer with an example.
12. Explain the four components of metric data: Subject, Stratum, Quantum, and Application. Use an example to explain your answer.
13. Explain few attributes of a good metric.
14. Explain GQMM with the help of an example.
15. How does “measures” help with good decision making? Explain.
16. You are the owner of a retail chain. You wish to enhance the productivity of your store’s employees. What metrics will you define to achieve this objective?
17. You are the owner of a private airline. Your business has been incurring losses. To combat the same, you have decided to cut cost. What metrics will you define to achieve this objective?
18. A supply chain is “Manufacturer → Whole seller → Retailers → Consumers”. What metrics can be defined for this supply chain?
19. “You cannot determine the direction in which the train went by looking at the tracks”. Comment giving an example.
20. “KPIs reflect the entire performance of the company and will be cascaded down to the level of individuals”. Explain giving an example.



# 9



# Basics of Enterprise Reporting

## BRIEF CONTENTS

|   |                               |
|---|-------------------------------|
| What's in Store   | Balanced Scorecard            |
| Reporting Perspectives Common to All Levels of Enterprise | Dashboards                    |
| Report Standardization and Presentation Practices         | How Do You Create Dashboards? |
| Enterprise Reporting Characteristics in OLAP World        | Scorecards vs. Dashboards     |
|   | The Buzz Behind Analysis...   |
|   | Unsolved Exercises            |

## WHAT'S IN STORE

You are now familiar with the key concepts relating to business metrics and KPIs, multidimensional modeling, and the big picture of a business enterprise. With this background, it's time to think about the basics of enterprise reporting and gain hands-on experience in using a simple reporting tool.

Reporting is an integral part of OLTP applications. We will summarize some of the best practices from the OLTP world that have influenced the evolution of standard reporting practices. Our focus is on OLAP-centric reporting. This class of enterprise reporting will objectively communicate facts relating to strategy, corporate/department performance against plan, status of critical initiatives, and metrics that matter to the stakeholders. These reports help leaders align their business activities to the vision and strategies of their enterprise and to monitor their performance against the organization's goals. We will introduce you to some analysis types, and we have a special section on the "Balanced Scorecard" as well.

This chapter is a “Must Read” for first-time learners interested to learn about the basics of corporate performance management, the key performance indicators, the balanced scorecard, and the enterprise dashboard. In this chapter we will also familiarize you with the difference between the balanced scorecard and the enterprise dashboard.

We suggest you refer to some of the learning resources suggested at the end of this chapter and also complete the “Test Me” exercises. You will get deeper knowledge by interacting with people who have shared their project experiences in blogs. We suggest you make your own notes/bookmarks while reading through the chapter.

## 9.1 REPORTING PERSPECTIVES COMMON TO ALL LEVELS OF ENTERPRISE

---

We have already introduced you to the organization of a large enterprise in Chapter 1, “Business View of Information Technology Applications”. Typically enterprises have headquarters and several regional centers. Several geographic location-focused operations may aggregate to the nearest regional center. Each geographic location may have “revenue generating – customer facing units” and “support units”. There could be regional or corporate level support functions as well. IT could be leveraged at the local operations level or the regional level or the entire corporate level. Hence, it is natural to expect IT-enabled reporting to occur at local, regional, or corporate levels. Let’s understand some common perspectives of reporting that apply at all levels of the enterprise.

- **Function level:** Reports being generated at the function level may be consumed by users within a department or geographic location or region or by decision makers at the corporate level. One needs to keep in mind the target audience for the reports. The requirements will vary based on the target audience. Departments such as HR, marketing, production, purchase, accounting, etc. will need specific standard reports to handle operational, tactical, and strategic challenges. Reports could be produced in many languages to meet global user needs.
- **Internal/external:** Sometimes the consumers of reports may be external to the enterprise. We are very familiar with the annual reports of organizations. Correctness as well as attractive presentation of the report is of paramount importance.
- **Role-based:** Today we are witnessing massive information overload. The trend is to provide standard format of report to similar roles across the enterprise, as they are likely to make similar decisions. For example, a sales executive responsible for strategic accounts will need similar information/facts for decision making irrespective of the country/products he/she handles.
- **Strategic/operational:** Reports could also be classified based on the nature of the purpose they serve. Strategic reports inform the alignment with the goals, whereas operational reports present transaction facts. The quarterly revenue report indicates variance with regard to meeting targets, whereas the daily cash flow summary indicates summary of day’s business transactions. When consolidated across several locations, regions, products/services, even this report will be of strategic importance.
- **Summary/detail:** As the name suggests, summary reports do not provide transaction-level information, whereas detailed reports list atomic facts. Even here several summaries could be aggregated to track enterprise-level performance.
- **Standard/ad hoc:** Departments tend to generate periodic reports, say, weekly, monthly, or quarterly reports in standard formats. Executives many times need ad hoc or on-demand reports for critical business decision making.

- **Purpose:** Enterprises classify reports as statutory that focus on business transparency and need to be shared with regulatory bodies. For example, a bank reporting to the Reserve Bank stipulated parameters of its operations. You might have even heard of audit reports that are produced to check the correctness and consistent application of business policies across global transactions. Analytical reports look into a particular area of operation like sales, production, and procurement, and they find patterns in historical data. These reports typically represent large data interpretations in the form of graphs. Scorecards are used in modern enterprises to objectively capture the key performances against set targets and deviation with reasons. These scorecards help kick off many initiatives that bring back business parameters under control.
- **Technology platform-centric:** Reporting in today's context need not use paper at all. Dashboards could be delivered on smartphones and tablets. Reports could be published in un-editable (secure) form with watermarks. Reports could be protected to be used by a specific person, during specific hours from specific device! Reports could be delivered to the target user in user-preferred formats such as worksheet, word document, PowerPoint Presentation, text file or HTML document, and so on. Reports could be just generated once and shared with many users through an email link as well. Security of data is a constant concern in large enterprises as reports represent the secret recipe of the business. Several tools have emerged in the marketplace to meet the reporting requirements of the enterprise. It is not at all uncommon for enterprises to use several tools and technologies to meet the reporting requirements of the enterprise. Some have even set up Reporting Centers of Excellence to handle this crucial function.

## 9.2 REPORT STANDARDIZATION AND PRESENTATION PRACTICES

---

Now let's turn our attention to some of the common best practices that most enterprises employ while considering reporting requirements. Enterprises tend to standardize reporting from several perspectives. Some report standardization perspectives are as follows:

- **Data standardization:** This standardization perspective enables enterprise users performing same role to receive common, pre-determined data sets that relate directly to their role. The data provided will also include facts needed for active collaboration with other functions within/ outside the enterprise.
- **Content standardization:** Enterprises focus on content standardization. This is tightly tied to the name of the report. For example, the shipping report from despatch will have information that helps users connected with shipping to make informed decisions.
- **Presentation standardization:** The third perspective of standardization is about report presentation. Here enterprises set standards on naming conventions, date formats, color/ black–white usability standards, use of logos, fonts, page formats, security classifications, cover pages, and so on.
- **Metrics standardization:** The next major focus of standardization will be typically on metrics. Enterprises' functions strive to find the metrics that best reflect the status of performance to help teams control the progress towards their goals. External benchmarking and purchasing threshold values for industry key metrics are common.
- **Reporting tools' standardization:** Another key perspective is about reporting tools. Enterprises deploy specific class of reporting tools for different requirements of departments/locations/ audience.

**Table 9.1** Features of good reporting

| Feature                       | Description  |
|-------------------------------|--|
| Report title                  | It is important to provide a crisp name for the report such that it reflects its purpose. Some teams may even add the target audience. Example:<br><b>Cash flow report for SOUTH Region</b><br><b>Product Shipping Report for Plant 2</b>  |
| Reporting period              | As the reports use data collected over a specific period, it is critical to state the same. The period format could be:<br>For week beginning March DD, YYYY<br>From DD/MM/YYYY to DD/MM/YYYY  |
| Header/footer                 | It is good to include report headers and footers that repeat on every page. The content could have elements like logo, function name, page number, confidentiality level, etc.   |
| Column headings               | The numeric data presented will need to be read based on the column names. Again keeping crisp but meaningful names is critical. These are different from RDBMS column names and need to be user-friendly.   |
| Column selection and sequence | Although reports are meant to be used by users in the same role, but across different locations, users have their own preferences when they want to see the information being presented. There needs to be flexibility for users to select the columns that they would like to see as well as the order or sequence from left to right. Example – Microsoft Explorer and Microsoft Outlook allow one to select from a list of available columns and also display it in a preferred sequence. |
| Filters                       | Users may not be interested to see all the lines simultaneously. They need to have flexibility to use standard filters/custom filters (user-defined) to view lines that meet specific criteria. Example:<br><b>Cash Flow for Department = “XXX”</b><br><b>Cash Flow for Amount &gt; 100000.00</b>  |
| Sort sequence                 | Users would like to view reports lines arranged in increasing or decreasing order for convenience of decision making. Example:<br><b>Names to be arranged in alphabetical order</b><br><b>Revenue Report to be in decreasing order of amount</b><br>It may be needed to sort lines in cascading fashion as well.<br>BY Department + BY Employee Name<br>BY Customer + BY Date + BY Invoice number  |
| Totals/group totals           | When data lines are sorted, they may need to be grouped or bunched together in chunks that make business sense. In these situations, users expect totals and cross-totals as well as overall totals to be computed/provided.   |
| Data field formatting         | Users also expect the stored data to be represented with formatting to enhance reading convenience.<br>Using currency symbols like \$, etc.<br>Date formatting like June 3, 2011   |
| Computed or calculated fields | Users may want to introduce new columns that are derived from existing columns and compute some value for decision making.   |

(Continued)

**Table 9.1** (Continued)

| Feature               | Description   |
|-----------------------|---|
| Highlighting breaches | Reports may highlight using color or font size/style to make the field seize the attention of the user.                           |
| Notes                 | Sometimes it may be essential to notify users of last-minute update messages that could answer typical questions raised by users. |

Report content and presentation related approaches and styles have evolved in synchronization with advancements in technologies relating to display and printing. Some features of the good reporting drawn from the OLTP world of reporting are given in Table 9.1.

Let us now look at a few common report layout types.

### 9.2.1 Common Report Layout Types

- **Tabular reports:** Tabular reports have a finite number of columns, typically representing the fields in a database. A tabular report has header and footer, and repeating detail rows. Data can be grouped on various fields. Each group can have its own header, footer, breaks, and subtotal. Table reports are typically used for logging detailed transactions. The tabular report depicted in Figure 9.1 is grouped on “Category” and “SubCategory”. It displays details of all products under each “SubCategory” and “Category”.
- **Matrix reports:** As discussed earlier, business reporting is about summarizing information for analysis. A matrix, cross-tab, or pivot report aggregates data along the *x*-axis and *y*-axis of a grid to form a summarized table. Unlike tabular report columns, matrix report columns are not static but are based on group values. Figure 9.2 depicts a matrix report that displays the total sales of products under each SubCategory and Category. The row grouping is on Category and SubCategory and the column grouping is on Month (Jan, Feb, Mar).

| Category    | SubCategory | Product                 | Order Quantity | Sales Account | Cumulative Total | Margin   | Margin % |
|-------------|-------------|-------------------------|----------------|---------------|------------------|----------|----------|
| Accessories |             |                         |                |               |                  |          |          |
| Bikes       | Helmets     |                         |                |               |                  |          |          |
|             |             | Sport-100 Helmet, Black | 81             | \$1,622.35    | 1622.3484        | 498.2142 | 30.71 %  |
|             |             | Sport-100 Helmet, Blue  | 69             | \$1,392.87    | 3015.2169        | 435.2727 | 31.25 %  |
|             | Locks       | Sport-100 Helmet, Red   | 45             | \$908.39      | 3923.6094        | 283.8735 | 31.25 %  |
|             |             | Cable Lock              | 50             | \$750.00      | 750.0000         | 234.3750 | 31.25 %  |
|             | Pumps       |                         |                |               |                  |          |          |
|             |             | Minipump                | 55             | \$659.67      | 659.6700         | 206.1455 | 31.25 %  |

**Figure 9.1** A tabular report.

| Category    | SubCategory       | Jan       | Feb       | Mar       | Total      |
|-------------|-------------------|-----------|-----------|-----------|------------|
| Accessories | Helmets           | 3371.1455 | 4360.2840 | 3923.6094 | 11655.0389 |
|             | Pumps             | 635.6820  | 467.7660  | 659.6700  | 1763.1180  |
|             | Locks             | 720.0000  | 735.0000  | 750.0000  | 2205.0000  |
|             | SubCategory Total | 4726.8275 | 5563.0500 | 5333.2794 | 15623.1569 |

**Figure 9.2** A matrix report.

- **List reports:** A list report has a single, rectangular detail area that repeats for every record or group value in the underlying data set. Its main purpose is to contain other related data regions and report items and to repeat them for a group of values. The list report, depicted in Figure 9.3, repeats

Clothing

| Category       | SubCategory       | Jan        | Feb        | Mar        | Total       |
|----------------|-------------------|------------|------------|------------|-------------|
| Clothing       | Jerseys           | 4787.5064  | 7498.504   | 5906.1858  | 18192.1962  |
|                | Bib-Shorts        | 6533.274   | 8477.058   | 6533.274   | 21543.606   |
|                | Shorts            | 3563.406   | 3923.346   | 3743.376   | 11230.128   |
|                | Tights            | 7828.956   | 9853.686   | 9859.0628  | 27541.7048  |
|                | Gloves            | 6817.938   | 11230.9888 | 7332.1706  | 25381.0974  |
|                | Caps              | 440.8525   | 700.1775   | 639.8479   | 1780.8779   |
|                | SubCategory Total | 29971.9329 | 41683.7603 | 34013.9171 | 105669.6103 |
| Category Total |                   | 29971.9329 | 41683.7603 | 34013.9171 | 105669.9103 |

Bikes

| Category       | SubCategory       | Jan         | Feb         | Mar         | Total       |
|----------------|-------------------|-------------|-------------|-------------|-------------|
| Bikes          | Road Bikes        | 672148.884  | 1345479.743 | 778022.7529 | 2795651.38  |
|                | Mountain Bikes    | 499562.0668 | 808888.6166 | 581875.2828 | 1890325.966 |
|                | SubCategory Total | 1171710.951 | 2154368.36  | 1359898.036 | 4685977.346 |
| Category Total |                   | 1171710.951 | 2154368.36  | 1359898.036 | 4685977.346 |

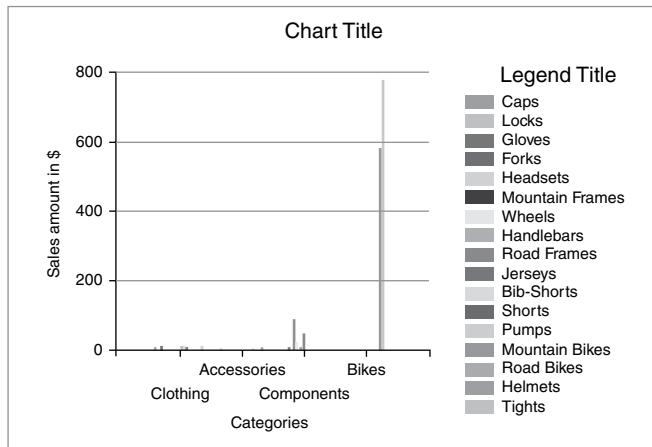
Components

| Category       | SubCategory       | Jan         | Feb         | Mar         | Total       |
|----------------|-------------------|-------------|-------------|-------------|-------------|
| Components     | Headsets          | 1784.682    | 4040.022    | 5111.88     | 10936.584   |
|                | Mountain frames   | 44023.2746  | 108360.7628 | 83686.3161  | 236070.3535 |
|                | Forks             | 2065.41     | 3717.738    | 4130.82     | 9913.968    |
|                | Road Frames       | 44494.5131  | 41904.0135  | 46293.1435  | 132691.6701 |
|                | Wheels            | 17155.503   | 22941.237   | 23089.089   | 63185.829   |
|                | Handlebars        | 1608.7395   | 2267.6475   | 2398.6      | 6274.987    |
|                | SubCategory Total | 111132.1222 | 183231.4208 | 164709.8486 | 459073.3916 |
| Category Total |                   | 111132.1222 | 183231.4208 | 164709.8486 | 459073.3916 |

**Figure 9.3** A list report.

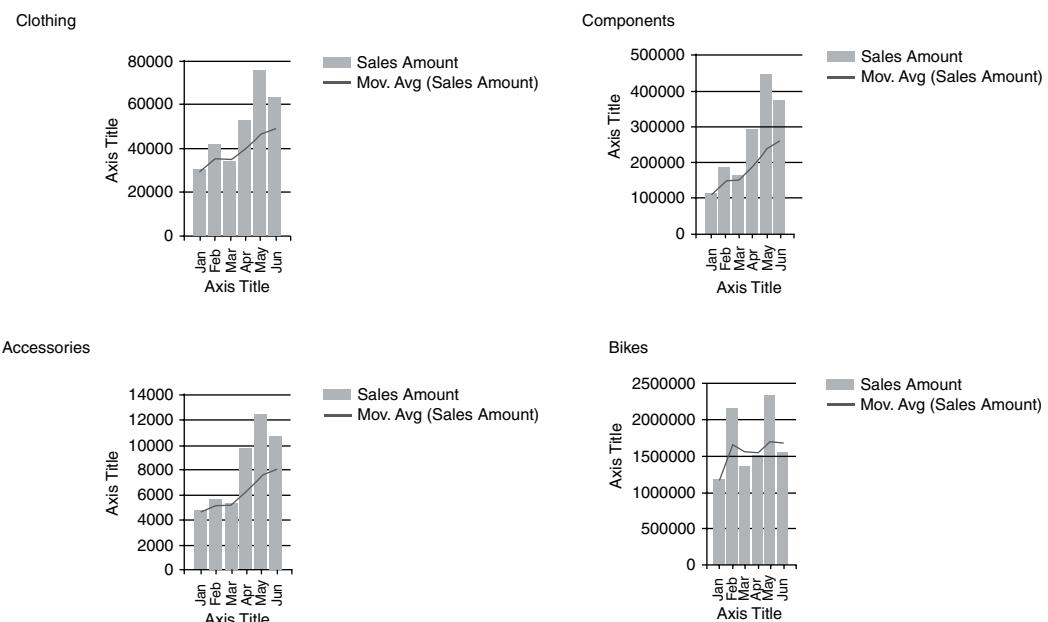
for each Category group. The records in the data set are grouped on Category and SubCategory. There are three Category Groups (Clothing, Bikes, and Components) in this list report.

- **Chart reports:** Chart reports provide a visual context for a lot of different kinds of data. There are several chart forms that can be used in the chart report such as bar chart, line graph, column chart, scatter plot, pie chart, etc. Figure 9.4 depicts a chart report of the total sales amount for each product in each Category and SubCategory. Figure 9.5 shows a chart report with trend lines for each category.



**Figure 9.4** A chart report.

**Note:** The colored figure is provided in the accompanying CD.



**Figure 9.5** A chart report with trend lines for each category.

| Student No. | Marks in SQL | Marks in SSRS | Marks in SSIS | Total | Percent | Grade |
|-------------|--------------|---------------|---------------|-------|---------|-------|
| A101        | 55           | 66            | 77            | 198   | 66.00%  | B     |
| A102        | 59           | 60            | 77            | 196   | 65.333% | B     |
| A103        | 70           | 80            | 90            | 240   | 80.00%  | A     |

**Figure 9.6** A gauge report.

- **Gauge reports:** These are the reports with gauge controls. If gauge controls are appropriately designed, one look at the gauge and you can say whether the enterprise is doing well, requires attention (not immediate though), or is in a bad state. Gauge reports depict values against a certain threshold. Figure 9.6 depicts a gauge report of student's performance. There are three color zones indicated on the gauge: red, amber, and green. Red depicts "immediate attention required", amber depicts "cause for concern but not urgent", and green depicts "things are going good". In our report on student's performance, students with total marks  $> 79\%$  are awarded "A" grade and are therefore in the green zone. Students with total marks  $>= 65$  but  $< 80\%$  are awarded "B" grade and are therefore in the amber zone. Students with total score less than 65% (not qualified) are awarded "C" grade and are in the red zone. In this example, the threshold value for each gauge is 65% with an upper limit of 100%.

### 9.2.2 Report Delivery Formats

In an attempt to ensure on-time delivery of information to the right decision makers, several technologies could be used for delivering reports in the enterprise. Here are some common formats of report delivery:

- **Printed reports:** Used only when really essential.
- **Secure soft copy:** You are already familiar with formats like un-editable PDF files, ZIP files, password-protected documents, documents with watermark, etc.

- **Email attachments:** Reports could be attached to emails.
- **Embedded emails:** Reports could be embedded into email and protected to ensure that they cannot be printed, saved, or forwarded.
- **FTP:** Reports could also be transferred to local systems by file transfers with security controls.
- **Link to reports:** The enterprise may choose to save reports to a central server and only provide a link through email.
- **Worksheet, PowerPoint Presentation, text:** Some users may need data for their own analysis and hence may need reports in Excel, PowerPoint, MS Word, or HTML/XML formats.
- **eBOOK:** Reports can now be grouped and formed into an eBOOK for publication to users. Some authorized users may be allowed to download the same onto their mobile devices.

Before we jump to the core of reporting in the OLAP world, let's quickly recall the process relating to production of reports and the role of reporting tools. Report development, like ETL or program development, has essential phases of requirements analysis, design, development, testing, production, and distribution. Typically, tools are used in design, development, and distribution phases. Design area includes several aspects like choosing data sources, columns from RDBMS, layout development, sample data generation, SQL query processing, and report preview. Typically, report specifications could be saved for future use. Every time the report needs to be generated, the user/admin user could either enter the variable parameters such as date, notes, filters, etc. at run time or provide these parameters in a file and include the file in the command for report generation. Enterprises use several reporting tools to meet the growing needs of different types of users and also for distribution/archiving of generated reports. Reporting tools also have the ability to generate only summary or detailed reports. Report developers may catalog common filters, sort sequences, columns for selection, and computed fields to enhance flexibility of reporting.

### 9.3 ENTERPRISE REPORTING CHARACTERISTICS IN OLAP WORLD

---

Enterprises invest money and efforts to help decision makers gain access to the right information at the right time on the right device. Some of the critical focus areas of enterprise reporting are as follows:

- **Single version of truth:** The value of providing the same “fact value” irrespective of the path the user has taken to reach for the data is of paramount importance in reporting.
- **Role-based delivery:** This feature is critical to avoid information overload.
- **Anywhere/anytime/any-device access:** Users have their own preferences and therefore flexibility needs to be built to ensure that users come to the same source of information again and again and don't find alternative ways to decision making.
- **Personalization:** Users' choices of delivery method, format like PDF/worksheet/CSV, book marking, customizing (in terms of language), etc. will need to be addressed.
- **Security:** Enterprises have huge concern over the unauthorized access to business-critical information, hacking by malicious sources, inadvertent leakage of business data, etc. The security framework needs to be thoroughly examined before implementing reporting.
- **Alerts:** Decision makers need immediate notification of threshold breaches of critical business KPIs. These alerts need to be delivered to various devices such as laptop, mobile devices in different forms like email, sound, voice message, SMS text, pop-up, etc. depending on user preferences.

- **Reports repository:** Many enterprises are attempting to create secure report repositories that service a wide range of users and have flexible delivery/viewing options. Managing the content of the repository itself is a task that needs specialized competencies and tools.

Enterprises have witnessed and reported huge gains arising out of business analytics reporting. Enterprise reporting is not about buying tools and converting existing reports to new platforms, but is all about building a culture of “fact-based decision making”. This is essentially change management. Some of the long-term benefits of these investments include:

- **Enhanced collaboration:** Teams around the globe will use facts to be on the same page and look at the same facts for designing powerful alternatives, collaboratively.
- **Objective communication:** Managers are no longer required to exercise “gut feel” to make “seat-of-the-pant” decisions. They can look at facts.
- **Reduced cost of audits/reviews:** As all members of the team are on the same page, there is no need to waste meeting time to bring them to the same level of understanding by briefings. They are already aware.
- **Reduced decision cycle time:** Enterprises adopting structure business analytics-based reporting will make consistent decisions and that too faster.
- **Better predictability and ability to influence goals:** Facts can be used to innovate and gain truly sustainable market leadership and competitive advantage.

## 9.4 BALANCED SCORECARD

---

We realize that “Measurement, Analysis, and Knowledge Management” (Malcolm Baldrige criteria) are critical for effective management of the organization and for a fact-based, knowledge-driven system for improving performance and competitiveness. It serves as a foundation for the performance management system.

How many business people take the strategy talk, that has an important link with the performance and competitiveness of their organization, seriously? Surprisingly, not many. According to a ***Fortune Magazine*** survey, “Only 5% of the workforce understands the strategy, 85% of executives spend less than one hour per month discussing strategy, only 25% of managers have incentives linked to strategy.” So, there is always a need for a strategic planning and management system which could

- Align business activities to the organization’s vision and strategies.
- Improve internal and external communication.
- Monitor the organization’s performance against its strategic goals.

The balanced scorecard is one such strategic planning and management tool used by organizations to align their business activities with their organization’s vision and strategies.

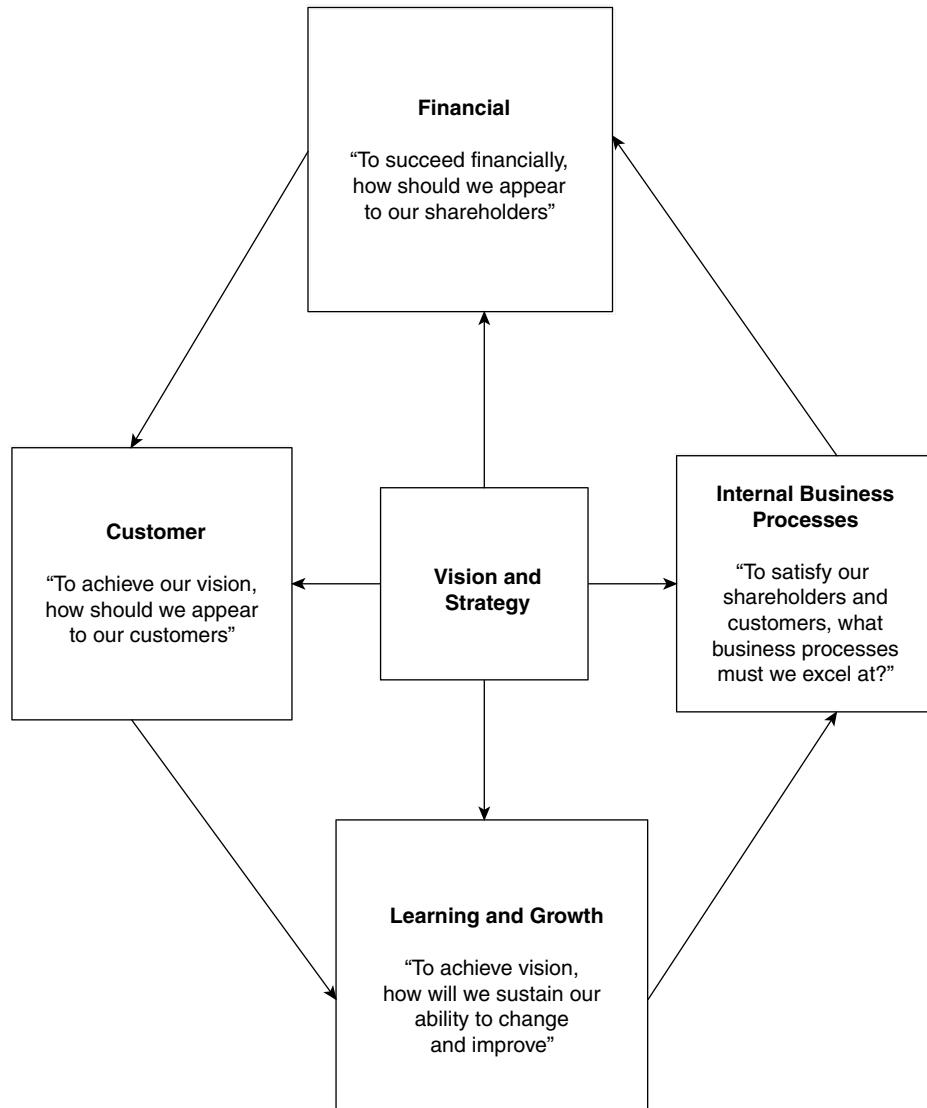
Dr. Robert S. Kaplan and David P. Norton gave the world the balanced scorecard in 1992. They have to their credit a book titled *The Balanced Scorecard* which was published in 1996. This was followed by their second book titled *The Strategy Focused Organization* in 2004. In this book, they have proposed the “Strategic Linkage Model” or “Strategy Model”.

The balanced scorecard is designed to identify the financial and non-financial measures and attach some targets to them so that at a later point in time during review it is possible to decide whether the organization’s performance has met the set expectations or not.

### 9.4.1 Four Perspectives of Balanced Scorecard

The balanced scorecard maps the organization's strategic objectives into the following four perspectives as depicted in Figure 9.7:

- **Financial perspective:** The financial perspective addresses the question of how shareholders view the firm and which financial goals are desired from the shareholder's perspective.



**Figure 9.7** The four perspectives of the balanced scorecard.

- **Customer perspective:** The customer perspective addresses the question of how the firm is viewed by its customers and whether the firm will be able to fulfil customers' expectations.
- **Internal business process perspective:** The business process perspective identifies the processes in which the organization must excel to satisfy its shareholders' expectations of good financial returns and also keep its customers happy and loyal.
- **Learning and growth perspective:** The learning and growth perspective identifies the competencies that the employees of the organization must acquire for long-term improvement, sustainability, and growth.

#### 9.4.2 Balanced Scorecard as Strategy Map

The balanced scorecard was earlier plotted as a four-box model. This model has evolved since then and is now plotted as a strategy map. The strategy map places the four balanced scorecard perspectives into a causal hierarchy. The causal hierarchy shows that the objectives support each other and that delivering the right performance in the lower perspectives will help achieve the objectives in the upper perspectives. The balanced scorecard strategy map depicts how the company intends to create value for its shareholders and customers.

Figure 9.8 shows the strategy map with the cause and effect relationship among four perspectives of the balanced scorecard. Each of the four balanced scorecard perspectives can be described in terms of the following parameters:

- **Objectives:** What is it that you wish to achieve?
- **Measurement:** How do you know if you have been able to achieve the stated objectives?
- **Target:** What is the level of performance expected or the level of improvement expected?
- **Initiative:** What is it that you would do to achieve your targets and thereby your objectives?

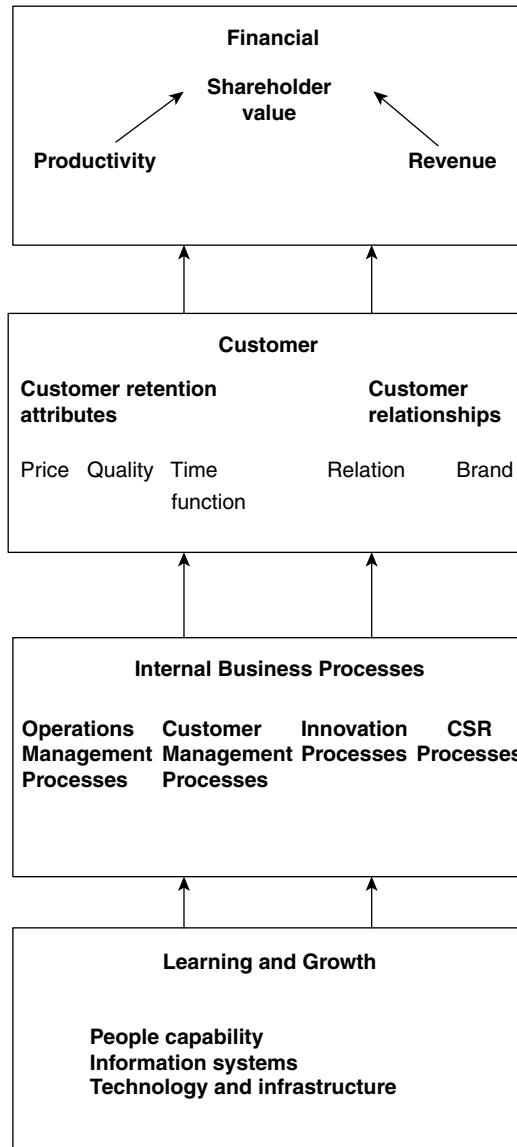
#### 9.4.3 Measurement System

The measurement system interconnects the objectives and the measures in the various perspectives so that they can be validated and managed. For example, let us consider an airline company XYZ which wishes to reduce its operating costs. It has decided to reduce the number of planes but at the same time increase the frequency of the flights among different cities to increase its revenue.

The question is: How will the airlines company retain its customers and increase its customer satisfaction rate?

An analysis of customers' data reveals that on-time delivery of services makes customer happy and satisfied. To ensure on-time departure of flights, the airline company has to ensure that the ground turnaround time is less. (Ground turnaround time is the time that the ground crew takes to clean and maintain the flight and also allow the passengers to get in and settle down.) But at the same time, the airline company has to ensure that the flights depart at the scheduled time.

Let us look how the airline improves the quality of service and reduces the ground turnaround time. The airlines achieves this by training and improving the skill sets of the ground crew. The cause and effect relationship among the four balanced scorecard perspectives concerning the airlines example is depicted in Figure 9.9. The tabular representation of the objectives, measures, target, and initiative concerning the airlines is given in Figure 9.10.

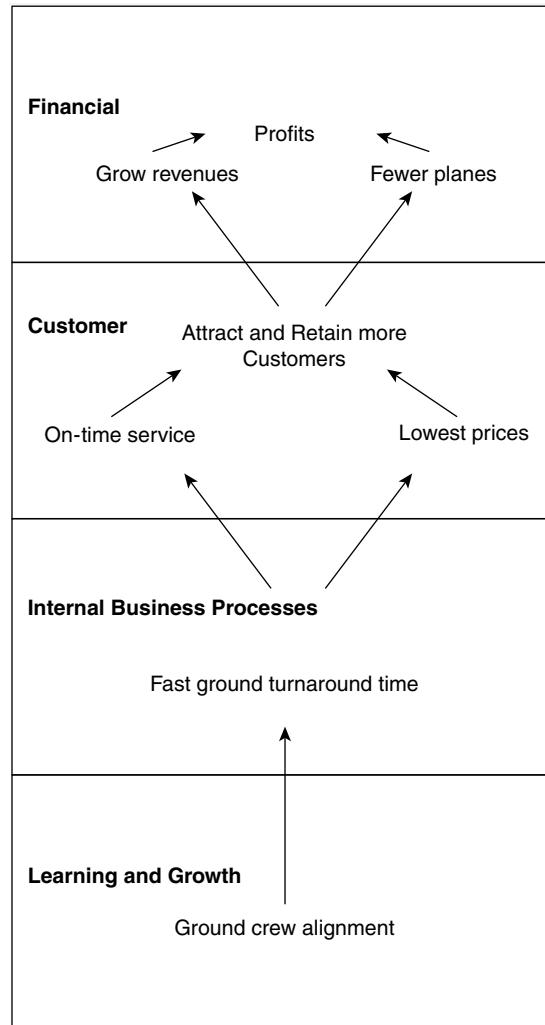


**Figure 9.8** Strategy map depicting four balanced scorecard perspectives in causal hierarchy.

#### 9.4.4 Balanced Scorecard as a Management System

The balanced scorecard translates an organization's missions and strategies into tangible objectives and measures. Figure 9.11 depicts the following four steps for designing the balanced scorecard:

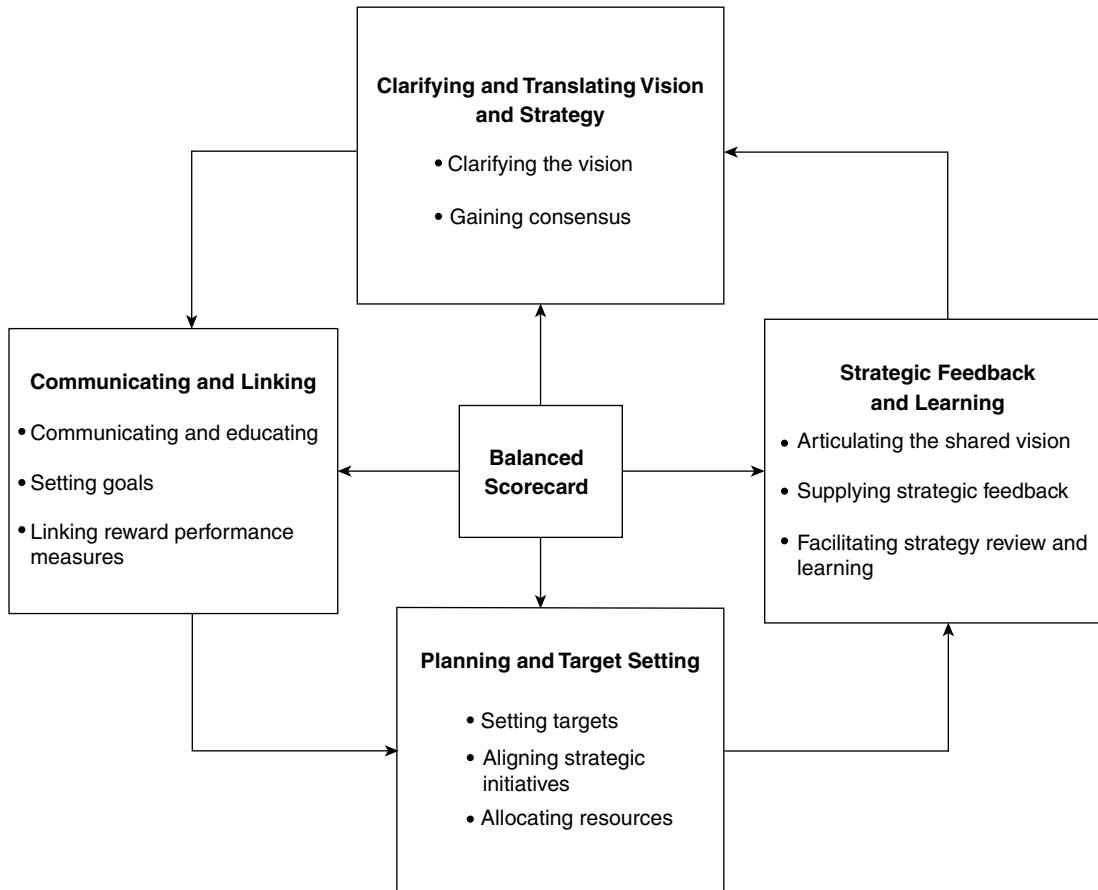
- Clarify and translate vision and strategy.
- Communicate and link strategic objectives and measures.



**Figure 9.9** The cause and effect relationship among the four balanced scorecard perspectives in the case of the airlines example.

| Objectives                  | Measurement                         | Target         | Initiative              |
|-----------------------------|-------------------------------------|----------------|-------------------------|
| Fast ground turnaround time | On-ground time<br>On-time departure | 30 mins<br>90% | Cycle time optimization |

**Figure 9.10** The tabular representation of the objectives, measures, target and initiative concerning the airlines example.



**Figure 9.11** Four steps for creating the balanced scorecard.

- Plan, set targets, and align strategic initiatives.
- Enhance strategic feedback and learning.

The balanced scorecard not only measures performance but also communicates and aligns the strategies throughout the organization. Some benefits of the balanced scorecard are as follows:

- Translating the organization's strategies into measurable parameters.
- Communicating the strategies to all the individuals in the organization.
- Alignment of individual goals with the organization's strategic objectives.

Altogether, the balanced scorecard translates the visions and strategies into a set of objectives and measures across a balanced set of perspectives. The scorecard includes measures of desired outcomes as well as processes that will drive the desired outcomes for the future.



### Remind Me

- The balanced scorecard is a strategic planning and management tool.
- The balanced scorecard is designed to emphasize both financial as well as non-financial aspects of the organization.
- The four perspectives of the balanced scorecard are: financial, customer, internal business process, and learning and growth.
- Each perspective can be described in terms of objectives, measures, targets, and initiatives.
- The four steps required for designing the balanced scorecard are:
  - Clarify and translate vision and strategy.
  - Communicate and link strategic objectives and measures.
  - Plan, set targets, and align strategic initiatives.
  - Enhance strategic feedback and learning.



### Connect Me (Internet Resources)

- *The Balanced Scorecard – Measures that Drive Performance*, Harvard Business Review, Feb. 1992
- *The Balanced Scorecard: Translating Strategy into Action*, Harvard Business School Press, Boston (1996)



### Point Me (Books)

- *The Balanced Scorecard*, Kaplan and Norton, 1996.
- *The Strategy-Focused Organization: How Balanced Scorecard Companies Thrive in the*

*New Business Environment*, Kaplan and Norton, 2004.



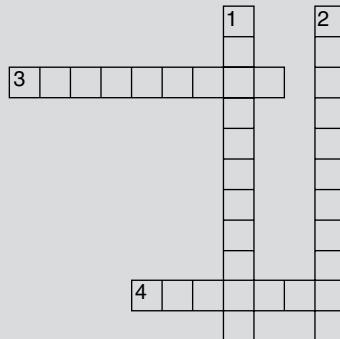
## *Test Me Exercises*

### **Answer me**

1. What are the four perspectives of the balanced scorecard?
2. Why is there a need to translate the balanced scorecard into a strategy map?
3. Why does the balanced scorecard take into consideration the non-financial measures as well?
4. What are the four basic steps required in the design of the balanced scorecard?
5. Can the balanced scorecard be plotted only for the organization, or can we plot it for a business unit as well?



## *Scorecard Puzzle*



### **ACROSS**

3. It is a strategic planning and management tool.
4. This perspective addresses the question of how shareholders view the firm.

### **DOWN**

1. It places the four perspectives into causal hierarchy.
2. It is what the organizations do to achieve their targets and thereby their objectives.

### **Solution:**

1. Strategy Map
2. Initiatives
3. Scorecard
4. Finance

## 9.5 DASHBOARDS

Corporate or enterprise dashboards are changing the way we look at information and the way we analyze our business. A well-constructed corporate dashboard answers four basic questions:

- Where?
- What?
- How?
- Why?

Instead of wading through pages of disparate operational data, dashboards portray critical operating and strategic information about an organization using a collection of powerful graphical elements. One quick glance at the dashboard tells users the key performance indicators and metrics used to measure and monitor the company's performance. Dashboards help in

- Better analysis.
- Better tracking.
- Proactive alerting.

### 9.5.1 What are Dashboards?

What is the first thing that comes to your mind when you hear the word “dashboard”?



Yes you guessed it right... It is indeed an automobile's dashboard!

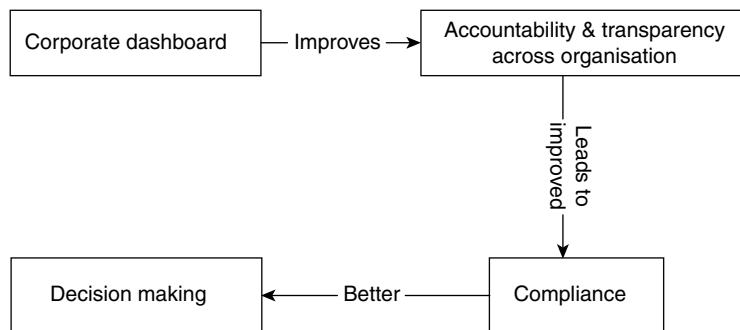
Dashboard is a control panel in an automobile that provides the driver with all the information regarding the operations and control of the vehicle. The dashboard used in Information Technology almost resembles that of an automobile, but is more interactive than an automobile dashboard.

So, what really is a dashboard? A dashboard is a graphical user interface that organizes and presents information in a way that is easy to read. It provides at-a-glance insight to what is actually happening in an organization. Dashboards have the following attributes:

- They display data relevant to their own objectives.
- They throw light on key performance indicators and metrics used to measure and monitor the organization's performance.
- Since dashboards are designed to serve a specific purpose, they inherently contain pre-defined conclusions that help the end-user analyze his or her own performance.

## 9.5.2 Why Enterprises Need Dashboards?

Figure 9.12 describes the benefits accruing to enterprises through dashboards.



**Figure 9.12** Importance of dashboards for enterprises.

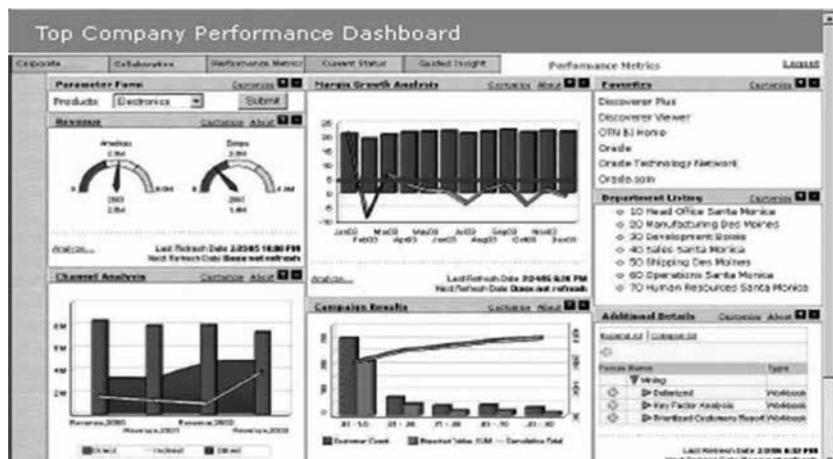
## 9.5.3 Types of Dashboard

### Enterprise Performance Dashboards

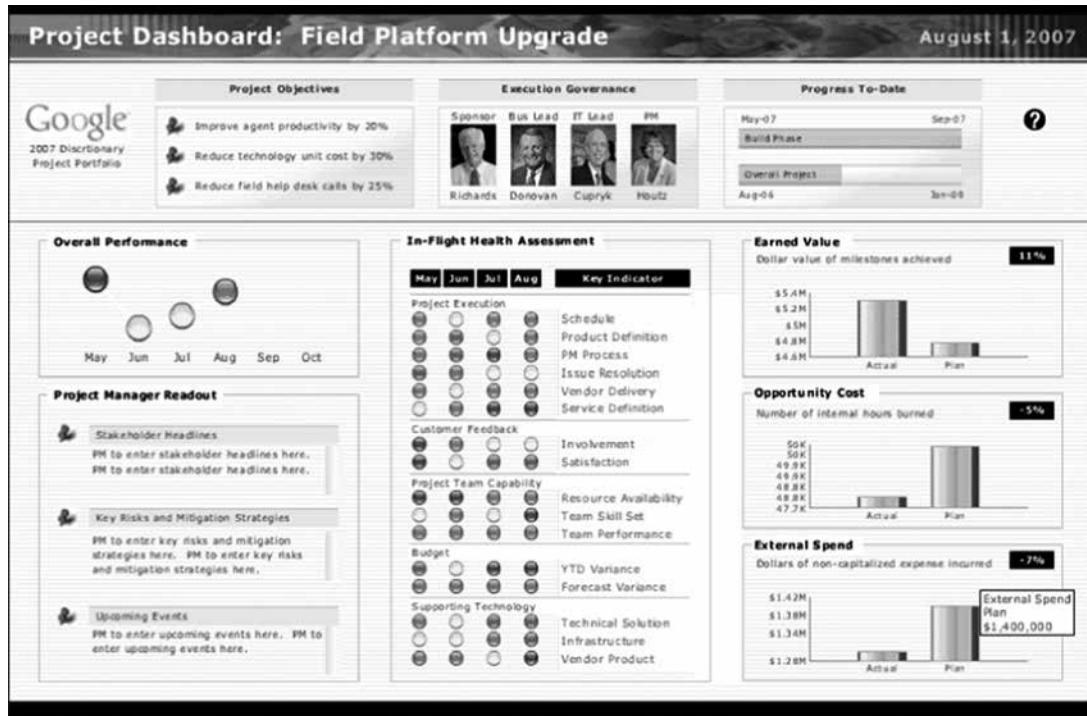
These dashboards provide an overall view of the entire enterprise, rather than of specific business functions/process. Typical portlets in an enterprise performance dashboard include:

- Corporate financials.
- Sales revenue.
- Business Unit KPIs (key performance indicators).
- Supply chain information.
- Compliance or regulatory data.
- Balanced scorecard information.

Figure 9.13 and 9.14 are samples of enterprise performance dashboards.



**Figure 9.13** A sample enterprise performance dashboard.



**Figure 9.14** Another sample enterprise performance dashboard.

### *Customer Support Dashboards*

Organizations provide this type of dashboard to its customers as a value-add service. A customer support dashboard provides customers their personal account information pertaining to the business relationship, such as

- Online trading.
- Utility services.
- Entertainment.
- B2B SLA (business-to-business service-level agreement) monitoring.

### *Divisional Dashboards*

These are one of the most popular dashboards used to provide at-a-glance actionable information to division heads, operational managers, and department managers. Each division has its own set of KPIs which can be visually displayed on the enterprise dashboard. Typical divisional dashboards include:

- Purchasing dashboards.
- Supply chain dashboards.
- Operations dashboards.
- Manufacturing dashboards.
- Quality control dashboards.
- Marketing dashboards.

- Sales dashboards.
- Finance dashboards.
- Human resources dashboards.

## **9.6 HOW DO YOU CREATE DASHBOARDS?**

---

Most dashboards are created around a set of measures or key performance indicators (KPIs). KPI is an indicator of the performance of a task, and it reveals the performance that is below the normal range so that corrective action can be taken. It draws attention to problem areas. The measures used in the dashboard should be relevant and support the initial purpose of the dashboard.

### **9.6.1 Steps for Creating Dashboards**

#### ***First Step***

Understand/identify the data that will go into an enterprise dashboard. Enterprise dashboards can contain either/both of the following mentioned data:

- Quantitative data.
- Non-quantitative data.

Quantitative data is the data that gives an idea of what is currently going on. Examples of quantitative data for an Education dashboard:

- No. of student batches.
- No. of learning programs.
- No. of students who have successfully qualified the internal certification.
- No. of students being trained on the various learning programs.

Examples of non-quantitative data for an Education dashboard:

- Salient features of the foundation learning program.
- Challenges faced by the instructor in classroom training.
- Users comments on the effectiveness of the learning program.

#### ***Second Step***

Decide on the timeframes. The various timeframes could be

- This month to date.
- This quarter to date.
- This year to date.
- Today so far.

#### ***Third Step***

Decide on the comparative measures. The comparative measures could be

- The same measure at the same point in time in the past.
- The same measure at some other point in time in the past.

- A distinct yet relative measure.
- A competitor's measure.

### Last Step

Decide on the evaluation mechanisms. The evaluation can be performed as follows:

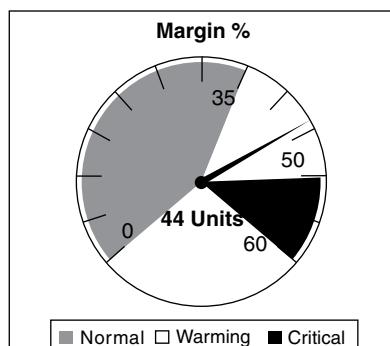
- Using visual objects, e.g. traffic lights.
- Using visual attributes, e.g. red color for the measure to alert a serious condition.

### 9.6.2 Tips For Creating Dashboard

- **Don't make your dashboard a data repository:** Avoid using additional and unwanted data. Focus on the primary goal of dashboard. Too much data makes the dashboard look cluttered and dilutes the actual information you are trying to present.
- **Avoid fancy formatting:** To communicate the actual information effectively through your dashboard, it is very important to present the data as simply as possible. It is important to focus on data rather than shiny graphics.
- **Limit each dashboard to one printable page:** Dashboards provide at-a-glance views into the key measures relevant to a particular objective. So, it is important to keep all the data in one page. It ensures better comparison between the different sections of the dashboard and process the cause and effect relationship more effectively. When the user has to scroll left, right, or down, these benefits are diminished. On the contrary, when dashboards bring all the information on one page then one glance can give a complete insight into the organization's performance. It also helps in identifying the problems where corrective actions are required.

Let us take one example. All organizations set certain goals that they wish to achieve. They select certain criteria to evaluate their performance. Suppose they want to visually see what their margin percent is and monitor their performance against the defined goals.

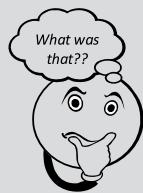
One way to visually depict this is by using a gauge where the indicator can clearly indicate if the goal was achieved or not. If the indicator is green then it must have met the goal, and if it is in red or yellow then corrective actions have to be taken. Figure 9.15 depicts a sample gauge indicator.



**Figure 9.15** A sample gauge indicator.

Dashboards have the following benefits:

- They place all critical information in just one screen. One need not flip through different pages to see the desired critical information.
- They help in improved decision making.
- They help in rapid problem detection.
- They help in better analysis of performance.
- They identify the trends and corrective actions to improve the organization's performance.



### *Remind Me*

- Dashboard is a graphical user interface that provides at-a-glance insight into what is actually happening in an organization.
- Types of dashboard.
  - Enterprise dashboard.
  - Customer support dashboard.
  - Divisional dashboard.
- Steps for creating a dashboard:
  - Identify the data that will go into an enterprise dashboard.
  - Decide on the timeframe.
  - Decide on the comparative measures.
  - Decide on the evaluation mechanisms.
- Benefits of a dashboard:
  - Better analysis.
  - Better tracking.
  - Proactive alerting.



### *Point Me (Books)*

- *Information Dashboard Design: The Effective Visual Communication of Data*, Stephen Few.
- *Say It With Charts: The Executive's Guide to Visual Communication*, Gene Zelazny.



### *Connect Me (Internet Resources)*

- [en.wikipedia.org/wiki/Dashboard\\_\(business\)](http://en.wikipedia.org/wiki/Dashboard_(business))
- [www.appsbi.com/what-are-dashboards](http://www.appsbi.com/what-are-dashboards)



## *Test Me Exercises*

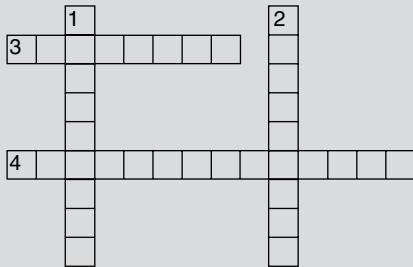
### **Answer me**

1. What are dashboards?
2. Why do organizations need a dashboard?
3. What are the various attributes of a dashboard?
4. What is the difference between quantitative data and non-quantitative data?
5. What makes a dashboard good or bad?



## *BI Crossword*

Dashboard



### **ACROSS**

3. Dashboards eases \_\_\_\_\_ making.
4. One of the chief benefits of dashboards

### **DOWN**

1. It helps monitor the performance of an enterprise.
2. Dashboard is a collection of powerful \_\_\_\_\_ elements.

### **Solution:**

- |  |  |
|--|--|
| <ol style="list-style-type: none"> <li>1. Scorecard</li> <li>2. Graphical</li> </ol> | <ol style="list-style-type: none"> <li>3. Decision</li> <li>4. Accountability</li> </ol> |
|--|--|

## **9.7 SCORECARDS VS. DASHBOARDS**

By now you have a fair understanding of dashboards and scorecards. Can the terms “dashboard” and “scorecard” be used interchangeably? Or is there a difference between the two?

Before we get into the differences, let us look at the commonality between a balanced scorecard and a dashboard. Both are measurement systems, built on integrated data, usually a data warehouse. Both provide

the organization with an insight/view on business performance. An ideal scenario is for the organization to measure its performance against its balanced scorecard, then drill down to the data warehouse (has data from several operational/transaction processing systems) to detect the possible causes of a problem. To know the current operational status of the problem area, you can revert to the dashboard. This was the tracking from the balanced scorecard to the dashboard. Let us look at a reverse scenario. We can start with a problem at hand. We view the current operational status of the problem at hand via the dashboard. We then take some remedial step at the operational level to counter the problem, monitor it through the dashboard, and then trace it back to the balanced scorecard to view the result of this action at the operational level.

A balanced scorecard is a business performance measurement (BPM) used mostly at the senior management level to view the business performance through indicators. It enables senior executives to perform a pulse-check on how the organization is performing in terms of accomplishing its strategic objectives. In contrast, an enterprise dashboard is equivalent to an automotive dashboard that displays real time changes in tactical information often displayed as charts, graphs, and gauges. A dashboard is a business activity (process) monitoring (BAM) or business process measurement (BPM) used most by the operational managers to monitor day-to-day operations through visualization.

### **9.7.1 KPIs: On Dashboards as well as on Scorecards**

A KPI is a metric that is tied to a target. KPIs usually indicate how far a metric is from its pre-determined target. KPIs are designed to let a business user know at a glance whether results are on target or off target. Balanced scorecards use KPIs almost entirely to measure success in each area of the business as defined in the strategy map. On the other hand, dashboards use KPIs to highlight milestones in operations.

### **9.7.2 Indicators: On Dashboards as well as on Scorecards**

Indicators, sometimes called icons, are graphical elements that give visual cues about performance. For example, traffic light symbols can be used as indicators – red to indicate a problem, yellow to indicate a potential concern, and green to show that performance is meeting or exceeding its goal.

What indicators are commonly used in dashboards and scorecards? Dashboards use mostly graphs, grids, gauges, and a variety of visualization techniques to highlight the operational data. On the other hand, scorecards commonly use symbols and icons.

An example would make it clearer. Let us look at the balanced scorecard of a manager, responsible for the customer service function of an enterprise. His scorecard might have the following indicators:

- Minimum resolution time.
- Maximum resolution time.
- Percentage of issues resolved at first attempt.
- Customer satisfaction survey.

All the above indicators will be considered for a period of time (generally a month or a quarter). The indicators will also be compared against the pre-defined goals and help analyze the manager's performance. The manager's dashboard might use the following indicators:

- Number of inbound calls (the calls customer initiates to the help desk) in queue.
- Number of calls in escalation.

- Average call resolution time.
- Current CSRs (customer service representatives) on-line.

Based on the above example, the differences between dashboards and balanced scorecards can be summarized as follows:

- Dashboards can provide tactical guidance while scorecards can assess the quality of execution.
- Scorecards inherently measure against strategic goals while dashboards present real time information.

Referring to Wayne Eckerson's "Performance Dashboards", the distinction between balanced scorecards and dashboards stated as follows:

*The primary difference between the two is that dashboards monitor the performance of operational processes, whereas scorecards chart the progress of tactical and strategic goals.*

Table 9.2 points out key differences between dashboards and scorecards. The enterprise dashboard is a useful weapon in the hands of today's managers looking to steer their business in the right direction and to always keep it on course. There are operational dashboards designed for operational managers to help them discharge their day-to-day operational activities with ease. There are strategic dashboards that are used by the C class (CEO, COO, CIO, CFO, etc.) and by business unit heads to assess metrics that represent corporate strategy and direction.

What next after the standard reports and enterprise dashboards. This brings us to the next topic, i.e. analysis.

**Table 9.2** Key differences between dashboards and scorecards

| Feature      | Balanced Scorecard       | Dashboards           |
|--------------|--------------------------|----------------------|
| Business use | Performance Measure      | Monitor Operations   |
| Users        | Senior Executives        | Operations Manager   |
| Used by      | Corporate/Unit           | Corporate/Department |
| Data         | Summary                  | Detail               |
| Refresh      | Monthly/Quarterly/Annual | Intra-day            |

## 9.8 THE BUZZ BEHIND ANALYSIS...

We will discuss three major kinds of analysis here:

- Funnel analysis.
- Distribution channel analysis.
- Performance analysis.

### 9.8.1 Funnel Analysis

Let us look at what is funnel analysis.

## Picture this...

You are a visitor on a popular website that sells books on-line. You are required to follow a set of steps to buy the book. Let us list down the set of steps...

- Visit the website that sells the book.
- Search for a particular book.
- Add the book to your cart.
- Buy the book by making the payment after careful validation, and complete the check-out process.
- Input the shipping address so that the book can be successfully shipped.

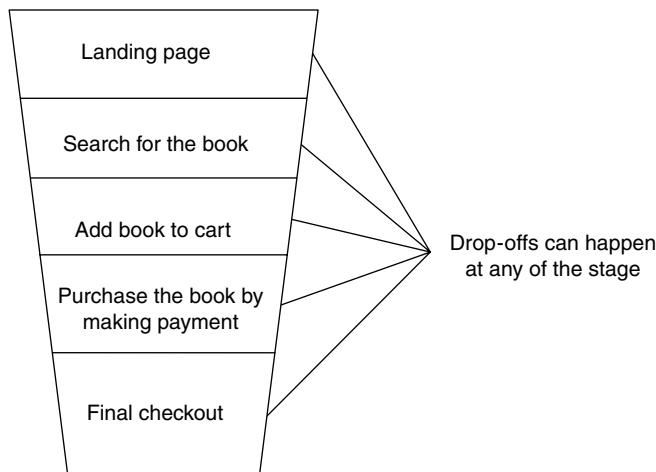
These are the steps from an end user's viewpoint, i.e. a visitor to the books' on-line site. In order to successfully complete the transaction, you are required to complete each step listed above – from visiting the website to making the payment and completing the check-out process. This is called the **conversion funnel**, shown in Figure 9.16. If you were able to successfully purchase the book, you have been converted from a visitor to a customer.

Now let us get into the shoes of a website analyst whose job is to analyze and recommend changes to the site in a way that maximizes the chances of the visitors eventually turning into customers. The website analyst will typically use a funnel mechanism to work out how the website is performing. The funnel analysis works on a simple philosophy. You work out how much you put in the top, and how much you get out, at the bottom. The more you put in the top, the more you'll get out, at the bottom.

The website analyst will try to analyze the bottlenecks at each step. He will want to zero down to the cause at every step that may potentially result in a "drop-off". Once the spots have been identified, next is to fix the same to improve the conversion rate.

Let us look at what bottlenecks can possibly occur at each of the steps mentioned above:

- The navigation to the home page of the website is difficult. Visitors rarely come in through the home page although they come in onto different pages of the website in as many different ways.
- The visitor doesn't really have a fascination for the search button on the website and struggles while searching for a particular book.
- The visitor faces problem in adding the book to the shopping cart.



**Figure 9.16** The typical funnel of book-selling website.

- He/she experiences problems while making the payment. Far too many details are asked for and the payment page times out ever so frequently.
- He/she faces difficulty in editing the shipping address once the address has been fed in.

So, what is required to have a good conversion rate?

- Easy navigation to pages to allow users to progress through every stage.
- The required action should be easy, very obvious, and compelling.

The essence of conversion rate optimization is to get a majority of visitors through the funnel. The focus should be on fixing the problem at the correct stage. If you are losing more people immediately after the landing page, it makes little sense to fix the problem somewhere down the line. However, if the drop-off at the landing stage is very small, then the problem is probably on one of the pages down the line.

One question that we must answer is: "Does every website have a conversion funnel, even those that do not sell anything?" The answer is "Yes". Assume a services company has a website. The website has a home page and through the home page, you can navigate to a bunch of other pages. Each page describes a particular service offered by the services company. Each page leads to a "contact us" page that details out the steps on "how one can get in touch with the services company". A successful conversion here will be when a visitor clicks on a button of the "contact us" form. However, it is not always as easy as it sounds. A vast majority of websites experience a "drop-off" between each stage in their conversion funnel. This means that for some reason visitors are failing to progress to the next stage and are not turning into customers. Your aim here is to lose as few people at any point in the process as possible.

In conclusion, there are two rules that can be followed:

**Rule 1:** Do away with/eliminate unnecessary steps to conversions as they just contribute to increased conversion funnel drop-off rates. The fewer the steps, the more likely a visitor will follow through with a conversion.

**Rule 2:** Use an effective call-to-action in every step of your path.

### 9.8.2 Distribution Channel Analysis

Distribution channels move products and services from businesses to consumers and to other businesses. Distribution channels are also known as marketing channels or channels of distribution. These consist of a set of interdependent organizations – such as wholesalers, retailers, and sales agents – involved in making a product or service available for use or consumption.

#### Picture this...

You would like to buy a personal computer. You can choose to buy it directly from the manufacturers by placing an order with them either in person or over the telephone (teleshopping) or over email or on-line through the Internet (online buying) or through several kinds of retailers including independent computer stores, franchised computer stores, and department stores.

Distribution channel structures usually range from two to five levels as described below:

- A two-level structure is directly from the manufacturer or provider to the consumer, i.e. manufacturer → consumer
- A three-level structure is: manufacturer → retailer → consumer
- A four-level structure is: manufacturer → wholesaler → retailer → consumer
- A five-level structure is: manufacturer → manufacturer's agent → wholesaler → retailer → consumer

This brings us to the question, “Is selling directly from the manufacturer to the consumer always the most efficient?” The answer is “No”. Intermediaries such as wholesalers, retailers, etc. provide several benefits to both manufacturers and consumers: benefits such as improved efficiency, a better assortment of products, routinization of transactions, and easier searching for goods as well as customers.

Another example from the hospitality domain is as follows: Several customers visit the “GoodFood Restaurant”. Some of these customers have read about the restaurant’s excellent ambience and quality food in newspapers, heard it as advertised over the television, or seen it on hoardings put up at prominent places and therefore have come in. A few others have just walked in to check out the restaurant, all by themselves. A few others came to know of the restaurant through their friends. Yet, a few others have heard it all from their colleagues, etc. What we are trying to convey here is that the restaurant attracts all sorts of guests.

Distribution channel analysis is the analysis of the various channels to determine which channel is the most efficient. In the restaurant example, it will be worthwhile to understand whether the word of mouth is fetching them very many customers, or is it their advertisement over television or the hoardings that is striking gold for them. The investment can then be very intelligently made by the restaurant owners on positioning their brand.

### 9.8.3 Performance Analysis

Let us start off with “why performance analysis?” And, let us try to explain this with an example. Alex was visiting his cousin, Justin, who had purchased an SUV (super utility vehicle) a couple of months back. Alex was also trying to arrive at a decision on which one should he go for at the end of the quarter when his company will award him some bonus. Conversations veered around to the performance of the SUV. Questions such as: “What’s the mileage like?”, “How is it to drive on rough terrains?”, “How much of maintenance is required?”, “What is the approximate total cost of ownership?”, etc. were being asked. The answers to these questions helped Alex understand the performance of the vehicle and also decide on the affordability of the vehicle.

Performance analysis is carried out “to improve a part of the organization/unit/function or to fix a problem that somebody has put forth”. Performance analysis helps uncover several perspectives of a problem or opportunity. It helps identify any or all drivers towards (or barriers) successful performance, and propose a solution system based on what is discovered.

Let us take another example. In almost every enterprise the world over, an employee is apprised of his/her performance in the performance review that is usually conducted annually or after every six months. This in turn will help him or her perform better because given the data, the areas of improvement are laid bare. For an organization, the performance analysis could be evaluating the performance indicators such as ROI (Return on Investment), ROA (Return on Assets), Return on Equity, etc. against those of its competitors in the domestic or global market.

Yet another example of performance analysis is an examination of the performance of current employees to determine if training can help reduce performance problems such as low output, uneven quality, excessive waste, etc.

Let us further look at the above explained three analyses in the light of a case study here.

#### Picture this...

A company “Infotech” with a major focus on the retail domain is hiring. The company recruits in large numbers from the national talent pool of Computer Science (CS) and Non-Computer Science (NCS) graduate engineers. These recruits have come in through various channels. Some came in through “Walk-in interviews”, some were “Referrals” (referred to by the employees of the company), some applied “On-line”

over the company's website, some came in through direct "Campus recruitment" (where in the company visited their college/university), etc. Some of these recruits (CS/NCS) have undergone "Industrial Training" with a corporate house while others are raw with no industrial training experience. Offers are made. Of the selected recruits, some accept the offer and others decline. Those who accept the offer are the prospective candidates and are identified using a "Candidate ID". Amongst those who accept the offer, some do not join the corporate/enterprise. Those who do join are called Trainees/Student Trainees and are on probation, and are put through a customized training program to prepare them for the floor. They are identified using a unique "Employee ID". No two employees in the organization can have the same "Employee ID". They are also assessed after the training gets over and before their actual life on the shop floor starts. Those who make it through the assessments are released to the shop floor/production/delivery. The "Infotech" company decides to perform some analysis to conclude the following:

- How many make it from the recruit stage to the employee stage? (**Funnel Analysis**)
- Which channel of recruitment ("On-line", "Campus Recruitment", "Walk-in interviews" or "Referrals") is the most profitable? (**Channel Analysis**)
- How many CS trainees make it through the training to eventually qualify for the shop floor? (**Performance Analysis**)
- How many NCS trainees make it through the training to eventually qualify for the shop floor? (**Performance Analysis**)
- How many trainees (CS/NCS, irrespective of their background) with an "Industrial Training" qualification in their kitty eventually qualify for the shop floor? (**Performance Analysis**)
- How many trainees (CS/NCS, irrespective of their background) without an "Industrial Training" qualification in their kitty eventually qualify for the shop floor? (**Performance Analysis**)
- How have the CS graduate engineers fared in their education starting from SSC → HSC/Diploma → Degree → Training? (**Performance Analysis**)
- How have the NCS graduate engineers fared in their education starting from SSC → HSC/Diploma → Degree → Training? (**Performance Analysis**)

We consider here a data set that is shared in an Excel sheet. The Excel sheet has 3796 rows and 18 columns of data. You can have a look at the sheet in the accompanying CD. Let us look at what the funnel analysis reveals:

The funnel analysis as depicted in Figure 9.17 clearly reveals that out of 3796 recruits who have been selected and made the offer, only 2597 accepted the offer and are the prospective candidates. Out of the

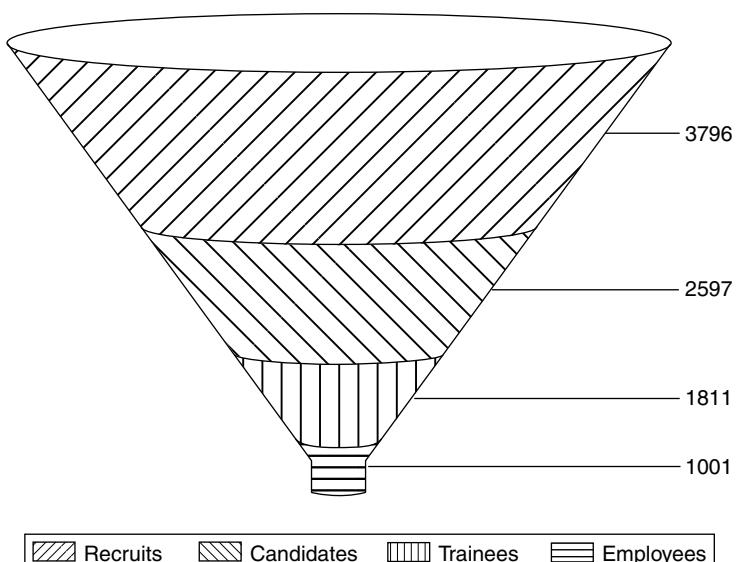
| S. No. | Column Name      | Column Description  |
|--------|------------------|---|
| 1.     | Candidate ID     | Used to uniquely identify a candidate. A candidate is a recruit who has accepted the offer made by "Infotech" |
| 2.     | Employee ID      | Used to uniquely identify an employee. A trainee becomes an employee upon joining the organization            |
| 3.     | First Name       | First Name of the Employee  |
| 4.     | Middle Name      | Middle Name of the Employee   |
| 5.     | Last Name        | Last Name of the Employee   |
| 6.     | DegreePercentage | Percentage scored in the Degree Examination   |

(Continued)

(Continued)

| S. No. | Column Name                         | Column Description   |
|--------|-------------------------------------|--|
| 7.     | 12 <sup>th</sup> /DiplomaPercentage | Percentage scored in the 12 <sup>th</sup> Grade or equivalent diploma              |
| 8.     | SSCPercentage                       | Percentage scored in SSC   |
| 9.     | University Name                     | The name of the university   |
| 10.    | Native State                        | The name of his/her native state   |
| 11.    | Background(CS/NCS)                  | Computer Science (CS) / Non Computer Science (NCS) background                      |
| 12.    | Industrial Training                 | Industrial training at the corporate house   |
| 13.    | TrainingExam1Percentage             | Percentage scored in the first training exam                                       |
| 14.    | Exam1Grade                          | Grade secured in the first training exam   |
| 15.    | TrainingExam2Percentage             | Percentage scored in the second training exam                                      |
| 16.    | Exam2Grade                          | Grade secured in the second training exam  |
| 17.    | IntoProduction                      | Whether released to production/delivery  |
| 18.    | Channel                             | Channels such as 'Campus Recruitment', 'Online', 'Referrals', 'Walk-in Interviews' |

2597 prospective candidates, only 1811 join the "Infotech" company in the capacity of an employee and undergo a customized training program. Out of the 1811 employee who underwent the training program, only 1001 employees qualified and were released to begin life on the shop floor.

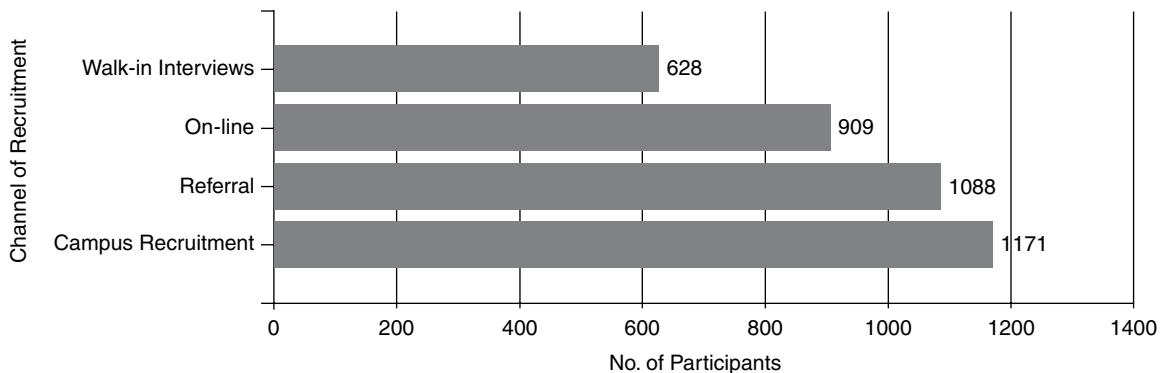


**Figure 9.17** Funnel analysis of employee recruitment by the "Infotech" company.

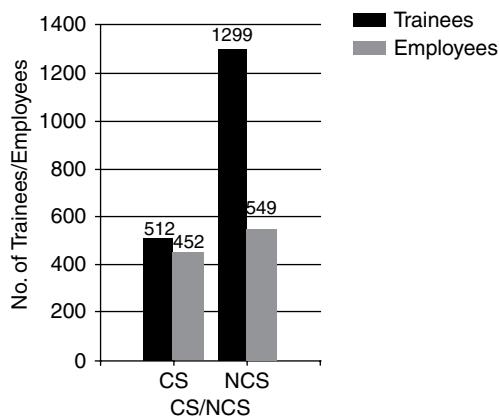
Let us now have a quick look at the distribution channel analysis of employee recruitment by the “Infotech” company. The company gets its recruits from four recruitment channels: “Campus recruitment”, candidates applying “On-line”, “Walk-in interviews”, and “Referrals” (where employees of the company refer candidates). In the channel analysis, depicted in Figure 9.18, it is obvious that “Campus Recruitment” fetches the biggest number. There were 1171 recruits through the “Campus Recruitment” program, 1088 through “Referrals”, 909 applied “On-line”, and only 628 came in through “Walk-in interviews”.

Now for the performance analysis:

As indicated by Figure 9.19, out of 515 CS trainees who underwent the training program, 452 were able to successfully qualify for the shop floor, and only 549 out of 1299 NCS trainees qualified the training program.



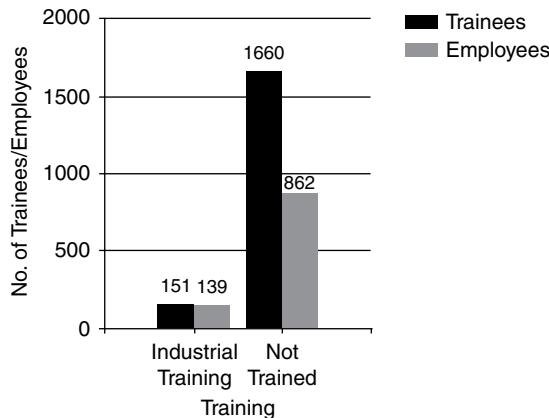
**Figure 9.18** Channel analysis of employee recruitment by the “Infotech” company.



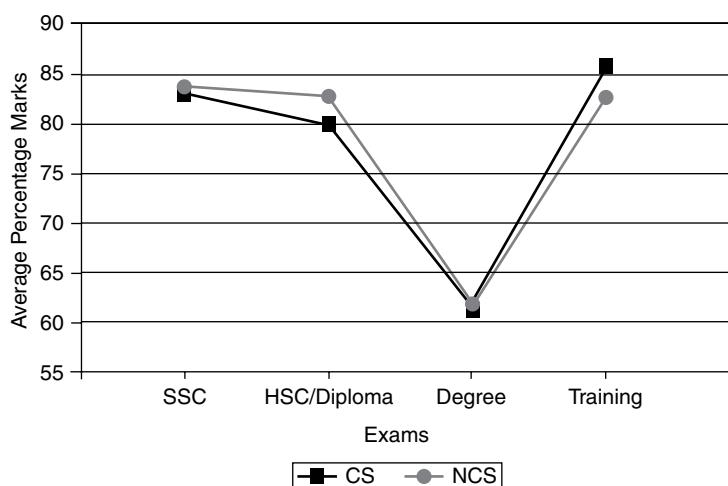
**Figure 9.19** Performance analysis of “Infotech” recruits (CS/NCS) in the training program.

As indicated in Figure 9.20, out of 151 trainees with an “Industrial Training” background, who underwent the training program, 139 were able to successfully qualify for the shop floor. And, of 1660 trainees without an “Industrial Training” background, only 862 qualified the training program.

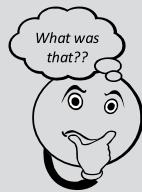
As is evident from Figure 9.21, both the CS and NCS groups show a decline from their SSC to HSC/Diploma with the dip being the highest at the Degree level, only to rise during the training once they join the “Infotech” company.



**Figure 9.20** Another performance analysis of “Infotech” (Industrial training/not trained) recruits in the training program.



**Figure 9.21** Another performance analysis of “Infotech” recruits (CS/NCS).



## Remind Me

- Scorecards measure performance against strategic goals, whereas dashboards present real time information using graphical elements.
- KPIs are designed to let a business user know at a glance whether results are on target or off target.
- An indicator gives visual cues about performance.
- Dashboards mostly use graphs, grids, gauges, and a variety of visualization techniques to highlight the operational data. On the other hand, scorecards commonly use symbols and icons.



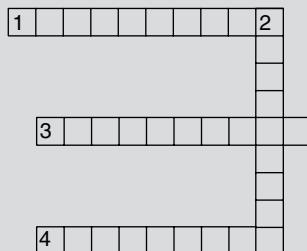
## Connect Me (Internet Resources)

- [www.klipfolio.com/satellite/dashboards-scorecards](http://www.klipfolio.com/satellite/dashboards-scorecards)
- [office.microsoft.com/.../what-is-the-difference-between-a-dashboard-and-a-scorecard-HA101772797.aspx - United States](http://office.microsoft.com/.../what-is-the-difference-between-a-dashboard-and-a-scorecard-HA101772797.aspx)



## BI Crossword

Scorecard vs. Dashboard



### ACROSS

- These are graphical elements that give visual cues about performance.
- Users of balanced scorecards.
- \_\_\_\_\_ is Business Activity Process Monitoring.
- \_\_\_\_\_ is a Business Performance Measurement.

### DOWN

**Solution:**

- |               |               |
|---------------|---------------|
| 1. Indicators | 3. Executives |
| 2. Scorecard  | 4. Dashboard  |

## UNSOLVED EXERCISES

---

1. Describe the functions that you think are there in a typical enterprise.
2. What is a balanced scorecard? Explain.
3. Why do you think companies or business units/functions or individuals should define and maintain a balanced scorecard?
4. Describe the Malcolm Baldrige Performance Excellence Framework.
5. “Texas Nameplate Company (TNC)” has won the Malcolm Baldrige award twice. Read more about the company and describe their focus on quality which led to their winning the award twice.
6. Why is it important for enterprises to go for “enterprise dashboard”? Explain.
7. How is a balanced scorecard different from an enterprise dashboard? Explain.
8. Is it possible to trace the progress on the operational tasks as depicted by the dashboard to the strategic objectives as defined by the balanced scorecard? Explain.
9. Create a balanced scorecard for a fictitious enterprise. Explain the rationale behind it.
10. Who is the balanced scorecard for? Explain your answer.
11. Who is the enterprise dashboard for? Explain your answer.
12. *“The primary difference between the two is that dashboards monitor the performance of operational processes whereas scorecards chart the progress of tactical and strategic goals.”* Explain giving an example.
13. Explain various types of analysis such as “performance analysis”, “channel analysis”, and “funnel analysis”. Give examples in support of your answer.
14. Think of your college/school/university results and cite the different analysis that can be performed on your results.
15. Why is “measurement, analysis, and knowledge management” so important for an enterprise? Give reasons to support your answer.
16. Assume you are the owner of a fast food chain. Give the different ways in which you will promote your fast food outlet. How will you perform the analysis?
17. Why is there so much emphasis on “internal processes” for any enterprise? Give reason to justify your answer.
18. Discuss “balanced scorecard as a strategy map”. Giving example in support of your answer.
19. Are KPIs plotted on a balanced scorecard? Explain.
20. Are KPIs plotted on an enterprise dashboard? Explain.



# 10



## Understanding Statistics

---

### BRIEF CONTENTS

|  |                                      |
|--|--------------------------------------|
| Role of Statistics in Analytics            | Matched Pair Groups in Data Sets     |
| Data, Data Description and Summarization   | Common Statistical Testing Scenarios |
| Getting to Describe Categorical Data       | Understanding Hypothesis and t-Test  |
| Getting to Describe Numerical Data         | Correlation Analysis                 |
| Association between Categorical Variables  | Regression                           |
| Association between Quantitative Variables | ANOVA                                |
| Statistical Tests                          | The F-Test                           |
| Paired and Unpaired Data Sets              | Time Series Analysis                 |

---

### WHAT'S IN STORE?

This chapter focuses on the understanding of Statistics. It will help you understand basic concepts associated with describing data sets and techniques used to visualize data. Further we will look into advanced concepts like Hypothesis and t-Test along with Correlation Analysis, Regression and ANOVA.

---

### 10.1 ROLE OF STATISTICS IN ANALYTICS

We have come across several branches of mathematics like arithmetic, algebra, geometry, and so on. Statistics and probability are also the subjects associated with mathematics that come very handy when we want to understand and analyze “Data”. Let us first study some data-related questions that may or may not need the concepts associated with statistics to answer and then define these subjects. Look at the following pairs of questions:

1. What is your monthly income?
2. Do IT programmer get paid more than the accountant?
3. How fast can your dog run?
4. Do dogs run faster than cats?
5. How much rain did Mumbai receive in December 2015?
6. Does it rain more in Mumbai than Bangalore?
7. What is the probability of rain this Friday?
8. What is the probability of train getting delayed over 10 minutes during weekends?

Some of these are just one single fact that you can recall/find and answer. But the questions numbered 2, 4, 6 have variability; there can be more data points and each of these situations need the application of statistics to answer those questions.

The last two questions need the use of probability concepts to predict the possibility using a large set of data points collected over a period of time.

Now let us look at some of the business-related practical scenarios and questions that could be answered using statistics.

1. You are a doctor trying to develop a cure for Ebola. Currently you are working on a medicine labeled D-X. You have data from patients to whom medicine D-X was given. You want to determine on the basis of those results whether D-X really cures Ebola.
2. You are the quality manager at a mobile phone factory producing over 10000 pieces each day. You observe that last Tuesday's assembly batch reported that there were several phone bodies that were slightly smaller than usual and hence got rejected. You want to find whether anything changed in the manufacturing line and it is an aberration.
3. You are the social media advertising manager at a product company and you have launched several digital campaigns to promote your product to get over 1000 online customers each week. What is the probability of getting such sales?
4. You are the service in-charge of a motorcycle repair shop. Of late, you have seen quite a few customers complain about quality of service. You would like to find how many customers are likely to switch your competitor over the next three months?

The "Numbers" provided by probability will help you make decisions for corrective actions in business. In the above examples, the drug researcher may need to focus on plan B if the drug D-X is not curing sufficient number of patients. The quality manager may need to schedule repair of machines that are contributing to the production of defective parts. The service in-charge may want to talk to dissatisfied customers immediately to prevent churn. Hence statistics and probability have many applications in analyzing business data and supporting decision-making.

Yes, predicting outcome of games, stock movement, etc. are all big time applications of statistics and probability as well.

For a layman, "statistics" means numerical information expressed in quantitative terms. This information may relate to objects, processes, business activities, scientific phenomena, or sports.

We can define statistics as science of collecting large number of facts (or real-world observations) and analyzing with the purpose of summarizing the collection and drawing inferences.

We can define probability as a measure or estimate of the degree of confidence one may have in the occurrence of an event, measured on a scale of impossibility to certainty. It may be defined as the proportion of favorable outcomes to the total number of possibilities.

The term statistics sometimes causes confusion and therefore needs explanation.

A statistic is just a number. There are two kinds of statistics: **(a) summarization or aggregation or descriptive statistics and (b) probability statistics or inferential statistics**. The most important summarization statistics are the total, averages such as the mean and median, the distribution, the range and other measures of variation. Inferential statistics uses descriptive statistics as its input and probability theory to predict outcomes.

**Descriptive statistics:** It allows one to show, describe or present data in a meaningful way such that patterns might emerge from the data. It can only be used to describe the group that is being studied. It cannot be generalized to a larger group. In other words, it is just a way to describe our data. Example:

1. Measures of central tendency such as mean, median and mode.
2. Measures of spread such as range, IQR (inter-quartile range), variance, standard deviation.

**Inferential statistics:** Inferential statistics helps to make predictions or inferences about a population based on the observation or analysis of a sample. However, it will be imperative that the sample be representative.

Inferential statistics can be used for two purposes: to aid scientific understanding by estimating the probability that a statement is true or not, and to aid in making sound decisions by estimating which alternative among a range of possibilities is most desirable.

It is important to note that if the raw data sets are of poor quality, probabilistic and statistical manipulation cannot be very useful. Hence decisions based on such erroneous foundations will be flawed.

There are many tools that are used in the field of statistics and probability such as t-test, Z-test, F-test, Histogram, Rank and Percentile calculation, Sampling, Curve fitting, Correlation, Covariance, Regression, Random number generation, ANOVA and so on.

Now, let us understand more about data and its variety.

## 10.2 DATA, DATA DESCRIPTION AND SUMMARIZATION

When we think about data, we find two fundamental types viz. alphanumeric and numeric. In the following sections we will understand concepts relating to these two types of data and possible ways to describe them.

### 10.2.1 Getting to Describe “Categorical Data”

Let us first recall some of the terms associated with data.

**Data** – It is a collection of facts that have similar attributes or characteristics.

- “Phone number list” is a named collection of, say, mobile phone numbers of your friends.
- “Email IDs list” is an example of collection of email IDs of your classmates.

**Measure** – Data with associated unit of measure (UOM) is typically termed as measure.

- “Service hours per month” has a numeric data associated with “time duration”.
- “Average product shipment time” is a measure derived out of multiple data points.

**Metric** – It is a system of measures based on standard UOM with a business context. The term business metric also refers to the same.

- “Product proliferation rate” by region is an example of measuring “what percentage of products were purchased by customers in different cities belonging to the region”.
- “Employee attrition rate” by quarter measures the percentage of employees leaving the company within each three-month period.

**Pattern** – Pattern in data is a predictable arrangement or feature.

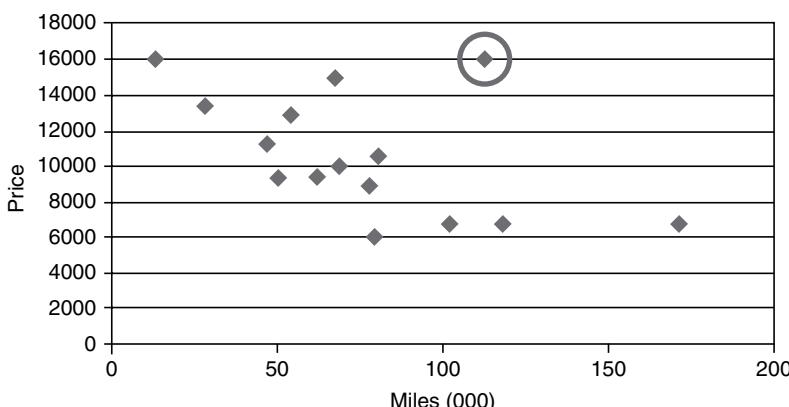
- Consumers who receive coupons buy more electronics than consumers without coupons is a pattern.

A pattern in a statistical model describes variations in data set.

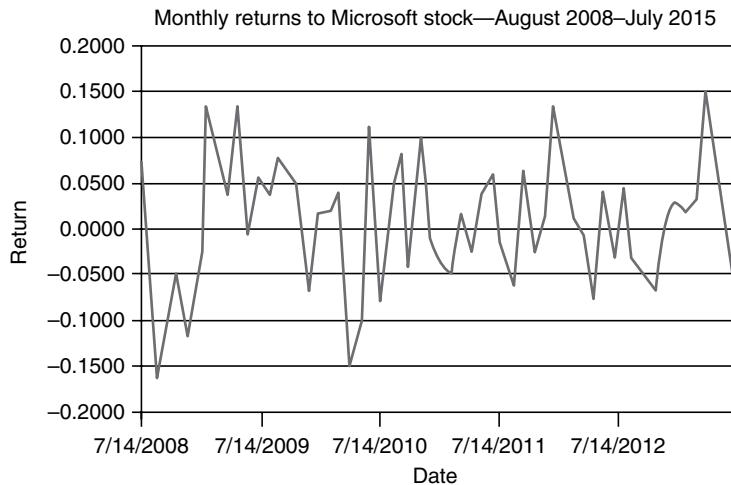
- If you were to tabulate the average price of used motorcycles by gathering data about used vehicles like Make-Model, Year of Manufacture, Mileage and Average price and draw a scatter plot, you can determine the suggested market price for any new entry with a different mileage (Figure 10.1).

We have seen data collected being entered in rows and columns, with column indicating a particular field and row representing each instance of column values. The column entries are called categorical or ordinal variables and represent potential groups of occurrences. Numerical variables describe quantitative attributes of the data like income, height, age, etc. Numerical variables have **unit of measurement**. The data collected for a numerical variable must share a common unit of measurement.

One of the key operations we perform on tables is **aggregation or totaling**. Aggregation generates fewer rows of summary. For example, if you have a table of monthly OLA cab hiring expenses along with type of car like Mini, Sedan and Van, you can sum or aggregate the total amount you have spent by the cab type. Here you will have just three rows for each type of cab with the total money you have spent in the month for that category.



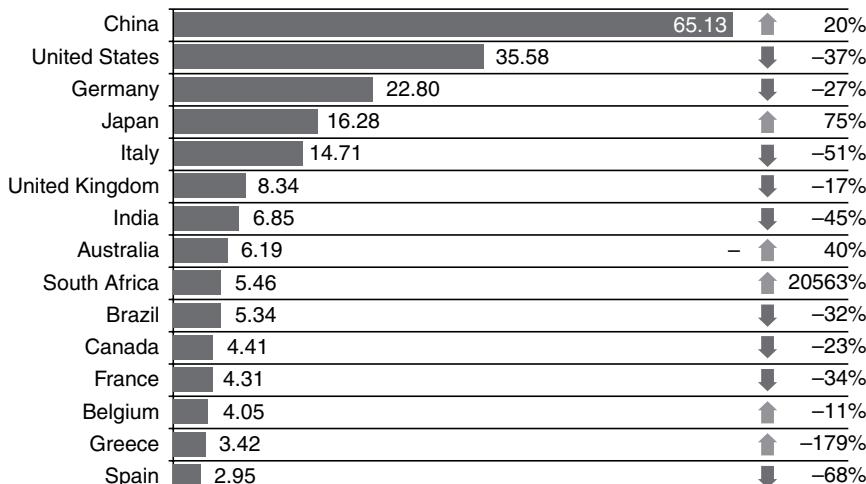
**Figure 10.1** Scatter plot showing the correlation between average price of used motorcycles and their mileage.



**Figure 10.2** Sample time series analysis.

A time series is a sequence of data that records an attribute at different times. The rows of a time series carry different meaning compared to the examples above. Here we record changing value of the same parameter/variable over a regular time interval called frequency (see Figure 10.2).

A bar chart can be used to display the distribution of categorical variables. Here the length of the bar will be proportional to the count of the category. Figure 10.3 shows a bar graph depicting world leading investments in clean energy initiatives. You need to note that the area occupied by each of the bars is proportional to the quantity it is showing. This graph gets cluttered if number of entries are too many. Sometimes pie chart is also used for the same purpose. Recent practices do not recommend use of pie charts. Pie charts should be used when the number of slices can be kept between 3 and 6, otherwise with too many slices it can get extremely difficult to read the pie chart.



**Figure 10.3** Sample bar chart.

Here are some terms that we need to be familiar with:

1. The **Mode** of a category is the most common category in the data set or the category with highest frequency. This will be the longest bar. Sometimes you may have more than one high frequency category and is termed as multi-modal graph.
2. **Median** of the category is the label of the **middle data point** when the values are **sorted**.

### 10.2.2 Getting to Describe “Numerical Data”

While handling a large collection of numerical data values, we use three basic methods to describe the numerical data – percentiles, histograms and box plots.

A **percentile rank** is the percentage of scores that fall below a given score. Median is the 50th percentile, the lower quartile is the 25th percentile and the upper quartile is the 75th percentile. The minimum is the 0th percentile and the maximum is the 100th percentile. The most familiar statistic is the mean or the average. The average squared deviation from the mean is the variance.

To calculate the percentile rank of  $n_2$  in the series  $n_1, n_2, n_3, n_4, n_5, n_6$  use the following formula:

$$\text{Percentile} = (\text{Number of data points below the } n_2 \text{ value}) / (\text{Total number of data points}) \times 100$$

Percentile ranks are not on an equal interval scale.

So far we have seen the different statistics that identify largest, smallest, middle, average and mode (highly repeated values) in a data set. There is another aspect to consider while describing the data set, viz., how far the different data points are spread from the center? Let us examine these concepts.

One measure of spread is the **Range**. The range is simply the difference between the smallest value (minimum) and the largest value (maximum) in the data.

1. Range is used in manufacturing industries for the statistical quality control of manufactured products in large scale like LED bulbs.
2. Range is useful in studying the variations in the prices of mutual funds, shares that are sensitive to price changes that fluctuate from one period to another.

The **Inter Quartile Range** (IQR) gives information about how the middle 50% of the data are spread. The interquartile range is the difference between the Q3 and Q1. Hence,  $\text{IQR} = \text{Q3} - \text{Q1}$ .

The difference between a data value and the mean is called the **deviation**. One way to measure how the data are spread is to look at how far away each of the values is from the mean. This could be positive or negative value.

The **standard deviation** is a measure of the average deviation for all of the data points from the mean. As the sum of mean always adds to zero, we will need to use the squared deviations to make each value positive. Standard deviation is a statistic used as a measure of the dispersion or variation in a distribution, equal to the square root of the arithmetic mean of the squares of the deviations from the arithmetic mean. **Variance** is calculated by taking the differences between each number in the set and the mean, squaring the differences (to make them positive) and dividing the sum of the squares by the number of values in the set. Hence, the variance is a numerical value used to indicate how widely individuals in a group vary. If individual observations vary greatly from the group mean, the variance is big; and vice versa.

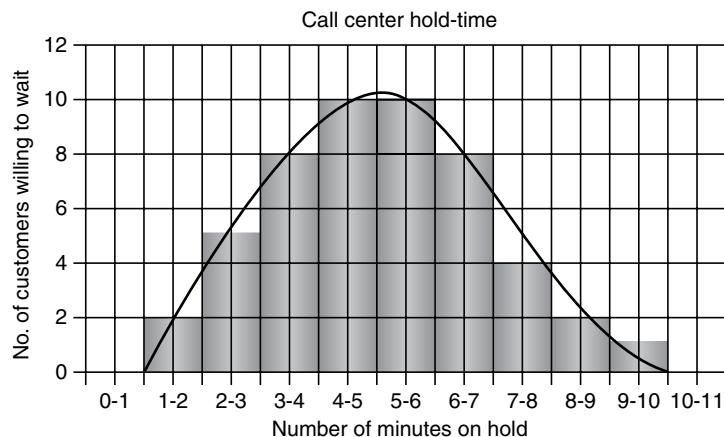
A **histogram** is a graph that shows the counts in a data set as the heights of the bar that are proportional to the areas of the bar distributed on a chosen interval scale. The interval will accommodate all

possible values of the numerical variable (Figure 10.4). An **outlier** is an observation that lies an abnormal distance from other values in a random sample from a population.

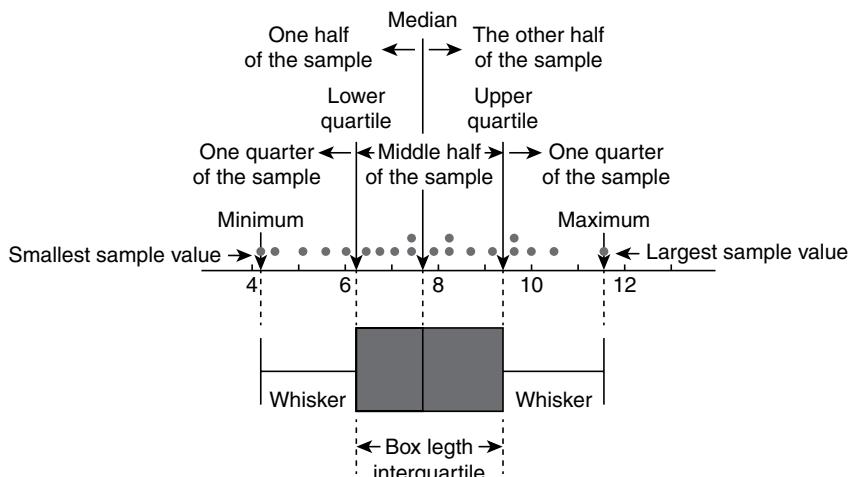
A **Boxplot** is a way of graphically summarizing groups of numerical data through their quartiles. Boxplots may also have lines extending vertically from the boxes (called whiskers) indicating variability outside the upper and lower quartiles (Figure 10.5).

The extremes at the right and left of the histogram where the bars become short are termed as the tails of the distribution. If one tail stretches out farther than the other, the distribution is Skewed (Figure 10.6).

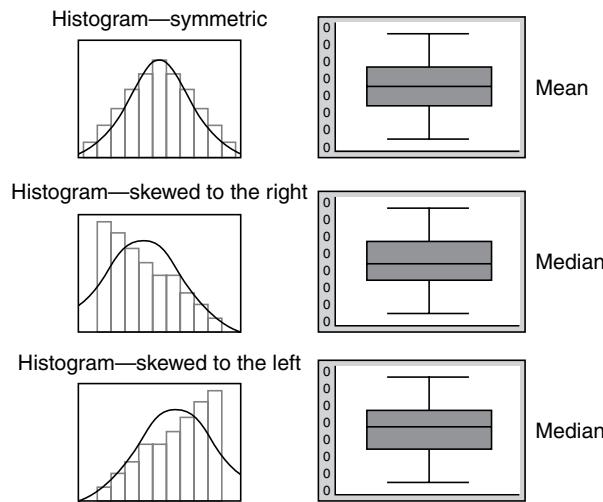
We may not have the opportunity to analyze the entire set of occurrences all the time. For example, getting data about all the people in India in the age group of 45–50 years who need vision correction.



**Figure 10.4** Sample histogram.



**Figure 10.5** Sample Box and Whisker Plot.



**Figure 10.6** Boxplot as a replacement for histogram.

Hence we use the following terminology to describe the boundaries of the data set we are using for data analysis:

1. **Population:** In every statistical analysis we aim to analyze information about some group of individuals or things. In statistical language, such a collection is called a population or universe. For example, we have the population of all cars manufactured by a company in the last 10 years. A population could be finite or infinite, depending on whether the number of elements is finite or infinite. In most situations, the population may be considered infinitely large.
2. **Sample:** A finite subset of population is called a sample. The definition of a sample is a small part of a large data set used to represent the whole or to learn something about the whole. An example of a sample is a small piece of chocolate offered free at a store to get you to buy a box of newly launched chocolate. Another example of a sample is a small quantity of blood that is taken to test in a lab.
3. **Sampling:** Sampling is concerned with the selection of a subset of items from within a statistical population to estimate characteristics of the whole population.

### 10.2.3 Association between Categorical Variables

**Contingency table** provides information about possible relationship between categorical variables. Used with Chi squared test and concepts of probability (discussed in subsequent sections), contingency tables are applied in various data analysis situations. For example,

*In order to discover whether online buyers subscribe to the retailer's mailing list, a question was posted on the website and responses collected. The results are shown in the following table:*

|              |            | <i>Join Mailing List</i> | <i>Decline to Join</i> | <i>Total</i> |
|--------------|------------|--------------------------|------------------------|--------------|
| <i>Buy</i>   | <b>YES</b> | 52                       | 12                     | <b>64</b>    |
|              | <b>NO</b>  | 343                      | 3720                   | <b>4063</b>  |
| <b>TOTAL</b> |            | <b>395</b>               | <b>3732</b>            | <b>4127</b>  |

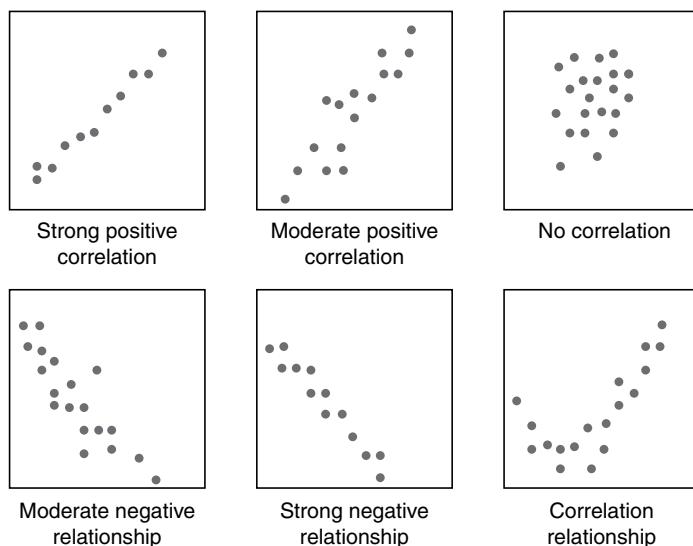
|              |            | <i>Join Mailing List</i> | <i>Decline to Join</i>   | <i>Total</i>           |
|--------------|------------|--------------------------|--------------------------|------------------------|
| <b>Buy</b>   | <b>YES</b> | 52<br>(81.25%)           | 12<br>(18.75%)           | <b>64<br/>(100%)</b>   |
|              | <b>NO</b>  | 343<br>(8.44%)           | 3720<br>(91.56%)         | <b>4063<br/>(100%)</b> |
| <b>TOTAL</b> |            | <b>395<br/>(9.57%)</b>   | <b>3732<br/>(90.42%)</b> | <b>4127</b>            |

A **contingency table** is a matrix that displays the frequency distribution of the categorical variables. They are heavily used in survey research, engineering and scientific research. The term “row percents” describes **conditional contingency** that gives the percentages out of each row total that fall in the various column categories. Similarly, column based contingencies are also computed; this provides a better picture for decision makers.

Two categorical variables are related in the sample if at least two rows **noticeably differ** in the pattern of row percentages. This is the same as saying that two categorical variables are related in the sample if at least two columns noticeably differ in the pattern of columns percentages.

#### 10.2.4 Association between Quantitative Variables

In order to study or investigate the possible influence of two numerical variables on each other we use scatter plots. Scatter plots show how much one variable affects another. The relationship between two variables is called their correlation. For example, we may want to understand if the consumption of



**Figure 10.7** Sample Scatter plots depicting correlation between variables.

electricity in household is related to day temperature? In such situations we use scatter plots. This idea is often times used to explore data for market segmentation. This is also the topic of correlation analysis.

To describe the association, start with the **direction**. In this example, colder the winter, the larger will be the electricity consumption. This pattern has **positive direction** because the data points tend to concentrate in the lower left and upper right corners. Another property of the association is the **curvature**. When the pattern appears like line, it will be **linear** and if the curve bends the association it will be **non-linear**. The third property of association is the **variation around the pattern**. Finally, the outliers in terms of their numbers and position have an influence on the line that represents the data pattern.

**Covariance** quantifies the strength of the linear association between two numerical variables. It measures the degree to which data points are concentrated along an imaginary diagonal line in the scatter plot.

**Correlation** is a more easily interpreted measure of linear association derived from covariance. Correlation does not have any units and it can reach  $-1.0$  or  $+1.0$  but these extremes are unusual.

The **Pearson correlation coefficient ( $r$ )** is a very helpful statistical formula that measures the strength between variables and relationships. In the field of statistics, this formula is often referred to as the Pearson R-test. Here is the formula to compute  $r$  value and the suggested way to interpret the result:

$$r = \frac{N \sum xy - (\sum x)(\sum y)}{\sqrt{[N \sum x^2 - (\sum x)^2][N \sum y^2 - (\sum y)^2]}}$$

where  $N$  is number of pairs of scores;  $\sum xy$  is the sum of the products of paired scores;  $\sum x$  is sum of  $x$  scores;  $\sum y$  is sum of  $y$  scores;  $\sum x^2$  is sum of squared  $x$  scores and  $\sum y^2$  is sum of squared  $y$  scores.

$r$  value = \_\_\_\_\_

|                |                                   |
|----------------|-----------------------------------|
| +.70 or higher | Very strong positive relationship |
| +.40 to +.69   | Strong positive relationship      |
| +.30 to +.39   | Moderate positive relationship    |
| +.20 to +.29   | Weak positive relationship        |
| +. 10 to +.19  | No or negligible relationship     |
| 0              | No relationship                   |

However, if a scatter plot shows a linear pattern, and the data are found to have a strong correlation, it does not necessarily mean that a cause-and-effect relationship exists between the two variables. A cause-and-effect relationship is one where a change in X causes a change in Y.

## 10.3 STATISTICAL TESTS

---

In the field of statistics it is usually impossible to collect data from all individuals of interest, that is, the Population. The standard approach is to collect data from a subset or sample of the population, but our real desire is to know the “truth” about the population. Quantities such as means, standard deviations and proportions are important values and are called “Parameters” when we are talking about a population. Since we usually cannot get data for the whole population, we cannot know the values of the parameters for that population. We can, however, calculate estimates of these quantities for our sample. When they are calculated from sample data, these quantities are called “statistics”. A statistic estimates a parameter. The field of statistics focused on statistical procedures or tests is called **parametric tests**.

1. **Parametric Statistical tests** are based on assumptions that observations are independent, the sample data have a normal distribution and data values in different groups have homogeneous variances.
2. **Non-Parametric Statistical tests** focus on statistical methods wherein the data is not required to fit a normal distribution. It uses data that is often ordinal, meaning it does not rely on numbers, but rather a ranking or order of sorts.

### 10.3.1 Paired and Unpaired Data Sets

Two data sets are “paired” when the following one-to-one relationship exists between values in the two data sets:

1. Each data set has the same number of data points.
2. Each data point in one data set is related to one, and only one, data point in the other data set.

An example of paired data would be a before–after treatment test. The researcher might record the medical parameters of each subject in the study, before and after a treatment is administered. These measurements would be paired data, since each “before” measure is related only to the “after” measure from the same subject.

This data is described as **unpaired or independent** when the sets of data arise from separate individuals or **paired** when it arises from the same individual at different points in time. For example, one clinical trial might involve measuring the blood pressure from one group of patients who were given a medicine and the blood pressure from another group not given the medicine. This would be unpaired data. Another clinical trial might record the blood pressure in the same group of patients before and after giving the medicine. In this case the data is “paired” as it is likely the blood pressure after giving the medicine will be related to the blood pressure of that patient before the medicine was given.

### 10.3.2 Matched Pair Groups in Data Sets

Many times we may need to design samples that use several pairs of subjects that have common attributes or profile. For example, let us say you want to compare the face-to-face and virtual learning effectiveness in a class. If one chooses different classes taught by different experts then there could be variations in lecture style, teaching approach, assessment methods and so on. On the other hand, if we pick a class taught by the same expert and pair the students with similar characteristics like learning style, gender and typical performance grade and then in each pair randomly assign them to take face-to-face and virtual sessions, we will be able to compare the performance more accurately. Hence the

findings of matched pair will be robust. Thus we can define a matched pairs design as a special case of a randomized block design. It can be used when the experiment has only two treatment conditions; and subjects can be grouped into pairs, based on some blocking variable. Then, within each pair, subjects are randomly assigned to different treatments.

A **proportion** refers to the fraction of the total that possesses a certain attribute. For example, suppose we have a sample of four pets – a bird, a fish, a dog, and a cat. We might ask what proportion has four legs? Only two pets (the dog and the cat) have four legs. Therefore, the proportion of pets with four legs is  $2/4$  or 0.50.

With this background we can learn about the common statistical tests.

Selection of appropriate statistical test is very important for analysis of gathered data and its type. Selection of appropriate statistical tests depends on the following two things:

1. What kind of data are we dealing with?
2. Whether our data follows normal distribution or not?

*What kind of data are we dealing with?*

Most often the collected data fall in one out of the following four types of data, that is, *nominal data*, *ordinal data*, *interval data*, and *ratio data*.

1. **Nominal data** is the collection of facts against a single name/categorical entity, for example, say Salary of Managers. Nominal data cannot be ordered or measured but can ONLY be counted. Data that consist of only two classes like male/female or owned/rented are called *binomial data*. Those that consist of more than two classes like tablet/capsule/syrup are known as *multinomial data*. Data of these types are usually presented in the form of contingency tables.
2. **Ordinal data** is also a type of categorical data but in this, categories are ordered logically. These data can be ranked in order of magnitude. One can say definitely that one measurement is equal to, less than, or greater than another. For example, data on average family spending in different income groups of India such as lower middle class income group, middle class income group, higher middle class income group etc.
3. **Interval data** has a meaningful order and also has the quality that equal intervals between measurements represent equal changes in the quantity of whatever is being measured. For example, room temperature: Freezing, 5–6°C; cool, 16–22°C; warm, 24–28°C and hot, >29°C. There is nothing called zero in range type of data.
4. **Ratio data** has all the qualities of interval data plus a natural zero point. For example, ratio of heights, lengths, etc.

*Whether our data follow the normal distribution or not?*

1. The data collected may follow normal distribution or different distribution pattern.
2. There are various methods for checking the normal distribution of data including plotting histogram, plotting box and whisker plot, plotting Q–Q plot or measuring skewness and kurtosis.

### 10.3.3 Common Statistical Testing Scenarios

While there could be several forms of data like nominal, ordinal, ratio, and interval data in the samples collected relating to different experiments, let us first learn about most commonly occurring data forms and methods of statistical analysis associated with these data forms. The following are some of the

common goals such as **description, comparison of two or more groups, measuring association or prediction** for using statistical tests.

1. Description of one group of observations with nominal data.
2. Description of one group of observations with ordinal data.
3. Comparison of a group with nominal data with a hypothetical value.
4. Comparison of a group with ordinal data with a hypothetical value.
5. Comparison of two unpaired groups with nominal data.
6. Comparison of two unpaired groups with ordinal data.
7. Comparison of two paired groups with ordinal data.
8. Comparison of two paired groups with nominal data.
9. Comparison of three or more unmatched groups with nominal data.
10. Comparison of three or more unmatched groups of equal or different sample sizes with ordinal data.
11. Comparison of three or more matched groups of equal or different sample sizes with ordinal data.
12. Comparison of three or more unmatched groups of equal or different sample sizes with ratio/interval data with normal distribution.
13. Measuring association between two variables with nominal data.
14. Measuring association between two variables with ordinal data.
15. Prediction from another measured variable with nominal data.
16. Prediction from another measured variable with ratio or interval data with normal distribution.
17. Prediction with several measured or binomial variables.

## 10.4 UNDERSTANDING HYPOTHESIS AND T-TEST

Hypothesis is a tentative explanation based on observations you have made. Example: Students in North India spend more on movies than students in South India OR adding fertilizer to a plant makes it grow better.

The basic logic of hypothesis testing is to prove or disprove the statistical research question such as the examples above. By allowing an error of 5% or 1% (termed as alpha values of 0.05 or 0.01) and making correct decisions based on statistical principles, the researcher can conclude that the result must be real if chance alone could produce the same result only 5% or 1% of the time or less.

The following are the major types of statistical hypotheses:

1. **H<sub>0</sub>: Null Hypothesis** – It is usually the hypothesis that sample observations result purely based on chance. A hypothesis that attempts to nullify the difference between two sample means (by suggesting that the difference is of no statistical significance) is called a null hypothesis.
2. **H<sub>1</sub>: Alternative Hypothesis** – It is the hypothesis that sample observations are influenced by some non-random cause.

**Null hypothesis** is a more formal statement of your original hypothesis. It is usually written in the following form: There is no significant difference between population A and population B. The reason we write it in this form is to prove a hypothesis false. In fact, you can never really prove that a hypothesis is true.

*Example:* There is no significant difference in spending for movies in North India vs. South India OR There is no significant difference in the growth of fertilized plants vs. unfertilized plants.

#### 10.4.1 The t-Test

We use this statistical test to compare our sample groups (A and B) and determine if there is a significant difference between their means. The result of the t-test is a ‘t-value’; this value is then used to determine the p-value demonstrating probability of hypothesis being true or false in the entire population.

#### 10.4.2 The p-Value

It is the probability that ‘t-value’ falls into a certain range. In other words, this is the value you use to determine if the difference between the means in your sample populations is significant. In general, a p-value  $< 0.05$  suggests a significant difference between the means of our sample population and we would reject our null hypothesis. A p-value  $> 0.05$  suggests no significant difference between the means of our sample populations and we would not reject our null hypothesis.

Unpaired t-test is used when you have independent samples. Paired t-test is used when your samples are related. For example, you collected data (pulse rate) of your subjects before and after they had 3 cups of coffee.

#### 10.4.3 Z-Test

A Z-test is a **hypothesis test** based on the Z-statistic, which follows the standard normal distribution under the null hypothesis. The simplest Z-test is the 1-sample Z-test, which tests the mean of a normally distributed population with known variance. For example, the manager of a Gems Candy wants to know whether the mean weight of a batch of candy packs is equal to the target value of 100 gm. From historical data, they know that the filling machine has a standard deviation of 5 gm, so they should use this value as the population standard deviation in a 1-sample Z-test.

**Z-test for single proportion** is used to test a hypothesis on a specific value of the population proportion. Statistically speaking, we test the null hypothesis  $H_0: p = p_0$  against the alternate hypothesis  $H_1: p >< p_0$ , where  $p$  is the population proportion and  $p_0$  is a specific value of the population proportion we would like to test for acceptance. Example: If you want to prove that tea and coffee are equally popular in the college campus you require this test. In this example,  $p_0 = 0.5$ . Notice that in this particular example, proportion refers to the proportion of tea drinkers.

**Z-test for difference of proportions** is used to test the hypothesis that two populations have the same proportion. Example: Suppose one is interested to test if there is any significant difference in the habit of tea drinking between male and female students in the college campus. In such a situation, Z-test for difference of proportions can be applied. One would have to obtain two independent samples from the college campus – one from males and the other from females and determine the proportion of tea drinkers in each sample in order to perform this test. You must know the standard deviation of the population and your sample size must be above 30 in order for you to be able to use the Z-score. Otherwise, use the t-score.

Before we move further to examine statistical tests associated with more than two groups of data, let us understand the two key concepts of correlation and linear regression in little more detail.

## 10.5 CORRELATION ANALYSIS

Correlation analysis measures the direction and strength of the relationship between two variables. Correlation can predict or calculate the value of one variable from the given value of the other variable. Thus, correlation is a measure of the degree to which two variables are related.

A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. For instance, if  $x$  and  $y$  are two variables, then correlation would be a linear association between them (a straight line graph) and would help determine the relationship between the two. The correlation coefficient lies in the range of  $-1.00$  to  $+1.00$  as a positive or negative probability that the members of a data pair relate to each other. It gives an estimate of the degree of association between two or more variables. Correlation analysis also tests the interdependence of the variables.

Some of the types of correlations we could think of include:

1. Positive correlation.
2. Negative correlation.
3. Simple correlation.
4. Multiple correlation.
5. Partial correlation.
6. Total correlation.
7. Linear correlation.
8. Non-linear correlation.

*Positive and negative correlation* depend on the direction of change of the variables.

1. If two variables tend to move together in the same direction, that is, an increase in the value of one variable is accompanied by an increase in the value of the other variable or a decrease in the value of one variable is accompanied by a decrease in the value of other variable, then the correlation is called *direct or positive correlation*. Some of the examples could be product price and supply, exercise and weight change, etc.
2. If two variables tend to move together in opposite directions so that decrease in the value of one variable is accompanied by the increase in the value of the other or vice-versa, then the correlation is said to be inverse or negative correlation, for example, product price and demand.

The study of the relationship of two variables is called *simple correlation*. However, in a *multiple correlation*, researchers study more than two variables simultaneously. An example of multiple correlations would be the relationship between demand, supply of commodity and price. The study of two variables excluding one or more variables is called *partial correlation*. In *total correlation*, all the variables are taken into account. If the ratio of change between two variables is uniform, that is, the value of interval of 2 data series is constant, then there exists a *linear correlation* between them.

To examine whether two random variables are interrelated, you should plot a scatter diagram.

## 10.6 REGRESSION

**Regression To Mean** (RTM) is a statistical phenomenon that occurs when repeated measurements are made on the same subject. It happens because values are observed with random error. By random error we mean a non-systematic variation in the observed values around a true mean. Systematic error, where the observed values are consistently biased, is not the cause of RTM.

Sir Francis Galton introduced the word “regression” in 1877 when studying the relationship between the heights of fathers and sons. He studied over 100 such pairs and expressed the opinion that short fathers had short sons while tall fathers had tall sons. He also found that the average height of the sons of tall fathers was less than the average height of the tall fathers. Similarly, he also found that the average height of the sons of short fathers was more than the average height of the short fathers. In general, when observing repeated measurements in the same subject, relatively high (or relatively low) observations are likely to be followed by less extreme ones nearer the subject's true mean. Galton referred to the tendency to regression as the “Line of Regression”. The line describing the average relationship between two variables is known as the line of regression. Regression analysis, when used for studying more than two or three variables at a time, is called as *multiple regression*.

The primary objective of regression analysis is to provide estimates of the values of the dependent variables from independent variables. A simple linear regression allows you to determine functional dependency between two sets of numbers. For example, we can use regression to determine the relation between cool drink sales and average outside temperature. Since we are talking about functional dependency between two sets of variables, we need an independent variable and one dependent variable. In this example, if change in temperature leads to change in cool drinks sales, then temperature is an independent variable and cool drink sales is a dependent variable. Prediction or estimation is very important in business and science. Using this statistical tool, you can predict the unknown values. Regression analysis is used in all the fields of statistics, where two or more relative variables have the tendency to go back to the average. It is used to estimate the relationship between two economic variables like income and expenditure. For example, if you know the income, you can predict the probable expenditure.

In regression analysis, given a bunch of points, we find a line that “fits” them the best. For any line you try, each point has a distance to that line. This is known as your “**error**”, since the further the point is from the line, the less good your line is at fitting that point. If you add up those errors, you have the total error. You are trying to find the line that makes that error as small as possible.

| Correlation   | Regression   |
|---|--|
| It is the relationship between two or more variables and varies with the other in the same or the opposite direction. | It is a mathematical measure of viewing the average relationship between two variables.  |
| Correlation identifies the degree of relationship between two variables.  | Regression identifies the cause and effect relationship between the variables.   |
| The coefficient of correlation is a relative measure and the range of relationship lies between $-1$ and $+1$ .       | The regression coefficient is an absolute figure. It helps to find the value of the dependent variable if we know the value of the independent variable. |
| If the coefficient of correlation is positive, then the two variables are positive correlated and vice versa.         | Regression indicates that decrease in one variable is associated with increase in the other variable.  |

It turns out that if instead of adding up the errors, you add up the “squared errors”, the math becomes really more accurate, and given any set of points you can just figure out what that line should be.

## 10.7 ANOVA

---

ANOVA (Analysis of Variance) analysis method was developed by Ronald Fisher in 1918 and is the extension of the t-test and the Z-test. When you have more than two groups to compare, for example in a drugs trial when you have a high dose, low dose, and a placebo group (so 3 groups), you use ANOVA to examine whether there are any differences between the groups.

The one-way analysis of variance (ANOVA) is used to determine whether there are any significant differences between the means of three or more independent (unrelated) groups. Specifically, it tests the null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

where  $\mu$  = group mean and  $k$  = number of groups. If, however, the one-way ANOVA returns a significant result, we accept the alternative hypothesis (HA), which is that there are at least two group means that are significantly different from each other.

ANOVA is based on comparing the variance between the data samples to variation within each particular sample. If the “between variation” is much larger than the “within variation”, the means of different samples will not be equal. If the “between and within” variations are approximately the same size, then there will be no significant difference between sample means.

The following are some of the popular types of ANOVA:

1. One-way between groups.
2. One-way repeated measures.
3. Two-way between groups.
4. Two-way repeated measures.

Sometimes, we will need to understand inter-relationships that have influence and sub-categories. Let us suppose that the HR department of a company desires to know if occupational stress varies according to age and gender. The variable of interest is therefore occupational stress as measured by a scale. There are two factors being studied – age and gender. Further suppose that the employees have been classified into three groups or levels: Age less than 40 years, age between 40 and 55 years and above 55 years. Factor age has three levels and gender two. In such situations we need to use two-way ANOVA for testing the hypothesis.

## 10.8 THE F-TEST

---

F-test is similar to t-test but useful to compare multiple groups and determine if a group of variables is jointly significant. The F-distribution is named after the famous statistician R. A. Fisher. F is the ratio of two variances. The F-distribution is most commonly used in Analysis of Variance (ANOVA) and the F-test (to determine if two variances are equal). The F-distribution is the ratio of two chi-square distributions, and hence is right skewed. It has a minimum of 0, but no maximum value (all values are positive). The peak of the distribution is not far from 0.

In summary, several specialized approaches have been designed to study the relationship among two or more numerical variables and fine-tuned by several renowned statisticians.

Table 10.1 provides the statistical testing approaches (described earlier) and methods.

**Table 10.1** Goal of statistical testing along with suggested statistical test to accomplish the goal

| <i>Goal of Statistical Testing</i>  | <i>Suggested Statistical Test</i>   |
|---|---|
| Description of one group with nominal data  | Proportion  |
| Description of one group with ordinal data  | Median, Interquartile range   |
| Comparison of a group with nominal data with a hypothetical value   | Chi-Square test or Binomial test  |
| Comparison of a group with ordinal data with a hypothetical value   | Wilcoxon test – can be used when comparing two related samples (matched samples, or repeated measurements on a single sample) to assess whether their population mean ranks differ.   |
| Comparison of two unpaired groups with nominal data   | Chi-Square test   |
| Comparison of two unpaired groups with ordinal data   | Mann–Whitney test – The test involves the calculation of a statistic called U, whose distribution under the null hypothesis is known.   |
| Comparison of two paired groups with ordinal data   | Wilcoxon test   |
| Comparison of two paired groups with nominal data   | McNemar's test – This test is applied to $2 \times 2$ contingency tables to determine whether the row and column marginal frequencies are equal.  |
| Comparison of three or more unmatched groups with nominal data  | Chi-Square test   |
| Comparison of three or more unmatched groups of equal or different sample sizes with ordinal data                                 | Kruskal–Wallis test – A significant Kruskal–Wallis test indicates that at least one sample dominates the other sample.  |
| Comparison of three or more matched groups of equal or different sample sizes with ordinal data                                   | Friedman test – It is used to detect differences in treatments across multiple test attempts. The procedure involves ranking each row (or block) together, then considering the values of ranks by columns.   |
| Comparison of three or more unmatched groups of equal or different sample sizes with ratio/interval data with normal distribution | One-way ANOVA   |
| Measuring association between two variables with nominal data   | Contingency coefficients  |
| Measuring association between two variables with ordinal data   | Spearman correlation – It assesses how well the relationship is between two variables that are strictly increasing or decreasing values (i.e., monotonic function). If there are no repeated data values, a perfect Spearman correlation of +1 or -1 occurs when each of the variables is a perfect monotone function of the other. |
| Prediction from another measured variable with nominal data   | Logistic regression   |
| Prediction from another measured variable with ratio or interval data with normal distribution                                    | Linear regression   |
| Prediction with several measured or binomial variables  | Multiple logistic regression  |

Microsoft Excel 2013 provides functions to perform the analysis discussed so far and more. Interested learners may want to explore these to get deeper understanding of the statistical testing concepts and its applications in real-life scenarios.

Till now, this chapter has given you quick overview of the power of statistic in understanding nature of data distribution, their association, comparison and prediction. You may study any of the methods more deeply by referring to specific approach of statistical testing references.

## 10.9 TIME SERIES ANALYSIS

---

A time series is a sequence of observations that are arranged according to the time of their occurrence. A univariate time series is a sequence of measurements of the same variable collected over time. Most often, the measurements are made at regular time intervals. The annual yield of wheat and their price per ton, for example, is recorded in agriculture. We have seen daily reports of stock prices, weekly bullion rates, and monthly rates of industry unemployment. Meteorology department records wind velocity, daily maximum and minimum temperatures, and annual rainfall. Seismographs continuously record earthquakes. ECG and EEG record series of waves of human being for study of potential health ailments. Governments record and analyze births, deaths, and entry into school, dropouts, and many other facts. Manufacturing industries record defects in batches and analyze data for quality improvement. An epidemiologist might be interested in the number of typhoid fever cases observed over some time period. In medicine, blood sugar measurements traced over time could be useful for evaluating drugs used in treating diabetes.

From these situations we can conclude that there must be good reasons to record and to analyze the data of a time series. Among these is the wish to gain a better understanding of the data generating mechanism, the prediction of future values, or the optimal control of a system. The characteristic property of a time series is the fact that the data are not generated independently, their dispersion varies in time, they are often governed by a **trend and they have cyclic components**. The analysis of data that have been observed at different points in time leads to new and unique problems in statistical modeling and inference. A seasonal effect is a systematic and calendar-related effect. Examples include the decrease in retail price of white goods, which occurs around December in response to the Christmas period or an increase in water consumption in summer due to warmer weather. Seasonal adjustment is the process of estimating and then removing from a time series influences that are systematic and calendar related.

The basic objective of time series analysis is to determine a model that describes the pattern of the time series. Benefits of such a model are:

1. To describe the important features of the time series pattern.
2. To explain how the past affects the future or how two time series interact.
3. To forecast future values of the series.

The time domain analysis approach focuses on modeling some **future value** of a time series as a parametric function of the current and past values. In this scenario, we begin with linear regressions of the present value of a time series on its own past values and on the past values of other series. A newer approach to the same problem uses **additive models**. In this method, the observed data are assumed to result from sums of series, each with a specified time series structure; for example, a series is generated

as the sum of trend, a seasonal effect, and error. The steps involved in time series analysis could be summarized as *description, modeling and prediction*.

The frequency domain approach assumes that the primary characteristics of interest in time series analyses relate to periodic or systematic sinusoidal variations found naturally in most data. These periodic variations are often caused by biological, physical, or environmental phenomena such as global warming due to El Nino effect.

In spectral analysis, the partition of the various kinds of periodic variation in a time series is accomplished by evaluating separately the variance associated with each periodicity of interest. This variance profile over frequency is called the power spectrum.

There are two basic types of “time domain” models:

1. Ordinary regression models that use time indices as  $x$ -axis variables. These can be helpful for an initial description of the data and form the basis.
2. Models that relate the present value of a series to past values and past prediction errors – these are called ARIMA models (for Autoregressive Integrated Moving Average) of several simple forecasting methods.

R programming language is widely used for time series analysis and interested learners could explore further about time series analysis and implementation using R.



### Remind Me

- The data set is described as unpaired or independent when the sets of data arise from separate individuals or paired when it arises from the same individual at different points in time.
- A proportion refers to the fraction of the total that possesses a certain attribute.
- Nominal data cannot be ordered or measured but can ONLY be counted.
- Ordinal data can be ranked in order of magnitude.
- Interval data has a meaningful order and also has the quality that equal intervals between measurements represent equal changes in the

- quantity of whatever is being measured.
- Ratio data has all the qualities of interval data plus a natural zero point.
- Correlation analysis measures the direction and strength of the relationship between two variables.
- A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another.
- The study of the relationship of two variables is called simple correlation.
- Regression To Mean (RTM) is a statistical phenomenon that occurs when repeated measurements are made on the same subject.

- ANOVA is based on comparing the variance between the data samples to variation within each particular sample.
- A time series is a sequence of observations that are arranged according to the time of their occurrence.
- Descriptive statistics allows one to show, describe or present data in a meaningful way such that patterns might emerge from the data.
- Inferential Statistics help to make predictions or inferences about a population based on the observation or analysis of a sample.
- Variance is calculated by taking the differences between each number in the set and the mean, squaring the differences (to make them positive) and dividing the sum of the squares by the number of values in the set.
- Standard Deviation is a statistic used as a measure of the dispersion or variation in a distribution, equal to the square root of the arithmetic mean of the squares of the deviations from the arithmetic mean.
- When the pattern appears like line, it will be linear and if the curve bends the association it will be non-linear.



### *Point Me (Books)*

- First Course in Probability by Sheldon Ross
- Discovering Statistics using R by Andy Field
- An Introduction to Probability Theory and Its Applications by William Feller

- A course in Probability Theory by Kai Lai Chung



### *Connect Me (Internet Resources)*

- <https://www.coursera.org/course/stats1>
- <http://online.stanford.edu/course/statistical-learning-winter-2014>
- <http://www.springer.com/us/book/9781461443421?token=prtst0416p>



## Test Me Exercises

### Fill in the blanks

- (a) A \_\_\_\_\_ is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another.
- (b) \_\_\_\_\_ quantifies the strength of the linear association between two numerical variables.
- (c) In order to study or investigate the possible influence of two numerical variables on each other we use \_\_\_\_\_.
- (d) The interquartile range is the difference between the \_\_\_\_\_ and \_\_\_\_\_.
- (e) A \_\_\_\_\_ time series is a sequence of measurements of the same variable collected over time.
- (f) \_\_\_\_\_ programming language is widely used for time series analysis.
- (g) The study of two variables excluding one or more variables is called \_\_\_\_\_ correlation.
- (h) \_\_\_\_\_ is a system of measures based on standard UOM with a business context.

- (i) \_\_\_\_\_ in data is a predictable arrangement or feature.
- (j) \_\_\_\_\_ provides information about possible relationship between categorical variables.
- (k) ANOVA (Analysis of Variance) analysis method was developed by Ronald Fisher in 1918 and is the extension of the \_\_\_\_\_ and the \_\_\_\_\_.

### Solution:

- (a) Correlation Coefficient
- (b) Covariance
- (c) Scatter plot
- (d) Q3 and Q1
- (e) Univariate
- (f) R
- (g) Partial
- (h) Metric
- (i) Pattern
- (j) Contingency table
- (k) t-test, Z-test

# 11



## Application of Analytics

---

### BRIEF CONTENTS

|   |   |
|---|---|
| Application of Analytics                | Analytics in Retail   |
| Analytics in Business Support Functions | Analytics in Healthcare (Hospitals or Healthcare Providers) |
| Human Capital Analytics                 | Analytical applications development                         |
| IT Analytics                            | Widely Used applications of analytics                       |
| Sales & Marketing Analytics             | Anatomy of Social Media Analytics                           |
| Analytics in Industries                 | Anatomy of Recommendation Systems                           |
| Analytics in Telecom                    |   |

---

### WHAT'S IN STORE?

This chapter deals with the application of analytics and looks at the application of analytics in business functions like HR, Sales, Marketing as well as Telecom, Retail and Healthcare. The chapter concludes with the Anatomy of Social Media Analytics and the Anatomy of Recommendation Systems.

---

#### 11.1 APPLICATION OF ANALYTICS

We have defined analytics as the computational field of examining raw data with the purpose of finding new insights, drawing conclusions and communicating inferences to support business decisions and actions. Analytics relies on the simultaneous application of statistics, operations research, programming and mathematical modeling techniques to quantify observations. It is evident that, to build analytical applications for businesses, you need to have different competencies. Let us first understand how different industries harness the power of analytics for business benefits. Then, we will look into common approaches used to build analytical applications. This will provide us with clues about common

algorithms that form the core of such analytical applications. Finally, with this background we will delve deep into some of the algorithms. We will not cover how these algorithms can be implemented in a language like R. We will only point to resources that can help to get to that level. This structured approach of starting from big picture of business application, then moving to identification of common algorithms and understanding the details of the algorithm will provide you good foundation to start your analytics journey.

First, let us look at how analytics is used in businesses from different perspectives.

1. How can analytics help decision making in business support or business enabling functions like HR, Finance, IT, Procurement, Marketing, etc.?
2. What are the common areas of analytics deployment in different industries like Retail, Health-care, Banking, Insurance, Telecom, etc.?
3. How analytics provides competitive advantage in business functions by focusing on customer facing functions like:
  - o Understanding customer or market segment.
  - o Customizing products/services to customers and market segments.
  - o Continuously listening to customer wants and needs.
4. Learn how social media analytics and recommendation systems are built.

### 11.1.1 Analytics in Business Support Functions

1. **Human Capital Analytics:** Every enterprise will have strategic and secure information about their human capital, that is, employees. The internal data sources may range from employee profiles, compensation and benefits, employee performance, employee productivity and so on, stored in variety of technologies like ERP systems, OLTP RDBMS, Hadoop ecosystem, spread-marts, data-marts and data warehouses. Some of the external data sources may include compensation benchmarks, employee sentiments, thought leadership contributions, etc. Enterprises have started gaining benefits from the following areas by deployment of analytics:
  - o **Workforce planning analytics** to acquire talent at the right time for right positions. Human capital investment analysis will lead to identification of positions that drive business results and critical competencies needed for those positions. Dow Chemical developed a custom modeling tool that predicts future hiring needs for each business unit and can adjust its predictions based on industry trends. (*Acquisition*)
  - o **Workforce talent development analytics** aligned to business goals. (*Development*)
  - o **Workforce sentiment analytics** for enhancing employee engagement. (Ability to simulate business impact of employee attrition) (*Engagement*)
  - o **Workforce utilization analytics** to ensure optimized deployment of right talent in right functions. This helps to connect employee performance to business results. (*Optimization*) Retail companies can use analytics to predict incoming call-center volume and release hourly employees early if it is expected to drop.
  - o **Workforce compensation analytics** helps to optimize benefits using big data sources including performance and benchmarks. (*Pay*)
  - o **Compliance analytics** helps to detect any anomalies relating to enterprise compliance policies and initiate proactive corrective actions. (*Compliance*)

2. **IT Analytics:** All enterprises use IT as business enabler. Enterprises invest in a variety of IT resources like data networks, servers, data center/cloud services, software licenses, maintenance of software, end user support and many productivity tools. IT operations of enterprises are becoming complex due to multiple technology platforms, outsourcing partners, complex demands of users and geographic spread of operations. Investing in right IT resources for business results is certainly a strategic decision.
  - o **IT infrastructure analytics** provide the insight into the health (availability and performance) of IT infrastructure. Service desks can prevent major outages by using predictive analytics.
  - o **Data network and storage utilization analytics** will lead to optimization of servers and bandwidth.
  - o **Security analytics** can provide vital clues about potential information security threats and alert teams.
  - o **Service quality analytics** will provide insights into root causes of SLA deviations and trigger process improvement initiatives.
  - o **IT assets analytics** supports optimal investment forecasts.
  - o **IT policy compliance analytics** can report policy enforcement deviations and trigger corrective actions.
3. **Sales and Marketing Analytics:** All enterprises leverage IT for many marketing activities. In its most basic form, marketing managers study reports relating to the customer segments, revenue share, revenue mix, marketing expenses trend, sales pipeline, marketing campaign performance and so on. Many enterprises use business intelligence solutions to slice-dice customer data, understand buyer behavior in various market segments and generate alerts against preset thresholds. In the world of analytics, these are termed as '**Descriptive Analytics**'. Managers are aware of what has happened and what is happening in the businesses. Analytics allows enterprises to move three steps further. *First*, '**Exploratory Analytics**' allows knowledge workers to find new business opportunities by discovering hidden data patterns. *Second*, mature organizations use '**Predictive Analytics**' to influence the future business. Finally, enterprises embark on '**Prescriptive Analytics**' to use the power of 'Algorithms, Machine learning and Artificial intelligence' techniques to make routine decisions almost instantaneous, thereby reducing the decision cycle times dramatically. Let us look at some of the common applications of analytics in sales and marketing functions of an enterprise.
  - o **Customer behavior analytics:** Customer behavior data, which tells decision makers what the customer does and where he/she chooses to do it, sits in multiple transaction systems across the company. Customer attitudinal data, which tells decision makers why a customer behaves in a certain manner or how he/she feels about a product, comes from surveys, social media, call center reports and so on. Using analytics to mine new patterns, conduct proof of concept analytics to find areas of business impact, develop predictive analytics solutions are some of the application areas.
  - o **Customer segmentation:** This application allows enterprises to define newer and sizable groups of target prospects using analytics. This enables enterprises to customize products and services to new segments and position them for competitive advantage. Segmentation can be more strategic, such as behavior-based profiling, predictive modeling, or customer-state and event-based segmentation.

- o **Modeling for pricing automation:** Deeper machine learning applications may fall in areas such as price elasticity modeling, channel affinity modeling, influence group link modeling and customer life event modeling.
- o **Recommendation systems:** Next best offer models can leverage many data sources and behavior of similar buyers to predict next best product or service your customer will look for and proactively recommend the perfect fit solution.

In the above examples we have considered examples of application of analytics for decision support in common enterprise functions like HR, Marketing and IT. Here are some of the important points to remember:

1. Data for analytics can be taken from many sources including OLTP data, data marts, data warehouses, big data sources and even data streams.
2. Analytics could be performed with the goal of **Discovery, Exploration, Prediction or Prescription**. Different visualization techniques will help decision makers interpret the results and initiate actions.
3. Analytics will need different types of tools for data transformation, exploration, modeling and visualization.
4. Knowledge of statistics, data mining, and business intelligence are some of the key skills needed to develop analytical applications. It is imperative that you have good functional knowledge of HR or Sales or Marketing or IT before you can design and develop the analytics.

Next let us understand application of analytics in different industries. It is also important to note that development of analytical applications will need hands-on experience in programming language, development methodology, concepts relating to architecture, security, user experience design and application performance requirements.

## 11.2 ANALYTICS IN INDUSTRIES

---

In this section let us look at how different industries apply analytics for business benefits. One needs to have some understanding of industry domain basics, trends, common current challenges in order to develop industry specific analytics solutions.

### 11.2.1 Analytics in Telecom

In the telecom industry, people and devices generate data  $24 \times 7$ , globally. Whether we are speaking with our friends, browsing a website, streaming a video, playing the latest game with friends, or making in-app purchases, user activity generates data about our needs, preferences, spending, complaints and so on. Traditionally, communication service providers (CSPs) have leveraged this tsunami of data they generate to make decisions in areas of improving financial performance, increasing operational efficiency or managing subscriber relationship. They have adopted advanced reporting and BI tools to bring facts and trends to decision makers. We have chosen some examples of strategic focus areas for deploying analytics, but this is not an exhaustive coverage of all possible areas of application of analytics. Let us look at the role of analytics in CSP business:

### ***Operational Analytics***

1. ***Network Performance:*** CSPs need to understand the bottlenecks in the network performance and optimize network utilization. They can use analytics to model capacity plans needed to meet service levels.
2. ***Service Analytics:*** This domain deals with analysis of customer problems, speed of resolution and identification of priority customers and ensures their satisfaction. Advanced analytics will deliver customer sentiment through social media analysis.
3. ***Regulatory Analytics:*** CSPs collaborate with other carriers and partners to support roaming, sharing infrastructure, etc. They need to track regulatory compliance as per agreed contract norms and handle deviations through anomalies detection.
4. ***Product Analysis:*** It involves analysis of data to enhance revenue, launch promotions, create campaigns, create new segments, strategize pricing and study churn.

### ***Subscriber Analytics***

1. ***Subscriber Acquisition:*** CSPs study customer behavior to identify the most suitable channels and sales strategy for each product.
2. ***Fraud Detection:*** Analytics helps to detect billing and device theft, cloned SIMs and related frauds as well as misuse of credentials.
3. ***Churn Analytics:*** This helps CSPs to not only model the loyalty programs but also predict churn and destination CSP.
4. ***Value Segment Prediction:*** Here CSPs will be able to enhance revenue by defining new subscriber base ahead of competition by matching their profitable offerings to subscribers needing them.

### ***Financial Analytics***

1. ***Infrastructure Analytics:*** CSPs study CAPEX and optimize investments in infrastructure and save money by considering utilization options.
2. ***Product Portfolio Analytics:*** This area provides information into the profitable products and services and helps to exit from loss making products.
3. ***Channel Analytics:*** Helps CSPs to optimize the commercial terms with partners to optimize distributor margins.
4. ***Cost Reduction:*** This area focuses on reducing service management cost, operations cost, compliance risks related cost, etc.

#### **11.2.2 Analytics in Retail**

Only some industries have greater access to data around consumers, products they buy and use, and different channels that sell and service products – and the lucky vertical is retail industry. Data coupled with insights are at the heart of what drives the retail business.

Technologies like Point of Sale (PoS), CRM, SCM, Big Data, Mobility and Social Media offer a means to understand shoppers via numerous digital touch points ranging from their online purchases, to their presence on social networks, to their visits to brick and mortar stores as well as tweets, images, video, and more. Even today retailers are grappling with how to meaningfully leverage and ultimately

monetize the hidden insights around huge amounts of structured and unstructured data about a consumer.

Value of analytics can come from three sources:

1. Gaining insight to improve processes and resource optimization.
2. Personalizing and localizing offers.
3. Creating community for branding and customer engagement.

### Gaining insight to improve processes and resource optimization

1. **Supply Chain Analytics:** Every retailer needs to optimize the vendors of products, its cost and quality. They need to constantly track the performance of supply chain and initiate proactive actions for competitive advantage.
2. **Pricing Analytics:** Helps retailers to optimize the product pricing, special offers, merchandizing, loyalty programs and campaigns that attract maximum number of consumers both from physical store and online store perspective.
3. **Buying Experience Analytics:** Retailers can gain insight into the path taken to purchase, complaints registered, help provided by store personnel, store layout/item search time, product details availability, pricing, etc. and enhance the buying experience and train personnel for enhancing consumer loyalty.

### Personalizing and localizing offers

1. **Inventory Analytics:** Retailers aim to fulfill consumer demand by optimizing stocks and ability to replenish when consumer demand increases due to seasonal effects or as a result of powerful campaigns. This area of analytics will alert store managers about the potential need for stocking highly moving items and reduce slow moving items.
2. **Consumer Analytics:** Every region around the world has people with different taste for goods and service levels. The purpose of consumer analytics is to equip store managers with insights to customize their products and services to the local consumer profile.
3. **Campaign Analytics:** All retailers will have digital marketing programs to entice consumers with value offers. Retailers invest in this area of analytics to design most effective campaigns that convert maximum number of consumers into buyers.
4. **Fraud Detection:** All retailers strive to eliminate fraud relating to payments, shipping, and change of price tags and so on. Analytics can study transactions in real-time to detect fraud and alert store personnel or online commerce teams.

### Creating community for branding and customer engagement

1. **Web Analytics:** Here the different perspectives of each consumer's online behavior such as surfacing traffic, visitor and conversion trends, location of smart devices, access to kiosks will be analyzed to recommend the best sales approach in response to each of the customer's real-time actions.
2. **Market Basket Analytics:** The promotion, price, offer, and loyalty dimension of shopping behaviors will be used to understand sales patterns, customer preferences, and buying patterns to create targeted and profitable product promotions, customer offers and shelf arrangements.
3. **Social Media Analytics:** Listening and learning from the social community dimension of each consumer's online behavior is the scope of this area of analytics. Here store taps into

customer-generated content with sentiment and behavioral analysis to answer key merchandise, service, and marketing strategy questions.

4. **Consumer Behavioral Analytics:** The focus area is consumer preferences such as channels, categories, brands, and product attributes; return and exchange patterns; usage level of service programs; and participation in loyalty programs.

### 11.2.3 Analytics in Healthcare (Hospitals or Healthcare Providers)

Healthcare is a very complex eco-system of multiple industries interconnected to achieve the healthcare goals of a country. These entities include healthcare providers, physicians, insurance companies, pharmaceutical companies, laboratories, healthcare volunteers, regulatory bodies, retail medicine distributors and so on centered on a patient. You can imagine the complexity, variety, volume, velocity of data that gets generated in each of these independent enterprises and multitude of interconnected heterogeneous IT applications. Analytics is applicable for all these enterprises, viz. insurance companies, pharmaceutical manufacturers, hospitals, etc. Here we will focus on how hospitals, that is, healthcare providers, can leverage analytics for goals like:

1. **Hospital Management Analytics:** It focuses on cost reduction, enhancing quality of care, improving patient satisfaction, improving outcomes (performance of diagnosis, testing and treatment), providing secure access to patient data (Electronic Health Records – EHR). Analytics in this area can support fact-based decisions in areas of reduction of medical errors, manage diseases, understand physician performance and retain patients.
2. **Compliance Analytics:** Provide healthcare compliance metrics to regulatory authorities and benchmark against world-class hospitals using Baldrige criteria. Improvement in widespread use of digital data will support audits, analytics and improve hospital processes needed for regulatory compliance.
3. **Financial Analytics:** This area of analytics will lead to enhance RoI (Return on Investment), improved utilization of hospital infrastructure and human resources, optimize capital management, optimize supply chain and reduce fraud.
4. **Predictive Models:** They can help healthcare professionals go beyond traditional search and analysis of unstructured data by applying predictive root cause analysis, natural language and built-in medical terminology support to identify trends and patterns to achieve clinical and operational insights. Healthcare predictive analytics can help healthcare organizations get to know their patients better, so that they can understand their individual patient's needs, while delivering quality, cost-effective life-saving services.
5. **Social Analytic:** It can help hospitals listen to patient sentiments, requirements, affordability, and insurance to model care and wellness programs customizing services by localization of needs.
6. **Clinical Analytics:** A number of other critical clinical situations can be detected by analytics applied to EHR such as:
  - o Detecting postoperative complications.
  - o Predicting 30-day risk of readmission.
  - o Risk-adjusting hospital mortality rates.
  - o Detecting potential delays in diagnosis.
  - o Predicting out of intensive care unit death.

### 11.2.4 Analytical Application Development

Before we dive deep to understand some of the common applications of analytics in businesses, it is important to know the development methodology. Here we are trying to summarize the design, development and deployment steps in their simplified form.

**Stage 1: Defining the problem.** After this step you should be able to explain – what business question(s) are you trying to answer? Once you understand this, you need to think about what data is available to you to answer the question:

1. Is the data directly related to the question?
2. Is the data you need even available within the enterprise or elsewhere?
3. What measure of accuracy and granularity are you going to use? Is that level of summaries good enough for the business users?
4. What criteria are you going to use to determine success or failure? Determine, up front, how you are going to measure the results.

**Stage 2: Setting-up technical environment and processing the data.** Collect the data and perform basic data quality checks to ensure accuracy and consistency. While this may end up taking the most time, it is critical and erroneous data will create erroneous results. You may need to transform the data to make it conducive for analysis. You will need to pick the analytics approach and possible choice of algorithms and visualization requirements.

**Stage 3: Running the initial analysis or model.** You may like to split the data set into a test data set and a validation data set. This is also the step whereby you will choose the method or methods by which you want to build the model and process the data. As you become more familiar with predictive modeling and with your own data, you will find that certain types of problems align with certain types of modeling approaches or algorithms.

**Stage 4: Evaluate the initial results.** Are the results in line with what you were expecting to see? Are you able to interpret the results? Do they answer the business question you are trying to answer? If the answer is yes, then move on to the next step. If the answer is no, then consider the following:

1. Try using different algorithms/models.
2. Consider collecting more or different data.
3. Consider redefining or reframing the problem, changing the question and the means to an answer as you better understand your data and your environment.

**Stage 5: Select the final model.** You may want to try a number of different models and then when you are satisfied with the results, choose the best one. Run the selected model or analysis and re-examine the results.

**Stage 6: Test the final model.** It is important to test the final model and the only way to do so is to take the selected model and run it against the validation data set and assess the results. Do not tweak or change the model in any way at this stage, as it will invalidate any comparison to the initial results. If the results are similar and you are satisfied with them you can move on to the final stage. If you are not, then go back (to stage 3) to reassessing the model and the data, make any necessary or desired changes and try re-running the model again.

**Step 7: Apply the model and validate by usage.** There could be some data exceptions that the analysis model may not handle. You may need to keep checking for impact of such data on results and

visualization. You may consider adding incremental features and newer visualizations once the analytical model is stable and provides consistent results for decision-making.

## 11.3 WIDELY USED APPLICATION OF ANALYTICS

---

After looking at some of the applications of analytics that are common to many industries and applications specific to industries, we are now in a position to find some general patterns of application. We will use examples of some powerful application areas of analytics and then understand “How these are built?”, “What are the components of such application?”, and “What algorithms are used to develop such applications?” We have provided detailed explanation of the algorithms used in these types of analytical applications in the data mining section. The only step remaining will be to understand how these algorithms are coded in language like R.

From the previous sections it is clear that most industries tend to look at applying analytics in areas of:

1. Processing ***social media data*** for business benefits – Telecom CSPs stand to understand the voice-of-the subscribers; HR will understand the sentiment of employees and partners; hospitals discover unmet needs of patients; IT function will understand the business user challenges and service level expectations. Hence we will take-up **Social Media Analytics** for diving deeper. Web analytics or digital analytics is another area that is leveraged by product and services enterprises.
2. All product and service enterprises would be striving to acquire new customers without any exception. Each one would like to customize and personalize offers so that prospects see value and make a buying decision. One of the most common approaches enterprises take is the ***recommendation system or engine*** to predict the most potential buyers. So we will see what is inside a recommendation engine. This common analytics paradigm is being used to recommend books, gift items for various occasions, doctors in a local area, household items for online purchase – the list is endless.

### 11.3.1 Anatomy of Social Media Analytics

Today, there are over 2 billion social media users. The major social networking platforms include:

1. Nearly 50% of Internet users are on Facebook.
2. Over 500 million tweets are sent each day.
3. The +1 button on Google Plus is hit 5 billion times each day.
4. 70 million photos and videos are exchanged in Instagram.
5. There are 39 million recent college graduates on LinkedIn.
6. 600 million Skype calls are active each day.
7. Similar number of users are on WhatsApp.

We can define social media as the technology-enabled interactions among peer-to-peer, employees-and-customers as well as among employees. Today, businesses use social media for achieving business goals like:

1. Deploying a contest to make robust consumer engagement.
2. Launching campaigns that are targeted to specific target segments.

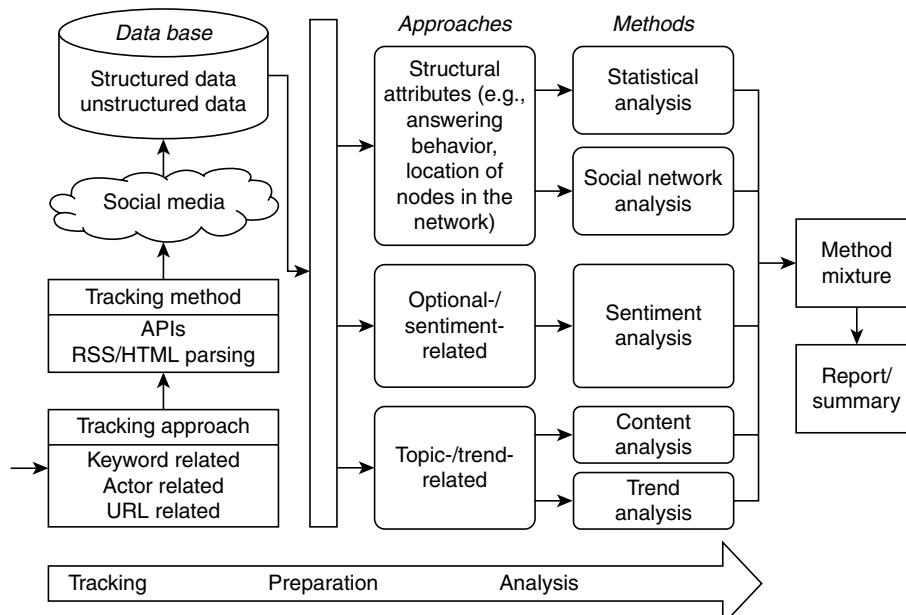
3. Providing customers with a wide range of products, offers, and initiatives via social media.
4. Recommending “e-commerce offers” to promote the online purchase.
5. Alerting customers about events they look forward to.
6. Managing risk and fraud by creating awareness.
7. Creating dedicated forums with topic to enable followers joining live discussions.

Social media has given business results like increased customer acquisition, increased cross-sell, enhanced brand image, accurately capture market requirements and reduced cost to serve or sell. The actions initiated by decision makers using social media analytics could be learning from consumers, reacting to consumer sentiments/feedback, supporting consumer ideas, driving product or service messages and serving multiple communities with specific profiles.

Using techniques like opinion mining, text mining, audio analytics, sentiment analysis, and predictive analytics allows you to look at historical patterns and make predictions about future behavior for specific individuals. By taking customer data that you hold internally and adding what people have said and done, you can predict what your customers are likely to do in similar set-ups.

Figure 11.1 shows a typical process for designing and developing social media analytics application. Some questions that would come to our minds include:

1. How to identify data sources that directly relate to business goals?
2. How to get social media data from the applications introduced above?
3. What technologies are available for storing and processing such data?
4. What are the common algorithms or analytics methods used to extract insight from such data?
5. What are some of the common tools used for social media analytics and how to report the results of such analytics?

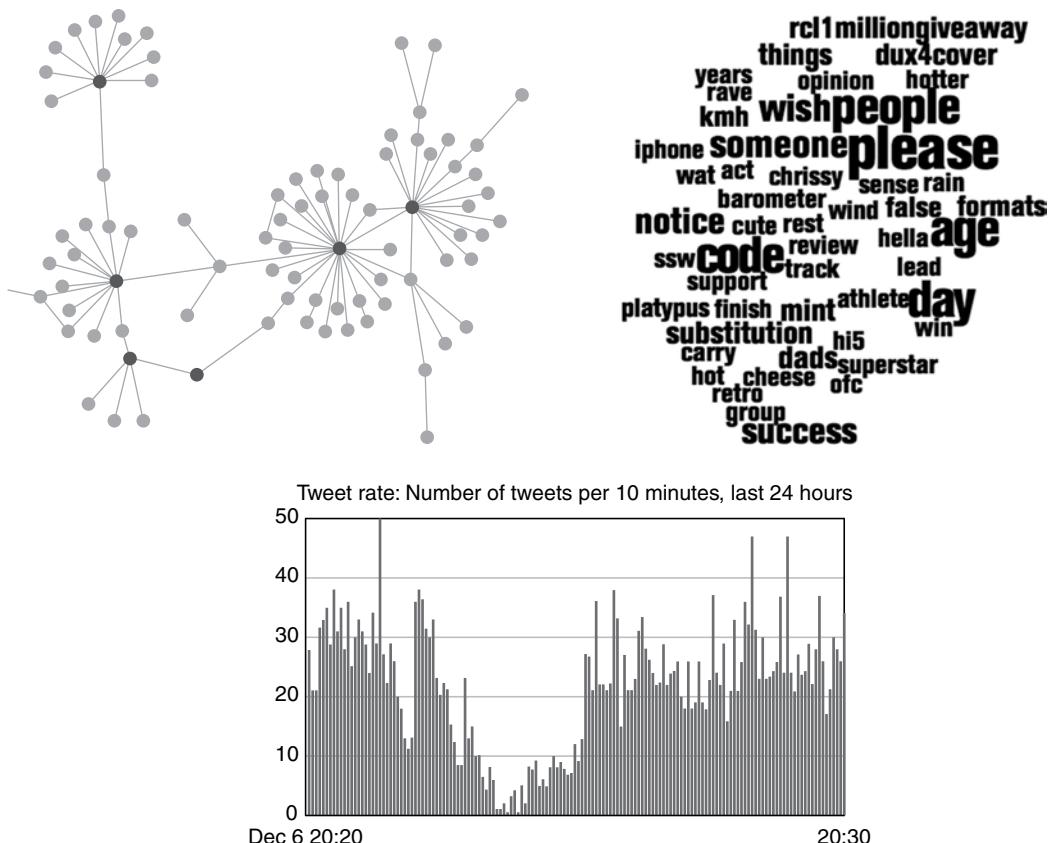


**Figure 11.1** A typical process for designing and developing social media analytics application.

The following paragraphs will exactly provide these insights.

Points to consider while designing and developing a Social Media Analytics Application:

1. **Relating business goals to data sources:** Typically, social media analytics projects start with identification of business goal. Some of the common business goals include customer insight management, product or brand reputation management, innovation management, general opinion tracking and product/service value management. These goals will have specific target audience with specific demographic and psychographic profiles. Not all social media applications attract all the population. Hence once you are clear about the business goal associated with target groups, the first job will be to identify the social media applications they use.
2. **Collecting and storing social media data:** The next step is to track the relevant data needed for the goal. Different social media applications provide different methods of data extraction from the huge amounts of social media transactions that happen in real time. Some applications allow data tracking by keywords while other use URLs. Depending on the social media platform(s), APIs, RSS, or HTML parsing can be used to track structured data or unstructured data like textual content in social media. There are many third-party applications that help you interface with social media platform and provide required data. REST API is a common method used for this purpose.



3. **Approaches to process gathered data:** Data analysis approaches depends on end result you need. For example, if you are interested in behavioral analysis using data attributes like location, demographic details, and influencers in the network, then you will use structured data store and queries. Statistical analysis on this data is also a common requirement. On the other hand, if the aim is to mine opinion or perform sentiment analysis, you will need to focus on sentiment indicators. You may need different set of approach to look into the content of the interactions to find the trend or analyze content. Text analysis is one of the common requirements in such situations.
4. **Data analysis methods and tools:** Some of the common methods used for data analysis include filtering, statistics, regression, social network analysis, text analysis, trend analysis and sentiment analysis. Many times multiple methods are employed to get the final results needed. In big data (Hadoop HDFS) scenarios, MongoDB could be used for storing Twitter data. You may collect user tweets using REST APIs and user data using open authentication (OAuth). Network diagrams, heat maps, tweet volume charts and word clouds are used to display the results of Twitter data analysis as shown below. R and Python languages are the commonly used for programming.

### 11.3.2 Anatomy of Recommendation Systems

Recommendation engines are not totally new; they take results from market basket analysis of retail business data to advanced analytic systems and suggest the next best offer or next best activity for a specific customer. They are also very popular for making suggestions or recommendations when an online store visitor starts looking at a product or service. Amazon is probably the most famous example that uses recommendation engine analytics. In the past, all types of recommendation-based analytics were quite difficult to automate as the data storage, preprocessing, model creation, visualization and integration were complex and generally needed multiple IT systems working together.

Today, we can see recommendation systems being used in a variety of scenarios such as:

1. A restaurant to dine in a new location you are visiting.
2. Music you may want to listen to next.
3. Newspaper or magazine articles to read next.
4. Right doctor in your neighborhood for treatment.
5. Best auto insurance policy to buy.
6. Next vacation travel spot.
7. Best online store for your grocery and so on.

Applications of recommendation systems have transcended customers' shopping experience. Market research has shown that recommendation systems bring in anything between 10% and 30% of additional revenue for a company. Early adopters of recommendation engine technologies such as Amazon and Netflix have outperformed their competitors by leveraging the unparalleled customer insights generated by their proprietary recommendation systems.

Both Recommendations and Pricing are classic topics for advanced analytic modeling, and both offer possibilities for real-time scoring. As we build more accurate models and train them with real-life data, the more accurate will be the recommendations and the prices the company can offer. It will be a great advantage for retailers to change the price dynamically to acquire more customers. When an item is desired, there is more of a willingness to pay a premium price. When an item is less desired, the

price the customer will pay, will play an important role in the decision-making process. Discounts are a classic way of helping customers not only choose a particular supplier, but to help a customer move from undecided state to purchase commitment. At the same time, discounts are expensive. They eat into the profit a company makes. In an ideal world, we would make discounting decision based on the confidence of closing the deal immediately.

Recommendation systems predict customer needs based on previous purchase history, online search and navigations, social media interactions content, ratings users have provided for a product/service, analysis of reviews of products/services and other personalized attributes captured. Such recommendation engines need to track each customer interaction such as log-in, price comparison, product selection to cart, actual purchase, comment or rating posted.

There are many commercial eCommerce platforms such as Baynote, Omniture and RichRelevance that provide multiple approaches to determine the most appropriate product or service for recommendation. Some of the common algorithms or mechanisms supported by these platforms include rule engine, modified rule engine, recommendations based on social media traffic like Facebook, Twitter, etc., recommendations based on reviews and ratings, Ad-Word, Internet search terms used, syndicated recommendations and collaborative recommendations.

Structures built in the recommendation system include item-to-item association, many items-to-item(s) association, person-to-person associations, person-to-item associations and user behavioral heuristics.

In the following section let us understand the generic model of such recommendation systems.

### 11.3.3 Components of Recommendation Systems

Recommender systems may be based on several different techniques such as collaborative filtering, content filtering, social filtering, or various hybrid approaches that use a combination of these. Though the design of these systems vary in their detail, it is possible to abstract their characteristics and behavior to arrive at a common structure:

1. **Tracking user actions and behavior:** The users' behavior as they use the application is observed to know the items they may be looking for, their specifications, the preferences, experience and feedback of the user, and so on. The techniques used to track user behavior include capturing and analyzing click sequences, tracking eye movement, tracking navigation sequences and measuring time spent in specific sections of the application, items searched, attributes specified in search, number of similar items typically viewed, etc. They help in identifying and/or deducing user interests and preferences.
2. **Capturing, cleansing, normalization and storing of data:** The users' actions result in generating data that represents their interests, preferences, feedback on the information displayed, etc. which helps to build a model of the user and to compute similarities with other users. This helps to profile the user to personalize the information presented to the user and to improve the recommendations made. The collection of data may be based on observing the users' explicit actions such as searching for items, ratings, feedback on recommendations provided, etc. or implicit behavior such as the time spent looking at a certain item, navigation patterns, bookmarking, etc. As data is collected across users and over a period of time, it is essential to eliminate any aberrations or contradictions that may have crept in and keep it consistent.

The data store typically also includes one or more indexes created based on full-text search of the contents of the items, and their description and other metadata. These indexes are usually pre-computed in order to improve the real-time performance of the recommender.

Biases and anomalies such as undue dominance of certain parameters are compensated for or eliminated by applying techniques such as term frequency-inverse document frequency (tf-idf). The process of index creation may also involve pruning of frequently occurring but non-significant terms and coalescing variants of terms through techniques such as stemming and substitution of synonyms with an equivalent term. Such steps are often collectively referred to as normalization.

The components employed in the implementation of this module may include a full-text search engine such as Lucene or Solr. Very high volume and volatile data together with stringent scalability requirements may call for the use of a distributed search technology such as Elasticsearch to sustain performance in the face of heavy user loads.

3. **Prediction of relevant items and their ratings:** The current actions and data inputs from the user are overlaid on the information in the data store to generate the predictions. As mentioned above, a variety of approaches such as collaborative filtering, item-to-item filtering, content filtering, and their hybrids using algorithms ranging from primitive Euclidean distance-based similarity measures to sophisticated ones based on advanced machine learning algorithms can be applied for this purpose. Typically, this results in a list of items scored and ranked on relevance to the items that the user is looking for in order to determine relevance and improve the recommendations made.

Components required for implementation of this module include feature analyzers/extractors, components implementing logic for analysis, dimensionality reduction and validation of high-value features, modules for user/item clustering, one or more types of similarity finders based on user models and item data, which implement prediction algorithms based on statistical and machine learning techniques, and so on, depending on the sophistication of the implementation. Comparison and similarity analysis of user models is especially important in collaborative filtering scenarios which are specifically suited for content which is not amenable to machine analysis such as images and videos.

Technologies used in the implementation of these modules typically consist of the Big Data tools like Hadoop, MapReduce, and Spark, leveraging a wide-array of NoSQL horizontally-scalable Big Data stores such as HBase, Cassandra, Neo4j. Machine learning technologies specifically built for Big Data like Mahout with several built-in algorithms for predictions are gaining popularity for generating real-time results while serving millions of simultaneous user requests.

4. **Recommendation based on the predictions:** This module consists of the logic to generate user-friendly and valuable recommendations based on the ranked or weighted predictions from the previous step. The predictions are scored and combined with some application context and user choices to generate a set of recommendations catering to the user's interest. For example, the recommendations may be items that have a similarity index above a certain threshold determined by the algorithm or specified by the user or the top “ $n$ ” similar items. Recommendations are often based on user “neighborhood” considerations which confines the “distance” between users to a certain computed or configured limit specified in terms of similarity parameters.

Considering the potentially very large number of combinations of diverse factors on which recommendations can be based, this module will typically allow extensive configuration at several levels – administrators, business users and consumers – for filtering, formatting and presentation of results.

The above description outlines the essential components of basic recommendation system. As mentioned earlier, real-world implementations may vary in their complexity and sophistication. Primarily, in addition to the components described above, these implementations may have components for caching data such as user profiles and user models, item data, computed similarity metrics, etc. for real-time performance optimization.

In addition, the design may be further fine-tuned for maximizing online performance through moving the data store maintenance offline or asynchronous operation. A recent innovation in recommendation systems is to improve the quality of recommendations by factoring in the user location and other types of context information that mobile devices are capable of delivering. In such cases, the recommendation system may include additional components to cater to such scenarios.

Thus in this chapter, we have seen application of analytics in business functions, various industries and also looked at two common analytics systems viz. social media analytics and recommendation systems. We have also shared the typical analytical application development process as well as technologies needed. We will familiarize you with the algorithms such as k-means, clustering, association, etc. in Chapter 12, Data Mining.



### *Point Me (Books)*

- Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, Or Die by Eric Siegel.
- The Elements of Statistical Learning: Data

Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) by Trevor Hastie, Robert Tibshirani and Jerome Friedman.



### *Connect Me (Internet Resources)*

- <https://www.informs.org/Recognize-Excellence/Community-Prizes-and-Awards/Analytics-Society/Innovative-Applications-in-Analytics-Award>
- <http://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/applying-advanced-analytics-in-consumer-companies>



# 12



## Data Mining Algorithms

---

### BRIEF CONTENTS

|   |                                |
|---|--------------------------------|
| Association Rule Mining                         | What is $k$ -means clustering? |
| Binary Representation                           | Decision Tree                  |
| Item Set and Support Count                      | What are the uncertainties?    |
| Why should you consider support and confidence? | What is a decision tree?       |
| Implementation in R                             | Where it is used?              |
| $k$ -Means Clustering                           | Advantages of Decision Tree    |
| Why should you learn about clustering?          | Disadvantages of Decision Tree |

---

### WHAT'S IN STORE?

The focus of this chapter is to build knowledge about Data Mining Algorithms. We will discuss Association Rule Mining,  $k$ -Means Clustering and Decision Trees. We will also discuss implementation of Association Rule Mining,  $k$ -Means Clustering and Decision Trees using **R** statistical tool.

We suggest you refer to some of the learning resources provided at the end of this chapter for better learning.

---

### 12.1 ASSOCIATION RULE MINING

***Picture this:***

You are at your favorite salon for a hair-cut. The hair-dresser offers you a rather appealing deal. A head massage, hair wash or hair coloring at a slightly more price. You think about it and find the offer too good to refuse. You settle for hair-coloring along with hair-cut. After all you have been wanting to color your hair a different color for quite a while.

You are shopping online. You are about to make a check-out. Just then, a rather relevant product is recommended to you at discounted price. You pull the product into the shopping cart and proceed to checkout.

What happened here is: that the hair-dresser at the salon and the online retailer just cross-sold to you.

### ***Why should you learn about Association Rule Mining?***

1. You wish to retain your existing customers and keep them happy.
2. You wish to enhance quality of customer experience by recommending the most relevant product.
3. You wish to have the deals made to your customers convert to sales.

### ***How will you accomplish the above stated?***

The answer is simple – by using the power of association rule mining. Association rule mining is also referred to as Market Basket Analysis (MBA). Few also prefer to call it as affinity analysis.

### ***What is Market Basket Analysis (MBA)?***

It is a data analysis and data mining technique. It is used to determine co-occurrence relationship among activities performed by individuals and groups. Wikipedia defines cross-selling as “an action or practice of selling an additional product or service to an existing customer”. Association analysis is mostly done based on an algorithm named “Apriori Algorithm”.

### ***What questions does MBA help to answer?***

1. Should detergents be stocked together with other cleaning agents such as window cleaning agents or floor cleaners? Where should they be stocked in the store to maximize sales?
2. Should floor mats and runners be placed alongside health and personal care products?
3. Are chips and wafers purchased alongside soft drinks? Does the brand of soft drinks matter?

### ***Few Examples of Market Basket Analysis***

Market basket analysis is widely used in retail wherein the retailer seeks to understand the buying behavior of customer. This insight is then used to cross-sell or up-sell to the customers.

If you have ever bought a book from Amazon, this should sound familiar to you. The moment you are done selecting and placing the desired book in the shopping cart, pop comes the recommendation stating that customers who bought book “A” also bought book “B”.

Who can forget the urban legend, the very famous beer and diapers example. The legend goes... there was a retail firm wherein it was observed that when diapers were purchased alongside there was purchase of beer as well by the customer. The retailer cashed in on this opportunity by stocking beer coolers close to the shelves that housed the diaper. This just to make it convenient for the customers to easily pick both the products.

An association rule has two parts: (a) An antecedent (if) and (b) a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

### ***Picture this:***

A retailer “BigDailies” wants to cash in on its customers’ buying patterns. They want to be able to enact targeted marketing campaigns for specific segments of customers. They wish to have a good inventory management system in place. They wish to learn about which items/products should be stocked together to provide ease of buying to customers, in other words enhance customer satisfaction.

Where should they start? They have had some internal discussions with their sales and IT staff. The IT staff has been instructed to design an application that can house each customer transaction data. They wish to have it recorded every single day for every single customer and for every transaction made. They decide to meet after a quarter (3 months) to see if there is some buying pattern.

Presented in Table 12.1 is a subset of the transaction data collected over a period of three months:

**Table 12.1** Sample transactional data set

| Transaction ID |  | Transaction Details                |
|----------------|--|------------------------------------|
| 1              |  | {bread, milk}                      |
| 2              |  | {bread, milk, eggs, diapers, beer} |
| 3              |  | {bread, milk, beer, diapers}       |
| 4              |  | {diapers, beer}                    |
| 5              |  | {milk, bread, diapers, eggs}       |
| 6              |  | {milk, bread, diapers, beer}       |

This table presents an interesting methodology called association analysis to discover interesting relationship in large data sets. The unveiled relationship can be presented in the form of association rules or sets of frequent items. For example, the following rule can be extracted from the above data set:

$$\{\text{Diapers}\} \rightarrow \{\text{Beer}\}$$

It is pretty obvious from the above rule that a strong relationship exists between the sale of diapers and beer. Customers who pick up a pack or two of diapers also happen to pick a few cans of beers. Retailers can leverage this sort of rules to partake of the opportunity to cross-sale products to their customers. Challenges that need to be addressed while progressing with association rule mining are as follows:

1. The larger the data set, the better would be the analysis results. However, working with large transactional data sets can be and is usually computationally expensive.
2. Sometimes few of the discovered patterns could be spurious or misleading as it could have happened purely by chance or fluke.

### 12.1.1 Binary Representation

Let us look at how we can represent the sample data set in Table 12.1 in binary format (see Table 12.2).

*Explanation of the below binary representation:* Each row of Table 12.2 represents a transaction identified by a “Transaction ID”. An item (such as Bread, Milk, Eggs, Diapers and Beer) is represented by a binary variable. A value of 1 denotes the presence of the item for the said transaction. A value of 0 denotes the absence of the item from the said transaction. Example: For transaction ID = 1, Bread and Milk are present and are depicted by 1. Eggs, Diapers and Beer are absent from the transaction and therefore denoted by zero. The presence of the item is more important than its absence, and for the same reason an item is called as an asymmetric variable.

**Table 12.2** Sample transactional data set represented in binary format

| <i>Transaction ID</i> | <i>Bread</i> | <i>Milk</i> | <i>Eggs</i> | <i>Diapers</i> | <i>Beer</i> |
|-----------------------|--------------|-------------|-------------|----------------|-------------|
| 1                     | 1            | 1           | 0           | 0              | 0           |
| 2                     | 1            | 1           | 1           | 1              | 1           |
| 3                     | 1            | 1           | 0           | 1              | 1           |
| 4                     | 0            | 0           | 0           | 1              | 1           |
| 5                     | 1            | 1           | 1           | 1              | 0           |
| 6                     | 1            | 1           | 0           | 1              | 1           |

### 12.1.2 Itemset and Support Count

Let  $I = \{i_1, i_2, i_3, \dots, i_n\}$  be the set of all items in the market basket data set.

Let  $T = \{t_1, t_2, t_3, \dots, t_n\}$  be the set of all transactions.

**Itemset:** Each transaction  $t_i$  contains a subset of items from set  $I$ . A collection of zero or more items is called an itemset. If an itemset contains  $k$  elements, it is called a  $k$ -item itemset. Example: the itemset {Bread, Milk, Diapers, Beer} is called a 4-item itemset.

**Transaction width:** Transaction width is defined as the number of items present in the transaction. A transaction  $t_j$  contains an itemset  $X$  if  $X$  is a subset of  $t_j$ . Example: Transaction  $t_6$  contains the itemset {Bread, Diapers} but does not contain the itemset {Bread, Eggs}.

**Item support count:** Support is an indication of how frequently the items appear in the data set. Item support count is defined by the number of transactions that contain a particular itemset.

Item support count can be expressed as follows: Number of transactions that contain a particular itemset.

*Example:* Support count for {Diapers, Beer} is 4.

Mathematically, the support count  $\sigma(X)$ , for an item set  $X$ , can be expressed as

$$\sigma(X) = | \{t_i | X \subset t_i, t_i \in T\} |$$

The symbol  $| - |$  denotes the number of elements in the set.

**Association rule:** It is an implication rule of the form  $X \rightarrow Y$ , where  $X$  and  $Y$  are disjoint items, that is,  $X \cap Y = \emptyset$ . To measure the strength of an association rule, we rely on two factors: the support and the confidence.

1. *Support* for an itemset is defined as:

$$\text{Support}(x_1, x_2, \dots) = \frac{\text{Number of transactions containing } (x_1, x_2, \dots)}{\text{Total number of transactions } (n)}$$

$$\text{Support for } X \rightarrow Y = \frac{\text{Number of transactions containing } x_1, x_2, \dots \text{ and } y_1, y_2, \dots}{\text{Total number of transactions } (n)}$$

*Example:* Support for {Milk, Diapers} → {Beer} as per the data set in Table 12.1 is as follows:

$$\text{Support for } \{\text{Milk, Diapers}\} \rightarrow \{\text{Beer}\} = \frac{3}{6} = 0.5$$

2. *Confidence* of the rule is defined as:

$$\text{Confidence of } (x_1, x_2, \dots) \text{ implies } (y_1, y_2, \dots) = \frac{\text{Support for } (x_1, x_2, \dots) \text{ implies } (y_1, y_2, \dots)}{\text{Support for } (x_1, x_2, \dots)}$$

$$\text{Confidence of } \{\text{Milk, Diapers}\} \rightarrow \{\text{Beer}\} = \frac{\text{Support for } \{\text{Milk, Diapers}\} \rightarrow \{\text{Beer}\}}{\text{Support for } \{\text{Milk, Diapers}\}}$$

Substituting we get

$$\text{Confidence of } \{\text{Milk, Diapers}\} \rightarrow \{\text{Beer}\} = \frac{0.5}{0.67} = 0.7462$$

*Why should you consider support and confidence?*

It is important to consider support and confidence owing to the following reasons:

A rule which has low support may occur simply by chance. To place big bets on it may prove futile. It may prove rather uninteresting and non-profitable from a business perspective because it does not make sense to promote items that customers seldom buy together. Support is used to chuck off uninteresting rules.

Confidence is a measure of reliability of inference of an association rule. For a given rule  $X \rightarrow Y$ , the higher the confidence, the more likely it is for  $Y$  to be present in transactions that contain  $X$ . Confidence of a rule can also be used to provide an estimate of the conditional probability of  $Y$  given  $X$ .

The results of association rule analysis should be considered with caution. It does not necessarily imply causality. Causality, in fact, requires knowledge about the causal and effect attributes in the data. It requires the relationship to be observed, recorded and studied over a period of time. For example, ozone depletion leads to global warming. The association rule mining is more in the nature of establishing co-occurrence relationship between items in the antecedent and consequent of the rule.

### 12.1.3 Implementation in R

```
> transdata <- read.csv("d:/trans. Csv")
> transdata
```

|    | Transaction. ID | Transaction. Details |
|----|-----------------|----------------------|
| 1  | 1               | bread                |
| 2  | 1               | milk                 |
| 3  | 2               | bread                |
| 4  | 2               | milk                 |
| 5  | 2               | eggs                 |
| 6  | 2               | diapers              |
| 7  | 2               | beer                 |
| 8  | 3               | bread                |
| 9  | 3               | milk                 |
| 10 | 3               | beer                 |

```
11          3           diapers
12          4           diapers
13          4           beer
14          5           milk
15          5           bread
16          5           diapers
17          5           eggs
18          6           milk
19          6           bread
20          6           diapers
21          6           beer

> AggPosData <- split(transdata$Transaction.details, transdata$Transaction.ID)
> txns<-as (AggPosData,"transaction")

> summary(txns)
transactions as itemMatrix in sparse format with
  6 rows (elements/itemsets/transactions) and
  5 columns (items) and a density of 0.7

most frequent items:
      bread diapers      milk beer   eggs (Other)
      5       5        5     4     2     0

element (item/transaction) length distribution:
sizes
2 4 5
2 3 1
    Min. 1st Qu.      Median      Mean 3rd Qu.      Max.
    2.0   2.5        4.0        3.5   4.0       5.0

includes extended item information - examples:
  labels
1      beer
2      bread
3      diapers

includes extended transaction information - examples:
  transaction ID
1                  1
2                  2
3                  3

includes extended transaction information - examples:
  transaction ID
1                  1
2                  2
3                  3

> rules <-apriori(txns, parameter=list(supp=0.05, conf=0.4))
Apriori
```

Parameter specification:

```
confidence minval smax arem aval origionalSupport support minlen maxlen target
      0.4      0.1     1 none   FALSE           TRUE      0.05       1     10    rules

ext
FALSE
```

Algorithmic control:

```
Filter  tree   heap   memopt   load   sort   verbose
  0.1    TRUE   TRUE   FALSE    TRUE     2     TRUE
```

Absolute minimum support count: 0

Warning in apriori(txns, parameter = list(sup = 0.05, conf = 0.4)):

You chose a very low absolute support count of 0. You might run out of memory! Increase minimum support.

```
set item appearances ....[0 items(s)] done [0.00s].
set transactions ....[5 items(s), 6 transaction(s)] done [0. 00s].
sorting and recoding items ....[5 items(s)] done [0. 00s].
creating transaction tree ....done [0. 00s]
checking subsets of size 1 2 3 4 5 done [0.00s].
writing ....[71 rule(s)] done [0.00s].
creating S4 object ....done [0.00s].
```

> inspect(rules)

|    | lhs          | rhs          | support | confidence | lift |
|----|--------------|--------------|---------|------------|------|
| 1  | { }          | => {beer}    | 0.67    | 0.67       | 1.00 |
| 2  | { }          | => {milk}    | 0.83    | 0.83       | 1.00 |
| 3  | { }          | => {diapers} | 0.83    | 0.83       | 1.00 |
| 4  | { }          | => {bread}   | 0.83    | 0.83       | 1.00 |
| 5  | {eggs}       | => {beer}    | 0.17    | 0.50       | 0.75 |
| 6  | {eggs}       | => {milk}    | 0.33    | 1.00       | 1.20 |
| 7  | {milk}       | => {eggs}    | 0.33    | 0.40       | 1.20 |
| 8  | {eggs}       | => {diapers} | 0.33    | 1.00       | 1.20 |
| 9  | {diapers}    | => {eggs}    | 0.33    | 0.40       | 1.20 |
| 10 | {eggs}       | => {bread}   | 0.33    | 1.00       | 1.20 |
| 11 | {bread}      | => {eggs}    | 0.33    | 0.40       | 1.20 |
| 12 | {beer}       | => {milk}    | 0.50    | 0.75       | 0.90 |
| 13 | {milk}       | => {beer}    | 0.50    | 0.60       | 0.90 |
| 14 | {beer}       | => {diapers} | 0.67    | 1.00       | 1.20 |
| 15 | {diapers}    | => {beer}    | 0.67    | 0.80       | 1.20 |
| 16 | {beer}       | => {bread}   | 0.50    | 0.75       | 0.90 |
| 17 | {bread}      | => {beer}    | 0.50    | 0.60       | 0.90 |
| 18 | {milk}       | => {diapers} | 0.67    | 0.80       | 0.96 |
| 19 | {diapers}    | => {milk}    | 0.67    | 0.80       | 0.96 |
| 20 | {milk}       | => {bread}   | 0.83    | 1.00       | 1.20 |
| 21 | {bread}      | => {milk}    | 0.83    | 1.00       | 1.20 |
| 22 | {diapers}    | => {bread}   | 0.67    | 0.80       | 0.96 |
| 23 | {bread}      | => {diapers} | 0.67    | 0.80       | 0.96 |
| 24 | {beer, eggs} | => {milk}    | 0.17    | 1.00       | 1.20 |

|    |                           |              |      |      |      |
|----|---------------------------|--------------|------|------|------|
| 25 | {eggs,milk}               | => {beer}    | 0.17 | 0.50 | 0.75 |
| 26 | {beer,eggs}               | => {diapers} | 0.17 | 1.00 | 1.20 |
| 27 | {diapers,eggs}            | => {beer}    | 0.17 | 0.50 | 0.75 |
| 28 | {beer,eggs}               | => {bread}   | 0.17 | 1.00 | 1.20 |
| 29 | {bread,eggs}              | => {beer}    | 0.17 | 0.50 | 0.75 |
| 30 | {eggs,milk}               | => {diapers} | 0.33 | 1.00 | 1.20 |
| 31 | {diapers,eggs}            | => {milk}    | 0.33 | 1.00 | 1.20 |
| 32 | {diapers,milk}            | => {eggs}    | 0.33 | 0.50 | 1.50 |
| 33 | {eggs,milk}               | => {bread}   | 0.33 | 1.00 | 1.20 |
| 34 | {bread,eggs}              | => {milk}    | 0.33 | 1.00 | 1.20 |
| 35 | {bread,milk}              | => {eggs}    | 0.33 | 0.40 | 1.20 |
| 36 | {diapers,eggs}            | => {bread}   | 0.33 | 1.00 | 1.20 |
| 37 | {bread,eggs}              | => {diapers} | 0.33 | 1.00 | 1.20 |
| 38 | {bread,diapers}           | => {eggs}    | 0.33 | 0.50 | 1.50 |
| 39 | {beer,milk}               | => {diapers} | 0.50 | 1.00 | 1.20 |
| 40 | {beer,diapers}            | => {milk}    | 0.50 | 0.75 | 1.90 |
| 41 | {diapers,milk}            | => {beer}    | 0.50 | 0.75 | 1.12 |
| 42 | {beer,milk}               | => {bread}   | 0.50 | 1.00 | 1.20 |
| 43 | {beer,bread}              | => {milk}    | 0.50 | 1.00 | 1.20 |
| 44 | {bread,milk}              | => {beer}    | 0.50 | 0.60 | 0.90 |
| 45 | {beer,diapers}            | => {bread}   | 0.50 | 0.75 | 0.90 |
| 46 | {beer,bread}              | => {diapers} | 0.50 | 1.00 | 1.20 |
| 47 | {bread,diapers}           | => {beer}    | 0.50 | 0.75 | 1.12 |
| 48 | {diapers,milk}            | => {bread}   | 0.67 | 1.00 | 1.20 |
| 49 | {bread,milk}              | => {diapers} | 0.67 | 0.80 | 0.96 |
| 50 | {bread,diapers}           | => {milk}    | 0.67 | 1.00 | 1.20 |
| 51 | {beer,eggs,milk}          | => {diapers} | 0.17 | 1.00 | 1.20 |
| 52 | {beer,diapers,eggs}       | => {milk}    | 0.17 | 1.00 | 1.20 |
| 53 | {diapers,eggs,milk}       | => {beer}    | 0.17 | 0.50 | 0.75 |
| 54 | {beer,eggs,milk}          | => {bread}   | 0.17 | 1.00 | 1.20 |
| 55 | {beer,bread,eggs}         | => {milk}    | 0.17 | 1.00 | 1.20 |
| 56 | {bread,eggs,milk}         | => {beer}    | 0.17 | 0.50 | 0.75 |
| 57 | {beer,diapers,eggs}       | => {bread}   | 0.17 | 1.00 | 1.20 |
| 58 | {beer,bread,eggs}         | => {diapers} | 0.17 | 1.00 | 1.20 |
| 59 | {bread,diapers,eggs}      | => {beer}    | 0.17 | 0.50 | 0.75 |
| 60 | {diapers,eggs,milk}       | => {bread}   | 0.33 | 1.00 | 1.20 |
| 61 | {bread,eggs,milk}         | => {diapers} | 0.33 | 1.00 | 1.20 |
| 62 | {bread,diapers,eggs}      | => {milk}    | 0.33 | 1.00 | 1.20 |
| 63 | {bread,diapers,milk}      | => {eggs}    | 0.33 | 0.50 | 1.50 |
| 64 | {beer,diapers,milk}       | => {bread}   | 0.50 | 1.00 | 1.20 |
| 65 | {beer,bread,milk}         | => {diapers} | 0.50 | 1.00 | 1.20 |
| 66 | {beer,bread,diapers}      | => {milk}    | 0.50 | 1.00 | 1.20 |
| 67 | {bread,diapers,milk}      | => {beer}    | 0.50 | 0.75 | 1.12 |
| 68 | {beer,diapers,eggs,milk}  | => {bread}   | 0.17 | 1.00 | 1.20 |
| 69 | {beer,bread,eggs,milk}    | => {diapers} | 0.17 | 1.00 | 1.20 |
| 70 | {beer,bread,diapers,eggs} | => {milk}    | 0.17 | 1.00 | 1.20 |
| 71 | {bread,diapers,eggs,milk} | => {beer}    | 0.17 | 0.50 | 0.75 |

## 12.2 k-MEANS CLUSTERING

*Picture this:*

An automobile retailer wishes to open its service centers across a city. They analyze the areas/locales within the city from where they get the maximum service requests/complaints.

1. They need to understand as to how many service centers will have to be opened to service customers in the area.
2. They need to figure out the locations for the service centers within all these areas in such a way that the entire city is covered.

*Another example:* Of late several accidents have been reported. A non-profit organization (NGO) wants to open a series of Emergency-Care wards within a region. They have worked with the police department and the traffic department to come up with a list of all the accident-prone areas in the region. They have to decide the number of Emergency Units to be opened and the location of these Emergency Units, so that all the accident-prone areas are covered in the vicinity of these Emergency Units.

They have a big task cut out for them, that of deciding on the location of these Emergency Units so that the whole region is covered. A clear case wherein *k*-means clustering comes to rescue!

**Why should you learn about clustering?**

1. You wish to offer your services to specific groups.
2. You wish to form groups around data that look similar.
3. You wish to understand the data sets better.
4. You wish to divide data into groups that are meaningful or useful.

**How will you accomplish the above stated?**

The answer is simple – by using the power of clustering. The terms cluster and group can be used interchangeably. Form clusters or groups in such a way that the group/cluster members are more similar than non-group members.

**What is *k*-means clustering?**

A *k*-means clustering means to form *k* groups/clusters. Wikipedia explains it as “*k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster”.

The *k*-means clustering is the simplest, unsupervised learning algorithm. It is unsupervised because one has to only specify number of clusters. *k*-means “learns” the clusters on its own without any information about which cluster an observation belongs to.

Raw data → Pass it through the clustering algorithm → Clusters of data

Given below are the steps in performing *k*-means clustering:

1. Selects *K* centroids. A cluster centroid is the middle of the cluster.
2. Assigns each data point to its closest centroid.
3. Recalculates the centroids as the average of all data points in a cluster (i.e., the centroids are *p*-length mean vectors, where *p* is the number of variables).
4. Assigns data points to their closest centroids.
5. Continues steps 3 and 4 until the observations are not reassigned or the maximum number of iterations (R uses 10 as a default) is reached.

## 12.2.1 Implementation in R

1. You can import data into the Environment as shown below. The name of the file is Cars.txt. This file contains entry for Petrol cars and its corresponding mileage in Kilometers.

The screenshot shows the RStudio interface. The 'Console' tab displays the R session starting with the standard copyright notice. The user then runs the command `> cars = read.table("D:/Cars.txt", header=TRUE)`, which imports a dataset named 'cars'. The 'Environment' tab shows a data frame named 'cars' with 7 observations and 2 variables: 'Petrol' and 'Kilometers'. The 'Data' section of the environment pane shows the following data:

|   | Petrol | Kilometers |
|---|--------|------------|
| 1 | 1.1    | 60         |
| 2 | 6.5    | 20         |
| 3 | 4.2    | 40         |
| 4 | 1.5    | 25         |
| 5 | 7.6    | 15         |
| 6 | 2.0    | 55         |
| 7 | 3.9    | 39         |

2. Apply  $k$ -means algorithm as shown below. The data set is split into 3 clusters and the maximum iteration is 10.

```
> cars3 = kmeans(cars, centers=3, iter.max=10)
> cars3
K-means clustering with 3 clusters of sizes 3, 2, 2

Cluster means:
  Petrol Kilometers
1  5.20    20.0
2  1.55    57.5
3  4.05    39.5

clustering vector:
[1] 2 1 3 1 1 2 3

within cluster sum of squares by cluster:
[1] 71.140 12.905  0.545
  (between_SS / total_SS =  95.3 %)

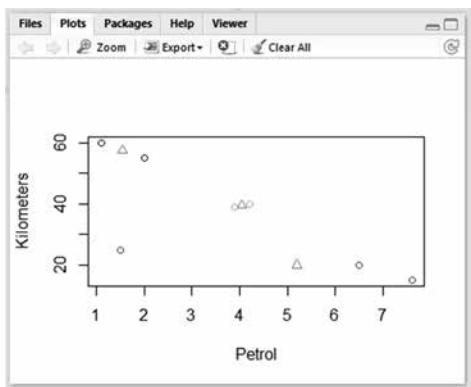
Available components:

[1] "cluster"     "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"   "size"         "iter"
[9] "ifault"
> |
```



3. Next, you can plot clusters as shown below:

```
> plot(cars[cars$cluster ==1, ], col = "red", xlim=c(min(cars[,1]),max(cars[,1])), ylim=c(min(cars[,2]),max(cars[,2])))
>
> points(cars[cars$cluster ==2, ], col ="blue")
>
> points(cars[cars$cluster ==3, ], col ="green")
>
> points(cars3$centers,pch=2,col="orange")
> |
```



## 12.3 DECISION TREE

### **Picture this:**

It is that time of the year again. The college fest is going to be next week. It is going to be a week-long affair. Your friends have started planning on the kind of stalls that they will put up. You too want to try out this stall thing. However, you are yet to decide on what stall you should go for. You have to communicate your decision to the organizing committee in a day's time. The time is short. The decision has to be made quickly. You do not want to end up with a wrong decision. It is your first time at putting up a stall and you want to go for maximum profit.

You have zeroed down your choice to either an ice-cream stall or a burger stall from a gamut of choices available. How about using a decision tree to decide on the same? Let us look at how we can go about creating a decision tree.

**Decision to be made:** Either an ice-cream stall or a burger stall.

**Payoff:** 350\$ in profit if you put up a burger stall and a 400\$ in profit if you put up an ice-cream stall.

### **What are the uncertainties?**

There is a 50% chance of you succeeding to make profit with a burger stall and a 50% chance of you failing at it.

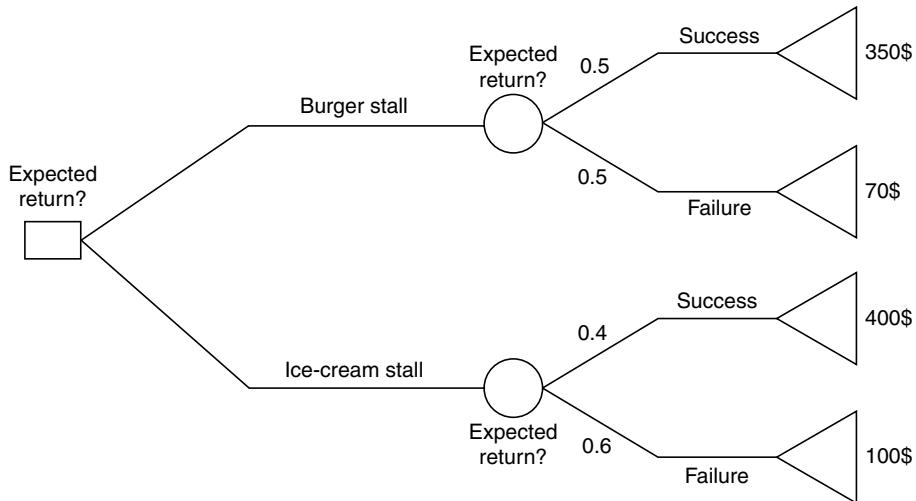
As per the weather forecast, it will be downcast sky and may drizzle or pour slightly throughout the week. Keeping this into consideration, there is a 40% chance of success and 60% chance of failure with an ice-cream stall.

Let us look at the cost of the raw materials:

*For Burger:* 70\$ for the burger buns, the fillings and a microwave oven to keep it warm.

*For Ice-creams:* 100\$ for the cone, the ice-cream and a freezer to keep it cold.

Refer Figure 12.1.



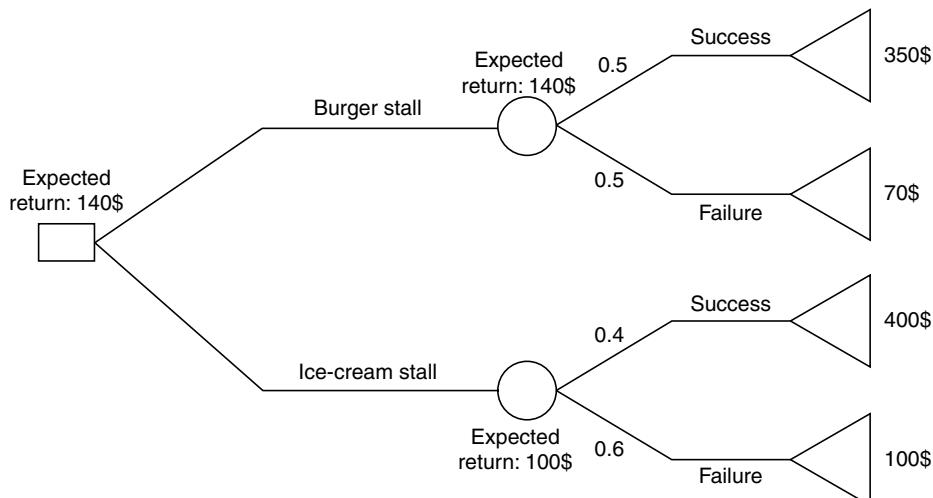
**Figure 12.1** A sample decision tree with expected return value yet to be arrived at.

Let us compute the effective value as per the below formula:

$$\text{Expected value for burger} = 0.5 * 350\$ - 0.5 * 70\$ = 140\$$$

$$\text{Expected value for ice-cream} = 0.4 * 400\$ - 0.6 * 100 = 100\$$$

Refer Figure 12.2.



**Figure 12.2** A sample decision tree with computed expected return.

The choice is obvious. Going by the expected value, you will gain by putting up a burger stall. The expected value does not imply that you will make a profit of 140\$.

Nevertheless, this amount is useful for decision-making, as it will maximize your expected returns in the long run if you continue to use this approach.

### ***Picture this:***

You have just completed writing the script of a romantic story. There are two takers for it.

1. The television network: They are interested in making a daily soap of it that will be telecast on prime time.
2. XYZ Movie Company: They have also shown interest.

You are confused. Should you sell the rights to the TV Network or XYZ Movie Company?

The TV network payout will be a flat 500,000 USD. XYZ Movie Company will pay in accordance to the audience response to the movie.

### ***Payouts and probabilities***

TV Network payout:

Flat Rate: 500,000 USD

### ***XYZ Movie Company Payout:***

Small Box Office: 250,000 USD

Medium Box Office: 600,000 USD

Large Box Office: 800,000 USD

### ***Probabilities:***

P (Small Box Office): 0.3

P (Medium Box Office): 0.5

P (Large Box Office): 0.2

For greater understanding, let us create a payoff table:

| <i>Decisions</i>               | <i>Small Box Office</i> | <i>Medium Box Office</i> | <i>Large Box Office</i> |
|--------------------------------|-------------------------|--------------------------|-------------------------|
| Sign up with TV Network        | 500,000 USD             | 500,000 USD              | 500,000 USD             |
| Sign up with XYZ Movie Company | 250,000 USD             | 600,000 USD              | 800,000 USD             |
| Probabilities                  | 0.3                     | 0.5                      | 0.2                     |

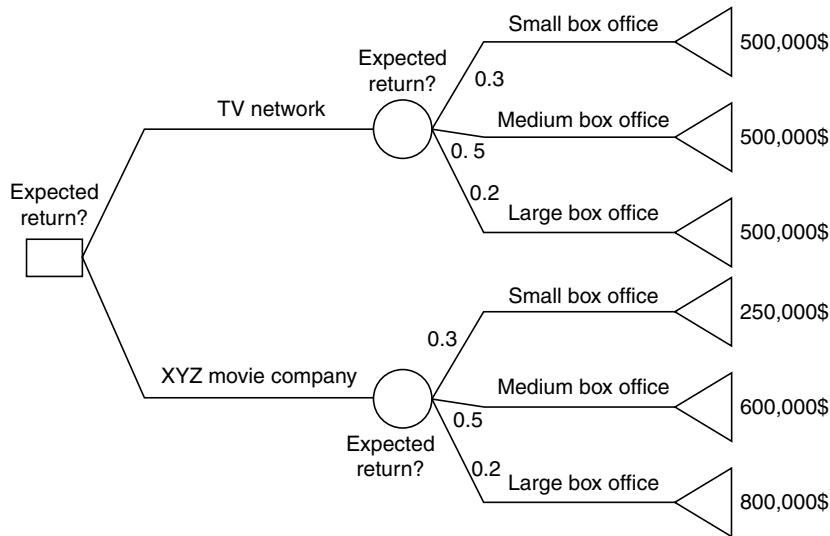
Refer Figure 12.3.

Let us compute the effective value as per the below formula:

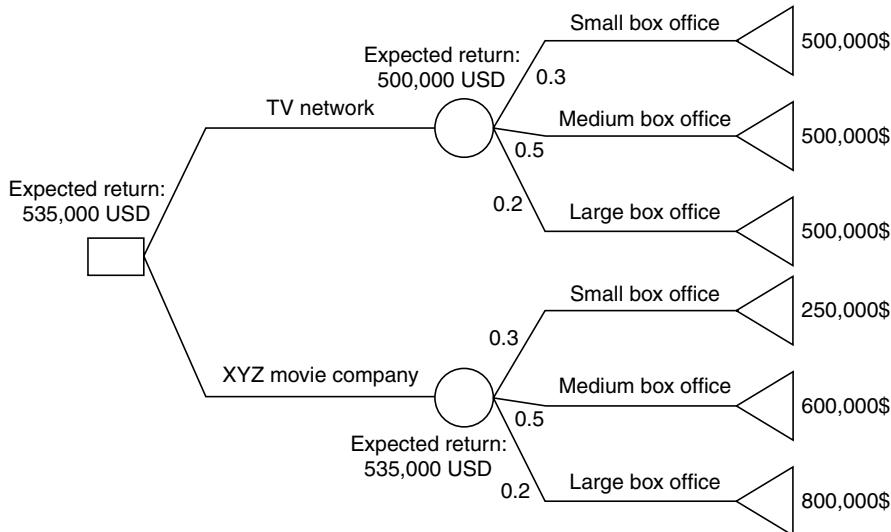
Expected value for TV Network =  $0.3 * 500,000 + 0.5 * 500,000 + 0.2 * 500,000 = 500,000$  USD

Expected value for XYZ Movie

Company =  $0.3 * 250,000 + 0.5 * 600,000 + 0.2 * 800,000 = 535,000$  USD



**Figure 12.3** A sample decision tree with expected return value yet to be arrived at.



**Figure 12.4** A sample decision tree with computed expected return.

The choice is obvious. Going by the expected value, you will gain by selling the rights of your script to XYZ Movie Company.

The expected value does not imply that you will make a profit of 535,000 USD.

### 12.3.1 What is a Decision Tree?

A decision tree is a decision support tool. It uses a tree-like graph to depict decision and their consequences.

The following are the three constituents of a decision tree:

1. **Decision nodes:** commonly represented by squares.
2. **Chance nodes:** represented by circles.
3. **End nodes:** represented by triangles.

### 12.3.2 Where is it Used?

Decision trees are commonly used in operations research, specifically in decision analysis. They are used to zero down on a strategy that is most likely to reach its goals. They can also be used to compute conditional probabilities.

### 12.3.3 Advantages from Using a Decision Tree

1. Easy to interpret.
2. Easy to plot even when there is little hard data. If one is aware of little data such as alternatives, probabilities and costs, it can be plotted and lead to useful insights.
3. Can be easily coupled with other decision techniques.
4. Helps in determining the best, worst and expected value for a given scenario or scenarios.

### 12.3.4 Disadvantages of Decision Trees

1. **Requires experience:** Business owners and managers should have a certain level of good experience to complete the decision tree. It also calls for an understanding of quantitative and statistical analytical techniques.
2. **Incomplete information:** It is difficult to plot a decision tree without having complete information of the business and its operating environment.
3. **Too much information:** Too much information can be overwhelming and lead to what is called as the “paralysis of analysis”.

### 12.3.5 Decision Tree in R

**Step 1:** Load the party package.

```
> library(party)
Loading required package: grid
Loading required package: mvtnorm
Loading required package: modeltools
Loading required package: stats4
Loading required package: strucchange
Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package: base':
  as.Date, as.Date.numeric

Loading required package: sandwich
```

Warning messages:

```
1: package 'party' was built under R version 3. 2. 3
2: package 'mvtnorm' was built under R version 3. 2. 3
3: package 'modeltools' was built under R version 3. 2. 3
4: package 'strucchange' was built under R version 3. 2. 3
5: package 'zoo' was built under R version 3. 2. 3
6: package 'sandwich' was built under R version 3. 2. 3
```

The above command loads the namespace of the package “party” and attaches it on the search list.

### **Step 2:** Check the data set “readingSkills”.

```
> readingSkills[c(1:100), ]
```

|    | nativeSpeaker | age | shoeSize | score    |
|----|---------------|-----|----------|----------|
| 1  | yes           | 5   | 24.83189 | 32.29385 |
| 2  | yes           | 6   | 25.95238 | 36.63105 |
| 3  | no            | 11  | 30.42170 | 49.60593 |
| 4  | yes           | 7   | 28.66450 | 40.28456 |
| 5  | yes           | 11  | 31.88207 | 55.46085 |
| 6  | yes           | 10  | 30.07843 | 52.83124 |
| 7  | no            | 7   | 27.25963 | 34.40229 |
| 8  | yes           | 11  | 30.72398 | 55.52747 |
| 9  | yes           | 5   | 25.64411 | 32.49935 |
| 10 | no            | 7   | 26.69835 | 33.93269 |
| 11 | yes           | 11  | 31.86645 | 55.46876 |
| 12 | yes           | 10  | 29.15575 | 51.34140 |
| 13 | no            | 9   | 29.13156 | 41.77098 |
| 14 | no            | 6   | 26.86513 | 30.03304 |
| 15 | no            | 5   | 24.23420 | 25.62268 |
| 16 | yes           | 6   | 25.67538 | 35.30042 |
| 17 | no            | 5   | 24.86357 | 25.62843 |
| 18 | no            | 6   | 26.15357 | 30.76591 |
| 19 | no            | 9   | 27.82057 | 41.93846 |
| 20 | yes           | 5   | 24.86766 | 31.69986 |
| 21 | no            | 6   | 25.21054 | 30.37086 |
| 22 | no            | 6   | 27.36395 | 29.29951 |
| 23 | no            | 8   | 28.66429 | 38.08837 |
| 24 | yes           | 9   | 29.98455 | 48.62986 |
| 25 | yes           | 10  | 30.84168 | 52.41079 |
| 26 | no            | 7   | 26.80696 | 34.18835 |
| 27 | yes           | 6   | 26.88768 | 35.34583 |
| 28 | yes           | 8   | 28.42650 | 43.72037 |
| 29 | no            | 11  | 31.71159 | 48.67965 |
| 30 | yes           | 8   | 27.77712 | 44.14728 |
| 31 | yes           | 9   | 28.88452 | 48.69638 |
| 32 | yes           | 7   | 26.66743 | 39.65520 |
| 33 | no            | 9   | 28.91362 | 41.79739 |
| 34 | no            | 9   | 27.88048 | 42.42195 |
| 35 | yes           | 7   | 25.64581 | 39.70293 |

|    |     |    |          |          |
|----|-----|----|----------|----------|
| 36 | yes | 8  | 27.71701 | 44.06255 |
| 37 | no  | 7  | 25.18567 | 34.27840 |
| 38 | yes | 11 | 30.78970 | 55.98101 |
| 39 | yes | 11 | 30.75664 | 55.86037 |
| 40 | yes | 11 | 30.51397 | 56.60820 |
| 41 | no  | 5  | 26.23732 | 26.18401 |
| 42 | no  | 5  | 24.36030 | 25.36158 |
| 43 | no  | 7  | 27.60571 | 32.88146 |
| 44 | no  | 10 | 29.64754 | 45.76171 |
| 45 | yes | 8  | 29.49313 | 43.48726 |
| 46 | yes | 7  | 26.92283 | 38.91425 |
| 47 | yes | 8  | 28.35511 | 44.99324 |
| 48 | no  | 6  | 26.10433 | 29.35036 |
| 49 | yes | 8  | 29.63552 | 43.66695 |
| 50 | yes | 8  | 27.25306 | 43.68387 |
| 51 | no  | 8  | 26.22137 | 37.74103 |
| 52 | yes | 6  | 26.12942 | 36.26278 |
| 53 | no  | 9  | 30.46199 | 42.50194 |
| 54 | no  | 7  | 27.81342 | 34.33921 |
| 55 | yes | 10 | 29.37199 | 52.83951 |
| 56 | yes | 10 | 29.34366 | 51.94718 |
| 57 | yes | 7  | 25.46308 | 39.52239 |
| 58 | no  | 10 | 28.77307 | 45.85540 |
| 59 | no  | 11 | 30.35263 | 50.02399 |
| 60 | no  | 8  | 29.32793 | 37.52172 |
| 61 | yes | 10 | 28.87461 | 51.53771 |
| 62 | no  | 7  | 26.62042 | 33.96623 |
| 63 | no  | 7  | 28.11487 | 33.39622 |
| 64 | no  | 11 | 30.98741 | 50.28310 |
| 65 | yes | 10 | 29.25488 | 50.80650 |
| 66 | yes | 5  | 24.54372 | 31.95700 |
| 67 | no  | 8  | 26.99163 | 37.61791 |
| 68 | no  | 11 | 30.26624 | 50.22454 |
| 69 | no  | 7  | 27.86489 | 34.20965 |
| 70 | yes | 10 | 30.16982 | 52.16763 |
| 71 | yes | 7  | 25.53495 | 40.24965 |
| 72 | no  | 7  | 26.75747 | 34.72458 |
| 73 | yes | 10 | 29.62773 | 51.47984 |
| 74 | no  | 5  | 24.41493 | 25.32841 |
| 75 | no  | 9  | 30.64056 | 42.88392 |
| 76 | yes | 7  | 26.78045 | 39.36539 |
| 77 | yes | 8  | 28.51236 | 43.69140 |
| 78 | yes | 5  | 23.68071 | 32.33290 |
| 79 | no  | 7  | 26.75671 | 33.12978 |
| 80 | no  | 10 | 29.65228 | 47.08507 |
| 81 | no  | 9  | 29.33337 | 41.29804 |
| 82 | no  | 9  | 26.47543 | 29.52375 |
| 83 | no  | 9  | 28.35925 | 41.92929 |
| 84 | no  | 8  | 27.15459 | 38.30587 |

|     |     |    |          |          |
|-----|-----|----|----------|----------|
| 85  | no  | 10 | 30.58496 | 45.20211 |
| 86  | yes | 9  | 30.08234 | 48.72401 |
| 87  | no  | 9  | 28.34494 | 42.42763 |
| 88  | yes | 11 | 29.25025 | 55.98533 |
| 89  | yes | 9  | 28.21583 | 48.18957 |
| 90  | no  | 8  | 28.10878 | 37.39201 |
| 91  | no  | 8  | 26.78507 | 37.40460 |
| 92  | yes | 10 | 31.09258 | 51.95836 |
| 93  | no  | 5  | 24.29214 | 26.37935 |
| 94  | no  | 7  | 27.03635 | 33.52986 |
| 95  | yes | 7  | 24.92221 | 40.19923 |
| 96  | no  | 6  | 27.22615 | 29.54096 |
| 97  | yes | 7  | 25.61014 | 41.15145 |
| 98  | yes | 10 | 28.44878 | 52.57931 |
| 99  | yes | 7  | 27.60034 | 40.01064 |
| 100 | yes | 11 | 31.97305 | 56.71151 |

“readingSkills” is a toy data set which exhibits a spurious/false correlation between a child’s shoe size and the score in his/her reading skills. It has a total of 200 observations on 4 variables, namely, native-Speaker, age, shoeSize and score. The explanation for the variables are as follows:

- nativeSpeaker: A factor that can have a value of yes or no. “yes” indicates that the child is a native speaker of the language in the reading test.
- age: Age of the child.
- shoeSize: This variable stores the shoe size of the child in cms.
- score: This variable has the raw score of the child in the reading test.

**Step 3:** Create a data frame “Inputdata” and have it store from 1 to 105 records of the “readingSkills” data set.

```
> InputData <- readingSkills[c(1:105), ]
```

The above command extracts out a subset of the observations in “readingSkills” and places it in the data frame “InputData”.

**Step 4:** Give the chart file a name.

```
> png(file = "decision_tree.png")
```

“decision\_tree.png” is the name of the output file. With this command, a plot device is opened and nothing is returned to the R interpreter.

**Step 5:** Create the tree.

```
> OutputTree <- ctree(
+ nativeSpeaker ~ age + shoeSize + score,
+ data = InputData)
```

ctree is the conditional inference tree. We have supplied two inputs: The first being the formula that is a symbolic description of the model to be fit; the second input “data” is to specify the data frame containing the variables in the model.

**Step 6:** Check out the content of “OutputTree”.

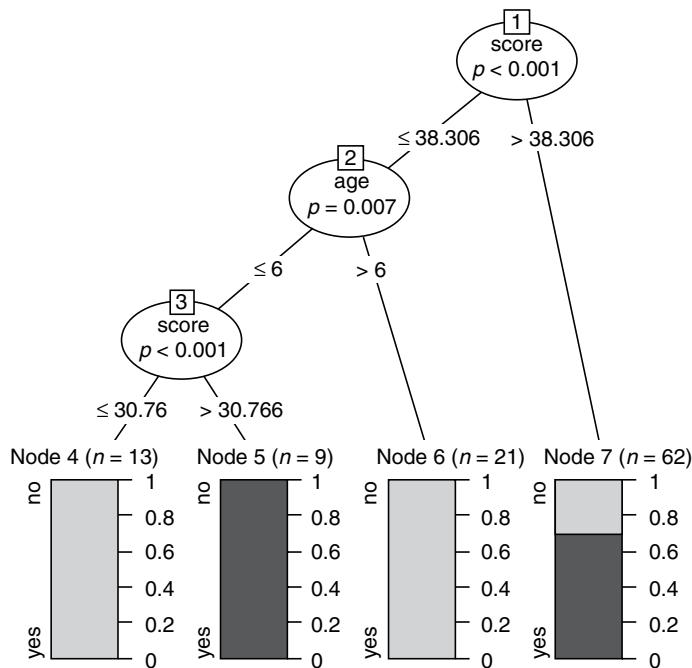
```
> OutputTree
      Conditional inference tree with 4 terminal nodes
Response: nativeSpeaker
Inputs: age, shoeSize, score
Number of observations: 105

1) score <=38.30587; criterion = 1, statistic = 24.932
   2) age <= 6; criterion = 0.993, statistic = 9.361
      3) score <= 30.76591; criterion = 0.999, statistic = 14.093
         4) * weights = 13
      3) score > 30.76591
         5) *weights = 9
   2) age > 6
      6) *weights = 21
1) score > 38.30587
   7) *weights = 62
```

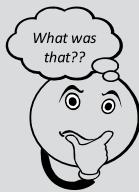
**Step 7:** Save the file.

```
> dev.off()
null device
1
```

This command is to shut down the specified device “png” in our example.  
The output from the whole exercise is as follows:



**Inference:** Anyone with a reading score  $\leq 38.306$  and age greater than 6 is NOT a native speaker.



### Remind Me (Forget Me Not!)

- Association rule mining is an example for item-based recommendation.
- Association analysis is mostly done based on an algorithm named “Apriori Algorithm”.
- $k$ -means clustering is the simplest, unsupervised learning algorithm. It is unsupervised because one has to only specify number of clusters.  $k$ -means “learns” the clusters on its own without any information about which cluster an observation belongs to.
- Decision tree is a decision support system and used in operation research



### Point Me (Books)

- *A Programmer’s Guide to Data Mining* by Ron Zacharski.
- Michael Berry and Gordon Linoff, *Mastering Data Mining*, John Wiley & Sons, 2000.
- K. Cios, W. Pedrycz, R. Swiniarski, and L. Kurgan, *Data Mining: A Knowledge Discovery Approach*, Springer, ISBN: 978-0-387-33333-5, 2007.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Verlag, 2001.



### Connect Me (Internet Resources)

- <https://www.coursera.org/course/ml>: Coursera “Machine Learning” course from Stanford
- [http://www.cs.cmu.edu/~wcohen/collab-filtering-tutorial.ppt?bcsi\\_scan\\_c9e2df1c0800a1cc=OVLGC8gZhmJbc+8+tEic9Ufto+IYAA=AAT1v/gg==:1](http://www.cs.cmu.edu/~wcohen/collab-filtering-tutorial.ppt?bcsi_scan_c9e2df1c0800a1cc=OVLGC8gZhmJbc+8+tEic9Ufto+IYAA=AAT1v/gg==:1)
- <https://www.coursera.org/learn/r-programming>: Coursera “R Programming” course from John Hopkins University



*Test Me*

#### A. Fill in the Blanks

1. Decision tree is a \_\_\_\_\_ method.
2. R is \_\_\_\_\_ tool.
3. \_\_\_\_\_ is a user-based recommendation algorithm.
4.  $k$ -mean splits data set in to \_\_\_\_\_ of cluster.

#### Answers:

1. Supervised learning
2. Statistical
3. Collaborative filtering
4. Fixed number

## UNSOLVED EXERCISES

1. Write an R program to implement Association Mining for frequently used item set. (Hint: You can construct your own data set.)
2. Write an R program to implement  $k$ -Means Clustering for a data set. (Hint: Download public data set from the Internet.)



# 13



## BI Road Ahead

---

### BRIEF CONTENTS

|                               |                                       |
|-------------------------------|---------------------------------------|
| What's in Store               | Business Intelligence for ERP Systems |
| Understanding BI and Mobility | Social CRM and BI                     |
| BI and Cloud Computing        | Unsolved Exercises                    |

---

### WHAT'S IN STORE

By now you are already familiar with the concepts relating to Business Intelligence (BI), i.e., data warehousing and the fundamentals of business analytics. With this background it's time to look ahead at various new possibilities in the field of BI, their applications and merits over the existing technologies.

In this chapter we will look ahead at the evolution of BI with mobility, cloud computing, ERP, and Social CRM. This chapter is a “Must Read” for those interested to learn about BI in depth and about its new horizon of possibilities.

We suggest you refer to some of the learning resources suggested at the end of this chapter and also complete the “Test Me” exercises. You will get deeper knowledge by interacting with people who have shared their project experiences in blogs. We suggest you make your own notes/bookmarks while reading through the chapter.

---

### 13.1 UNDERSTANDING BI AND MOBILITY

Business Intelligence (BI), as a concept, is not new. Nor is the practice of storing data in databases, analyzing that data, and revealing useful information out of it in the form of reports or through more

advanced data visualization techniques. BI has leveraged business extensively over the past two decades through ever-improving data management practices and through new technologies that together comprise what is now called the DSS or Decision Support System.

Parallel to the development of BI technologies and methodologies continued rapid research and new innovations in mobile technology. Mobile technology offered a solution to people who wanted to do things on the move. Mobility had two major offerings that became its major selling points:

- **24×7 connectivity:** Ability to stay in contact with others (mobile phones, wireless Internet, etc.) even when travelling or away from office/home.
- **Mobile workability:** The convenience of being able to work (using laptops, smartphones, etc.) from anywhere.

### 13.1.1 The Need for Business Intelligence on the Move

With the ever-increasing volumes of enterprise data (that began to run into thousands of terabytes!) coupled with the fast-paced world of modern business, intelligent decisions needed to be taken much faster. This meant that there was a need for better and faster transfer of information extracted by decision support systems to the people who consumed that information, i.e. the managerial and administrative-level population. No longer did they want to be limited to their office-based desktops to access data and useful information through DSSs and applications designed only for PCs.

Fortunately, the pioneers in the field of BI and analytics were people with good foresight. They saw the huge potential in the rapidly developing area of mobile technologies. What could be better than the power to be able to view performance metric reports, KPIs, etc. anywhere, anytime on your hand-held mobile device, and hence make quicker decisions for your business? As soon as this was thought up, research began in this area. Mobility gave business people an option to access real time data and make immediate decisions. Today, the enormous progress in the field of mobile BI and analytics can be seen all around.

Let us follow the gradual progress of BI mobility over the years.

### 13.1.2 BI Mobility Timeline

#### *The Antediluvian Era*

- Initially, BI was generally delivered to the end-users by a system that consisted of a wired local computer network. BI applications on a computer would connect to a database on the network and provide information through a channel, such as a web browser, or certain other software.
- Later, as mobile devices such as pagers and mobile phones came into the picture, they could receive data that was pushed through SMS service. However, not only would such messages contain a very limited and minimal amount of information, they also would not give the user any interactivity at all.
- BI applications designed for mobile devices were primitive, cumbersome to develop, and their maintenance cost a lot. For the time and money spent on them, the information they delivered was too less.

### ***Taking Up the Challenge***

- Then came the era of smartphones, and with them came various applications designed to read tables of data and a few charts too. However, there were still a few things that bothered their users – small screens that couldn't provide highly detailed information on charts, incapable mobile browsers, poor connectivity, etc.
- Laptops were a better alternative to smartphones as far as the need for better data visualization was concerned. However, the advantage with smartphones was that they were smaller, lighter, and hence less cumbersome to carry around. If data visualization could be improved on smartphones, history would be written.

### ***The Roadblocks***

- Small screen for viewing reports and KPIs; hence lack of detail.
- Poor resolution.
- Poor connectivity.
- Poorly equipped browsers.
- Information is sent to recipients on fixed schedule. However, recipient has no control over it.
- Small amount of transmitted data.
- Limited user interactivity: no analysis capability; no drill-down; no drill-through to sub-reports; no freedom to query either.
- Low on memory and processing power.
- Very limited functionality on keyboard.

...and many more like these.

### *What does one expect from mobile business intelligence technology?*

According to Catriona McGauchie, who wrote an article on mobile BI at [www.dashboardinsight.com](http://www.dashboardinsight.com), there are three major expectations from the adoption of mobile BI technology:

- *Device maturity*, i.e. the extent of the quality of information that the mobile device can show the user.
- *End-user expectations*, i.e. user-friendliness, user-interactivity, compatibility of mobile applications with desktop applications.
- *Connectivity* should be robust and secure.

### ***Overcoming the Shortcomings***

- In the 2000s, Blackberry smartphones began to establish their stronghold over the corporate and governmental market. Their key selling points were:
  - Wireless email.
  - Larger screens.
  - Advanced browser.
  - Advanced MOS (mobile operating system).
  - QWERTY keyboard and thumbwheel that provided better user-interactivity.
  - A native application for the device that is specially designed for the mobile screen. This provides superior interactivity.

- Another way of approaching the mini-computer-cum-mobile phone concept was to make laptop computers smaller, lighter, and more easily portable. The ultimate aim was to combine the advantages of smartphones (small size, lightweight, portability) with the computing power of laptops.

### ***The Present: A Result of Endless Research and Efforts***

As per [www.dashboardinsight.com](http://www.dashboardinsight.com), MBI (Mobile BI) generally speaks of three usage models:

- **Exceptions and alerts:** At times some events may happen which were unexpected or unaccounted for. The concerned user should be alerted about such events that may hamper the business promptly. For example, if a delivery/shipment is delayed somehow, the sales executive must receive an alert about it on his mobile for immediate action.
- **Push reporting:** Sometimes KPIs, or certain other specific reports, are pushed to executives on a regular basis according to a pre-determined schedule. Decision to push such reports to the user is taken at the source. It could be daily (at EOD), weekly (e.g. every Friday), or monthly too. For example, every Monday morning (8:30 am to be precise), a report on the sales performance figures of last week is delivered to the senior marketing and sales executives.
- **Pull reporting:** Here, the decision to pull/generate a report is initiated by the end-user. The user gives inputs through his mobile device and can ask for information from a central server-based system. For example, an executive uses an application to get the list of his top 5 sales executives.

### ***Mobile Devices/Applications as of Today***

- **MOS (Mobile Operating Systems):** Blackberry OS, Windows Phone, Symbian, Android, etc. are today the most popular mobile operating systems available in the market. They are best known for their flexibility, operability, and compatibility with desktop OSs. Also, it is easy to develop mobile applications for these MOSSs. Currently, the most popular MOS is Android. Over 80,000 applications are available for mobile devices operating on Android OS.
- **Apple iPhone:** iPhone is a revolutionary device that set a new standard and a greater level of expectation from mobile devices. iPhone became hugely popular and sold many units within the first few months of its release into the market. Apple released the software development kit for building apps that can run natively on iPhone and iPad.
- **Apple iPad:** iPad came to be hailed as the harbinger of the future of mobile computing. It combines the portability of a smartphone with the computational power and boasts of a larger screen and more interactive display of a computer. iPhone and iPad have transformed the way data is viewed on mobile devices. BI applications can now generate reports that can be converted into mobile dashboards for delivery to an iPhone or iPad. iPad is virtually a laptop without a physical keyboard (though that functionality can be accessed using the on-screen virtual keyboard). Besides, the iPad OS and applications are highly compatible with desktop applications. Navigating through KPIs and dashboards using touch screen is user-intuitive and user-friendly, which is why it appeals to the majority of the market.
- **Examples of small BI mobile apps dealing with local data analysis:**
  - A simple example of mobile BI would be the calorimeter and pedometer apps on your Sony Erickson cellphone. The phone has a motion sensor apparatus within it. When you go jogging,

the phone is capable of measuring the distance you cover through the pedometer app. The phone also measures the heart-rate (if you are holding it in your hand). The calorimeter app then combines the data from the pedometer with the heart-rate data and calculates the approximate number of calories you have burnt during your jog using a complex algorithm. The accuracy of the app is usually about 95%, which is acceptable enough for the purpose that you use it for. In this example, the mobile phone collects real time data by itself and uses BI applications to deliver information to the user. So, the data source and application system are on the same device.

- Another example would be the iPod touch's "music suggestions" functionality. It scans your entire music collection, recognizes the genres of music you listen to, and if the iPod is connected to the Internet (through its Wi-Fi), it can suggest you songs that you may find to your taste.
- Yet another example would be the iPhone. It provides a feature of map integration that shows the path you have taken for your jog/walk and graphically provides the gradient/pace changes using GPS capabilities.
- Various KPIs and dashboards are now designed specifically for the smartphone, iPad, and other mobile devices. They include user-interactivity features (like drill-down) too. If the dashboard is too wide for the screen, and you wish to view an individual sub-report on the dashboard in more detail, then you can do so using the drill-through functionality. Clicking/selecting a certain report on the dashboard will open up that report only over the entire screen. However, you can see the report up to a much greater level of detail.
- Blackberry smartphones are capable of displaying a variety of reports such as column charts, pie-charts, drillable tables, KPIs, trend graphs, etc. that are custom-designed for smartphone displays.

### 13.1.3 Data Security Concerns for Mobile BI

According to an article by Maribel D. Lopez on <http://us.blackberry.com>, data security must be provided at three levels:

- **Device security:** Best to let the source data stay on centralized servers rather than on individual mobile devices. That way, if the device is lost, the data is still secure. Also, access to the data center is only permitted within the network. Most mobile device manufacturers today provide encryption for email and phone memory, antivirus and firewall software, etc.
- **Transmission security:** Since the information is transmitted wirelessly to mobile devices, and it also generally involves third-party members in the network, data security during transmission becomes a top priority. Some measures taken to ensure this are SSL (secure sockets layer), VPN (virtual private network) connection, cryptographic encryption of data transmitted, etc.
- **Authorization, authentication, and network security:** It refers to controlling which user can access which data by assigning access privileges to users through IDs and/or passwords, all stored in an encrypted database storing user credentials.

## 13.2 BI AND CLOUD COMPUTING

### 13.2.1 What is Cloud Computing?

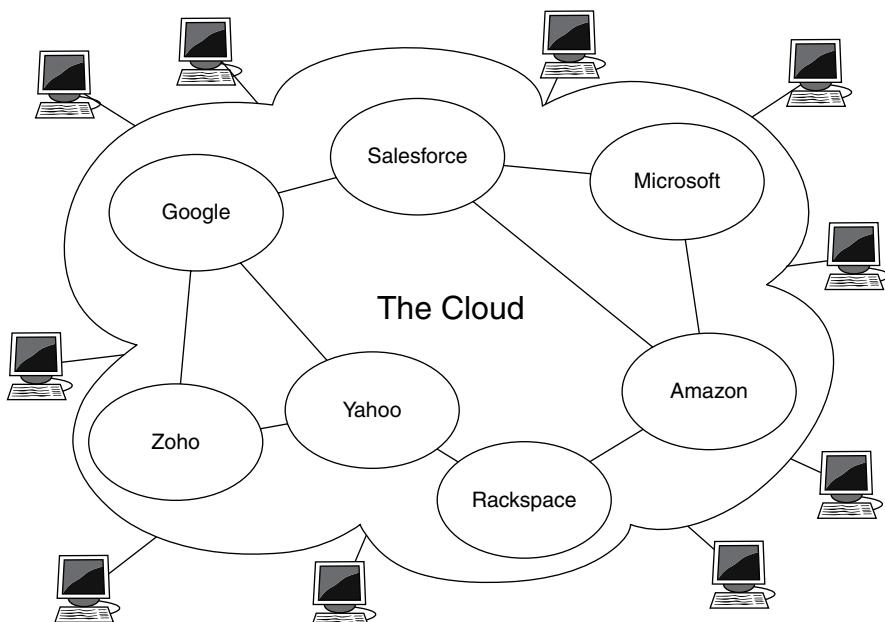
Let us consider the job of an executive, Mike, in a large organization. As an executive, he is responsible to make sure that all the employees have the right hardware and software to do their jobs. Buying

computers for every employee wouldn't suffice; he will also have to purchase software licences to provide employees the tools with which to perform their job. Whenever a new employee joins the company, the executive has to purchase additional software licences or make sure that the present software licence can be used by this new recruit. Mike's is a growing organization, with recruits joining the company almost every alternate fortnight. We agree that making available the requisite hardware and software for all the new employees is indeed a stressful job for Mike.

Wouldn't it be nice if instead of installing a suite of software for each computer, we have to load just one application on the system? That application would allow employees to log into a Web-based service which hosts the required programs that would enable the user to do his or her job. This remote server would be owned by another company, and it would be their responsibility to make available everything from word processing to email to complex data analysis programs. The world calls this arrangement cloud computing.

Cloud computing can be defined as location-independent computing, whereby shared servers provide data, software, and services to computers and other devices as and when required. These services are offered through data centers all over the world, which collectively are referred to as the "cloud". Any user who has access to the Internet can use the cloud and the services provided by it. Since all these services are connected, users can share information between multiple systems as well as with other users.

In a cloud computing system, there's a significant shift of workload. Local computers need not take the heavy load when it comes to running applications. The network of computers that forms the cloud handles the load instead. Due to this, hardware and software demands reduce on the user's side. The only thing the user's computer needs to run the cloud computing systems interface software is a Web browser, and the cloud's network takes care of everything else.



**Figure 13.1** Some prominent users of cloud computing.

A familiar example of cloud computing is any Web-based email service such as Hotmail, Yahoo! Mail, or Gmail. Instead of running an email program on computer, we log in to a Web email account on a remote system. The software and storage for the account doesn't exist on our computer; it's on the cloud computer of the service provider.

Another example is that of Google Docs. Through Google Docs we can have multiple people collaborating on the same document in real time. It is also possible to share these documents with people outside your organization. When it comes to public cloud computing, this is one of the more basic examples.

### 13.2.2 Why Cloud Computing?

The answer is simple: Rapid implementation, ease of use, and subscription pricing. Cloud computing customers generally do not set up their own physical infrastructure; instead they go for renting usage from a third-party provider to reduce capital expenditure. They use resources as a service, and they only pay for the part that they have used. This model is analogous to traditional utility services such as electricity or water. However, some cloud service providers' bill on subscription basis. The cloud is becoming very popular with small and medium enterprises (SMEs) because it is difficult for them to afford the large capital expenditure of traditional IT.

Some benefits of using cloud computing are

- Software as a subscription.
- Reduced software maintenance.
- Increased reliability.
- Increased scalability.
- Cost reduction.
- Matches current computing trends.
- Portability/accessibility.
- Efficient use of computer resources.
- Version-less software.
- Environment-friendly.
- Pay per use.

What are the barriers to adoption of cloud-based applications and platforms? There seem to be two prime concerns: data privacy and data security. Most of the enterprises are still hesitant to move their critical business data off their premises.

### 13.2.3 Why Business Intelligence should be on the Cloud?

In today's economy, smart businesses are trying to utilize every available opportunity to increase their performance and reduce cost. BI tools are thus continuously gaining popularity among companies of all sizes as they try to increase efficiency and effectiveness and thus gain a competitive edge. While large investments in conventional BI solutions are impractical and unattractive to most businesses, the popularity of software-as-a-service or cloud BI solutions is increasing enormously. Consider a scenario where a business unit of an IT company (let us call it "A") received an ETL (extract, transform, load) project for building a data warehouse, which requires 10 people to work for one month on an ETL tool, the licences of which are very costly. The business unit "A" purchases the licence for the same and starts

working on the project. Now assume that after some time another business unit (let us call it "B") of the same company bids for and gets another ETL project which requires 20 people to work for two months on another ETL tool. The business unit "B" is now required to purchase the licences for this ETL tool which again leads to a huge investment. Also after the completion of the respective projects, there is no use of the licences unless another project with a similar requirement arrives. The result: a huge wastage of resources and capital investment. This is where the cloud comes to the rescue. Instead of purchasing licences, it is a better option for the company to set up a private cloud and allow all its business units to use the software present on the cloud as a service as and when required, and pay for it on the basis of usage. This is a more efficient and cost effective way of using resources.

Cloud BI has a great potential to substantially disrupt the existing BI market because of its low cost, flexibility, and scalability. Cloud BI solutions offer all the benefits of traditional BI solutions while substantially reducing the investment required. Cloud BI solutions not only provide powerful and flexible business insight, but are also faster, easier, and cheaper than traditional BI solutions. The benefits of cloud BI are appealing and real. As compared to traditional business intelligence, cloud BI offers the following business benefits:

- **Increased access, maximum results:** Traditional BI solutions are costly and require IT resources. These limitations restrict their availability to professional analysts. On the other hand, cloud solutions are easier and less costly to deploy, and also require little expertise to operate. As a result, they are more accessible to non-technical users. Analysts and others who have to struggle with Excel for making sales forecasts and resource management, and for servicing customer accounts can make use of cloud BI tools to perform the analysis and for quick and easy reporting. There would be no need to install the required BI tool on their systems; instead, they can just access the cloud and conduct their work. This helps them to discover opportunities for performance improvements that are hidden in the data using BI.
- **Faster ROI:** Cloud BI offers a quick deployment. Unlike traditional BI implementation, which may take 12 to 18 months or more, cloud BI solutions can typically be setup in just few weeks. This is due to the reason that there is no extra hardware required to be installed and no database to be set. With the solution up and running in a short time, companies can start getting a return on their investment (ROI) quickly. Also the maintenance and customization is faster and easier. Since the hardware and infrastructure are maintained by the vendor, software upgrades and architectural changes are also handled by the vendor and delivered to the customer automatically.
- **Lower implementation costs:** Cloud BI providers manage the entire infrastructure required for their service as well as for hosting their applications, so we are spared the hardware and setup costs required to deploy a BI solution. Also, as software-as-a-service BI solutions can be set up in a small time as compared to traditional solutions, the time and resources required for a finished solution are drastically reduced.
- **Lower on-going costs:** In cloud BI solutions, there are no servers to maintain, no on-going software maintenance, no patches to be installed, and minimal IT resources are required. This leads to low on-going costs. Cloud BI vendors generally charge a subscription fee which is decided according to the application service, maintenance, and support. These subscriptions are usually based on the number of users accessing the cloud, the amount of data analyzed, and the software usage, and are much lower as compared to the cost of purchasing a conventional on-premises BI solution. This billing approach makes sure that customers have to pay only for what they need or what they use. Hence the customer retains financial control of the project thus maintaining the flexibility to scale up as the needs expand. As the customer's solution is running on a shared infrastructure, this increased financial control and flexibility comes at lower cost.

- **Scalability:** Cloud solutions are made to support a large number of users simultaneously. This means an enterprise can expand its cloud solution easily and quickly just by requesting a larger account size or access for more users. Unlike on-premise BI solutions, cloud solutions can be expanded without buying more hardware or installing different software. Since the vendor is responsible for capacity, organizations can begin with a small number of users and a small set of data, and can later increase the number as and when required.
- **Flexibility:** Unlike traditional solutions, cloud BI solutions are more flexible. A cloud BI solution can be easily changed. So it is easy for non-technical users to quickly add new reports and dashboards, data sources, and analysis. On the contrary, traditional BI solutions would take weeks or months to change and will also involve significant IT resources.
- **Greater visibility:** Cloud applications are run over the Internet, so a user is able to share data with others easily, both inside as well as outside the organization. The users can integrate data from various sources in different parts of the world, from other business units, and also from partners of the organization. This is important for any firm which has multiple sites at various locations.
- **Data warehouse on cloud:** Cloud BI also provides a lot of data warehousing options to its customers, from SaaS (software-as-a-service) or DaaS (data-as-a-service) offerings, which provide software, to PaaS (platform-as-a-service) and IaaS (infrastructure-as-a-service) solutions on which you can build your data warehouse. The cloud has proved to be a fertile ground for managing large volumes of data. Once the data is on the cloud, it is easy and quick to access it and also it is not location-specific, i.e. it can be accessed from any location.

With all the above-stated advantages, cloud BI has proved to be a more beneficial and better solution than traditional on-premise BI solutions. These are reasons enough for the world to move from traditional BI solutions towards cloud BI. Cloud computing allows one to focus on their key competencies and worry less about the IT infrastructure cost.

### 13.3 BUSINESS INTELLIGENCE FOR ERP SYSTEMS

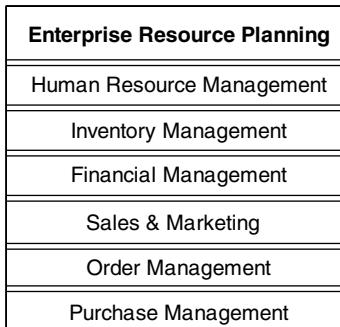
BI and ERP grew as two separate entities, but their coming together is akin to bringing two sides of the critical business coin together.

#### Picture this...

“PlusSales”, a typical manufacturing enterprise, has a chaotic mix of business applications. While a few of these applications exist in silos, a few others are tied together with the help of complex interface programs. Owing to some applications existence in silos, there is a huge probability that the same data may exist at more than one place, raising questions about the accuracy and consistency of data. The senior management of “PlusSales” felt the need for a system that could integrate and automate the business processes from end to end, for example from planning to manufacturing to sales. Enter the ERP software.... The ERP system provides several business benefits. Here, we enumerate the top three:

- Consistency and reliability of data across the various units of the organization.
- Streamlining the transactional process.
- A few basic reports to serve the operational (day-to-day) needs.

Figure 13.2 depicts a typical ERP system.



**Figure 13.2** A typical ERP system.

### 13.3.1 Why BI in ERP?

The mammoth growth in the volume of data today is compelling organizations all over the world to start harnessing the power of data to make better and faster decisions. An ERP system was able to solve some, if not all, the information needs of the organization. It was able to successfully integrate several business processes across the organization's supply chain. It was adept at capturing, storing, and moving data across the various units smoothly. What it lacked was the ability to integrate data from other existing applications and external sources, which is imperative to serve the analytical and reporting needs of the organization. Let us explain this further with the help of our example of the company "PlusSales".

"PlusSales" has humungous amount of data in a multitude of systems. Most, if not all, of the organization's financial data is safely housed in the ERP system. But Kevin, a senior executive, looking to gain a quick view of how the business is progressing requires not only the financial data but also the data from sales, inventory, CRM, etc. His needs can be satisfied from a merged data set that combines data from all the systems. Business Intelligence could do just that, i.e. it can include data from internal and external sources; it can support information access to external parties, vendors, and customers; it can provide real time actionable business insights; it can provide single version of the critical information; it can support the analytics and reporting needs of the organization; and much more.

One way to bring order to the chaos prevalent in the organization is by an extension of ERP to BI. It is time to think of BI as a layer which sits on top of or is embedded within ERP and other applications which stockpile giant repositories of data. Today is an era of an extended enterprise. The term "extended enterprise" implies that not just the employees, members of the board, managers, and executives make a company but a company also comprises its business partners and investors, its suppliers and vendors, and even its customers and prospects. The extended enterprise can only be successful if all of the component groups and individuals have the information they need in order to do business effectively. Refer to Figure 13.3.

When we say that the ERP systems alone cannot cater to the reporting and analytical needs of the organization, it is not to connote that the ERP systems cannot generate reports, but the built-in standard reports of the ERP systems are very basic and pretty generic across industries. The ERP systems initially were using data from live operational and transactional systems in queries and reports. One problem they posed was that most of their reports were hard-coded. Hence the stage was set for BI. Few of the forward-thinking and progressive ERP vendors started off by adding useful and powerful analytical and reporting capabilities to their suite to add on to the value provided to their customers. One such vendor was SAP which came out with SAP BW (Business Warehouse) in 1997.

| Enterprise Resource Planning | Extended Enterprise   |
|------------------------------|-----------------------|
| Human Resource Management    | Partners & Investors  |
| Inventory Management         | Mobile Access         |
| Financial Management         | Business Intelligence |
| Sales & Marketing            | Enterprise Logistics  |
| Order Management             | Customers & Prospects |
| Purchase Management          | Vendors & Suppliers   |
|                              | Managers & Executives |

**Figure 13.3** Components of an extended enterprise and ERP.

### 13.3.2 Benefits of BI in ERP

- ERP today is considered as “Trusted Data Source” and hence basing decisions on ERP data is considered more “Safe” by decision makers.
- Employs customized analytics to meet business needs.
- Enables users to perform “what if” forecasting.
- Enables users to develop customized reports.
- Presents key information in dashboard views using charts, pivots, gauges, etc.
- Drills down to view data at a more detailed level.
- Drills across from dimension to dimension at the same hierarchical level.
- Merges information from multiple varied systems to provide a “unified source” of data.
- Performs trend analysis using historical data.
- Searches for hidden patterns.

There are packaged BI solutions available with the ERP tools for implementing the key components required for a BI solution. Let us look at the key components of a BI solution:

- **Reporting tool:** One of the key strengths of BI solutions is to provide a robust front-end that allows users to view and analyze data in a dashboard view (summary level) and then drill-through/drill-across into data elements to launch detailed reports. Many BI solutions also provide ready-made customizable templates for users to create their own reports.
- **ETL tool:** It is a known fact that the data model for ERP is dramatically different from the data model for BI. ETL (extract, transform, and load) tools packaged with ERP solutions can be leveraged to transform, load, and merge external data for a comprehensive BI solution. There are pre-packaged scripts available for ETL which can be customized and implemented for data cleansing and merging.

### 13.3.3 ERP Plus BI Equals More Value

Realizing the potential of BI to leverage ERP systems for performance improvement, a few ERP vendors such as Oracle, SAP [SAP NetWeaver 7.0; ERP and BI in one system: MCOS (Multiple Components One System)], PeopleSoft, etc., have already ventured to provide BI support with ERP. There are several BI tools and applications available in the market, and it will be interesting to see how they will be able

to unlock the data that is available in the ERP systems, integrate it with the data from other internal and external sources and support the analytics and reporting needs of the organization.

But this is not as easy as it sounds. ERP and BI are essentially different entities which came into existence for different purposes, but can they join hands to leverage the power of each other. There is the need to transform data to information and information to intelligence. ERP can transform data into information, but BI tools are required to complete the transformation from information to intelligence.

### **13.4 SOCIAL CRM AND BI**

---

Just the other day, I happened to be with one of my friends when she received a call from a shop in an uptown mall informing her that the particular style of clothing is available with them and she should make the time to check it out. I asked my friend if she had placed one such order and left her contact number with them. She answered in the negative and simply said that she was a regular at their mall and they know her for quite some time now.

That was the beginning of the understanding the one-to-one relationship that the shop had managed to forge with its customers. This had certainly not happened overnight. Let us zero down to four essentials that could have improved the shop's customer relationship and thereby its business:

- Notice/observe the customer's requirements, habits, choices and preferences, etc.
- Remember the customer's behavior over time.
- Learn from the past interactions with its customers.
- Act on what it has learned to make customers more profitable.

A small business will allow the liberty to notice/observe all its customers. This comfort, however, is neither affordable nor possible with large firms/enterprises. This is where technology comes to the rescue. Let us explore further on the four essentials of customer relationship stated above.

- **Notice/observe the customer's requirements, habits, choices and preferences, etc.:** The details of each and every interaction with the customer can be recorded using a transaction processing system. These recordings might have been done for operational needs of the firm but few can debate the fact that these are the customer touchpoints where information about the customer behavior first makes its way into the enterprise.
- **Remember the customer's behavior over time:** The transaction processing system can collect humungous amount of data, but this enormous amount of data will simply remain data mounds if it is not carefully cleaned, sorted, merged, organized, and summarized. That leads to the data warehouse. Information gleaned from multiple disparate sources is stored in a data warehouse in a friendlier format than that maintained by operational systems. The data warehouse thus allows the bigger firms/enterprises to remember the customer's behavior over time. In other words, the data warehouse serves as the enterprise's memory.
- **Learn from the past interactions with its customers:** Now is the time to apply intelligence to memory. This will help recognize patterns, propose hypothesis, accept or dismiss hypothesis, make predictions, etc. The data warehouse is used to capture the differing needs, preferences, choices, propensities, etc. of the customers. Data mining allows corroborating questions such as "Which promotional scheme will work best for which customer segment?", "Who will remain a loyal customer and vouch for the product or service offered?", "Who will also prefer product Z?", etc. In other words, data mining is about skimming through the colossal data to find patterns. The task is arduous, more so because the signals sent out by customers are not always very clean.

In conclusion, although it is easier for small firms/organizations to know their customers, it is not impossible for large firms/organizations to learn about their customers and build a strong relationship with them. Customer Relationship Management (CRM) systems have been around for about 20 years now and are used by big and small organizations alike. Few examples of popular CRM systems are: Salesforce, Zoho, and Sugar CRM.

In traditional CRM one essentially starts out with information on a list of people/companies whom you know, how you know them, when and how you have interacted with them, etc. The next step is to classify or categorize people in your CRM tool as leads, friends, prospects, etc. that helps you define who that person is and how you know him/her. So now, you have data about the people/companies whom you know and you use that data to help manage your relationship with them. What is the advantage of having a CRM system in place? CRM systems are designed to create a process around the interaction that your company has with its customers in the hope of more efficiently closing a sale or resolving some sort of an issue.

Social CRM requires dealing with conversations and relationships with social customers in addition to the data or information that you might have about them. These conversations and relationships take place not just from the company to the consumer but also from consumer to consumer.

Think of Facebook and Twitter (the social media). Facebook has more than 500 million registered users who spend more than 3 billion hours a month on the Facebook site. It's a veritable interaction hub, where many businesses have a significant presence. The requirement is now for the companies to use Facebook in a way that's integrated with their other communication channels.

Let us look at Twitter as another example. Assume you are a large brand on Twitter such as XYZ Airlines. You are in the process of building relationships with your followers. At the same time you have the opportunity to build relationships with and listen to (and engage) the customers having conversations about you. Traditional CRM didn't work with Twitter or Facebook or with any other social platform; it was just a collection of data and information. So again, the big difference between CRM and Social CRM is that we now have all these conversations and relationships to consider.

Paul Greenberg, a leader in Social CRM, defines it as:

*CRM is a philosophy and a business strategy, supported by a technology platform, business rules, workflow, processes and social characteristics, designed to engage the customer in a collaborative conversation in order to provide mutually beneficial value in a trusted and transparent business environment. It's the company's response to the customer's ownership of the conversation.*

The time is right for the organizations to consider the following:

- How must the organization take action based on the conversations and relationships that it fosters or engages in with its customers?
- How to structure the organization in a way that is both efficient and scalable to take advantage of Social CRM?
- How to take all of the unstructured data from the social web and structure it in a way that allows you to get actionable insight?
- How to use Social CRM to empower your customers and grow your customer base?

There is a huge opportunity in this space and BI vendors are starting to see the growing opportunity. Enterprise business intelligence tools play a key role in successfully integrating and measuring the unstructured and semi-structured data that drives the social space. A few vendors such as SAS, SAP Oracle, IBM, DataFlux, Information Builders, etc. have already started addressing the challenges that Social CRM poses.



## Remind Me

- Mobility had two major offerings that became its major selling points:
    - **24 × 7 connectivity:** Ability to stay in contact with others (mobile phones, wireless Internet, etc.) even when travelling or away from office/home.
    - **Mobile workability:** The convenience of being able to work (using laptops, smartphones, etc.) from anywhere.
  - There are three major expectations from the adoption of mobile BI technology:
    - Device maturity, i.e. the extent of the quality of information that the mobile device can show the user.
    - End-user expectations, i.e. user-friendliness, user-interactivity, compatibility of mobile applications with desktop applications.
    - Connectivity should be robust and secure.
  - Mobile devices and mobile software/applications in the market, as of today:
    - Mobile operating system (MOS)
    - Apple iPad
    - Apple iPhone
  - Cloud computing can be defined as location-independent computing, whereby shared servers provide data, software, and services to computers and other devices as and when required.
  - A familiar example of cloud computing is any Web-based email service like Hotmail, Yahoo! Mail, or Gmail.
  - Why cloud computing? The answer is simple: Rapid implementation, ease of use, and subscription pricing.
  - Some benefits of using cloud computing are
    - Software as a subscription.
    - Reduced software maintenance.
  - Increased reliability.
  - Increased scalability.
  - Cost reduction.
  - Pay per use.
- Cloud solutions are made to support a large number of users simultaneously.
- Barriers to cloud-based solutions and platforms:
  - Two prime concerns: Data privacy and security. Most of the enterprises are still hesitant to move their critical business data off their premises.
- ETL tools available with ERP systems can be leveraged to extract, transform, and load external data for a comprehensive BI solution.
- BI enables users to perform “what if” analysis.
- ERP systems cannot cater to the analytics and reporting needs of the organization, but together with BI they can support analytics and reporting.
- ERP can transform data into information, but BI tools are required to complete the transformation from information to intelligence.
- ERP plus BI equals more value.
- Social CRM requires dealing with conversations and relationships with social customers in addition to the data or information that you might have about them. These conversations and relationships take place not just from the company to the consumer but also from consumer to consumer.
- A few vendors such as SAS, SAP Oracle, IBM, DataFlux, Information Builders, etc. have already started addressing the challenges that Social CRM poses.



## Connect Me (Internet Resources)

- <http://www.information-management.com/issues/20010601/3492-1.html>
- <http://www.information-management.com/infodirect/20000721/2499-1.html>
- <http://www.information-management.com/issues/20000701/2352-1.html>
- [http://www.cio.com/article/498904/ERP\\_and\\_BI\\_A\\_Match\\_Made\\_in\\_Heavy\\_If\\_You\\_re\\_in\\_Data\\_Hell](http://www.cio.com/article/498904/ERP_and_BI_A_Match_Made_in_Heavy_If_You_re_in_Data_Hell)



## Test Me Exercises

### Fill me

1. To further enhance its ERP package with BI capabilities, SAP came up with \_\_\_\_\_.
2. BI supports SAP in meeting the \_\_\_\_\_ and \_\_\_\_\_ needs of the organization.
3. ERP software provides several business benefits. For example, it provides \_\_\_\_\_ and \_\_\_\_\_ of data across various units of the organization.
4. ERP stands for \_\_\_\_\_ \_\_\_\_\_ \_\_\_\_\_.
5. Few examples of popular CRM systems include \_\_\_\_\_, \_\_\_\_\_, and \_\_\_\_\_.
6. Social CRM requires dealing with \_\_\_\_\_ and \_\_\_\_\_ with social customers.

7. Some benefits of using cloud computing are: \_\_\_\_\_, \_\_\_\_\_, \_\_\_\_\_, \_\_\_\_\_, etc.

### Solution:

1. SAP BW (Business Warehouse)
2. Analytics and reporting
3. Consistency and reliability
4. Enterprise Resource Planning
5. Salesforce, Zoho, Sugar CRM
6. Conversations and relationships
7. Software as a subscription, reduced software maintenance, increased reliability, increased scalability, cost reduction

### Match me

| Column A                 | Column B                         |
|--------------------------|----------------------------------|
| It is for data input     | Basic operational reports        |
| It is for data retrieval | Advanced analytics and reporting |
| BI                       | ERP                              |
| ERP                      | BI                               |
| Charts, gauges, matrix   | Dashboard                        |

**Solution:**

| <i>Column A</i>          | <i>Column B</i>                  |
|--------------------------|----------------------------------|
| It is for data input     | ERP                              |
| It is for data retrieval | BI                               |
| BI                       | Advanced analytics and reporting |
| ERP                      | Basic operational reports        |
| Charts, gauges, matrix   | Dashboard                        |

**UNSOLVED EXERCISES**

1. “BI + ERP = increased value”. Explain giving example.
2. What are the key data security concerns for mobile BI? Explain.
3. Why is it important for Business Intelligence space to take into consideration the Social CRM? Explain.
4. “BI on the cloud”. Why do you think this is important? Explain.
5. What are the benefits of BI in ERP? Explain.
6. What according to you are the differences between ERP and BI? Explain.
7. List a few popular CRM tools.
8. What are the benefits of integrating social media information into BI applications? What are the drawbacks?
9. Will organizations that embrace social BI be at a competitive advantage compared to companies that don’t?
10. Given the variability of freeform text, how can social media information be used?



# Glossary

**Attribute:** The literal meaning is quality, characteristic, trait, or feature. In the context of RDBMS (Relational Database Management System), a set of attributes are used to describe an entity. An entity is a physical or abstract object about which we can store information. The relationship between entities is depicted using an ER (Entity Relationship) model. While physically implementing the ER model, an entity gets converted into a table/relation and the attributes get converted into columns or fields. For example, think about a bank system having several entities such as “Loan”, “Account”, “Customers”, etc. Let us describe the entity, “Customers” using attributes such as CustomerID (10 digit integer, mandatory field), CustomerEmailID (text variable length, case sensitive, typical format x@y), CustomerName, CustomerAddress, etc.

*Also refer to terms: Entities, ER (Entity Relationship) model, Relation and Relationship.*

**Balanced Scorecard (BSC):** It is a strategic performance management framework used by organizations to measure and communicate their business performance. The concept of BSC was introduced by Dr. Robert S. Kaplan and David P. Norton. It suggests business performance to be looked at from four perspectives: (i) financial – how do we look to our shareholders? (ii) customer – how do we look to our customers? (iii) internal processes – what processes we must excel at? (iv) learning and growth – what should be done to build the competencies of the employee base?

*Also refer to terms: Dashboards, KPIs, Metrics, KRIs.*

**Business Analytics:** The analysis of enterprise data from multiple perspectives to gauge and enhance the performance of business. Business analytics is heavily dependent on data. For its successful implementation, business analytics requires a high volume of high quality data. The challenge faced by business analytics remains the storage, integration, reconciliation of data from multiple disparate sources across several business functions and the continuous updates to the data warehouse.

*Also refer to term: BI component framework.*

**Business Driver:** This is a term used in the Management field to describe the influence of a particular function or activity on business performance. Business drivers are the force that causes/propels the businesses to move/change in a certain way. Examples: changing economy, changing technology, changing workforce, changing labor laws, changing competitive markets, etc. In the context of

“Business Intelligence component framework”, business driver is one of the four components analyzed to gather data-related requirements. (Business requirements can be studied with reference to *business drivers, business goals, business strategies, and business tactics.*)

*Also refer to term: BI component framework.*

**Business Goal:** In the context of “BI component framework”, business goal refers to the objectives that businesses set out to achieve such as increase in customer base, increased market share, increase in productivity, increase in profitability, retention of employees, etc.

**Business Intelligence (BI):** Business Intelligence is about making available the right information in the right format at the right time to the right decision makers. It supports fact-based decision making. It focuses on ensuring “single version of truth”. BI helps provide 360 degree perspective on the business. According to Howard Dresner (he coined the term Business Intelligence in 1989), BI is a set of concepts and methodologies to improve decision making in business through use of facts and fact-based systems.

*Also refer to terms: Business component framework, Data warehouse, Data marts, Data mining, and Business analytics.*

**Business Rules:** This is a term typically used in the area of Business Process Management (BPM) and workflow automation. Traditionally, business rules or business logic is closely tied-to program logic. Businesses have seen huge program maintenance load as and when formula or computation logic changes. Hence the industry evolved techniques to represent business rules outside the program and have program code dynamically take the latest business rule for computation or workflow. Example: Insurance claims processing workflow is associated with many rules for computing eligible claim amount. These rules may change often times based on competitive products, government regulations, new business models, and so on.

**Business Strategies:** Strategy is an approach designed to achieve a particular goal. In the context of BI component framework, it is the methodology adopted to accomplish business goals such as use of technology to achieve the business goal of increase in productivity. Example: to achieve the business goal of employee retention, businesses can look at initiatives such as employee welfare program, etc.

*Also refer to term: BI component framework.*

**Business Users:** The users who leverage IT applications to conduct business operations or evolve new strategies. In the context of BI, business users or business executives request IT for data to make informed decisions.

*Also refer to terms: Casual users, Power users.*

**Business Value:** Typically, business value is measured in terms of ROI (Return on Investment), ROA (Return on Asset), TCO (Total Cost of Ownership), TVO (Total Value of Ownership), etc. Example: ROI is the returns that accrue from capital investment, and also includes the operational expenses. ROA is the returns that accrue from capital investment only. TCO is the total cost incurred from the date of purchase of an asset to the date of retirement. TVO differs from TCO in that it considers the benefits of alternative investments. It is a comparative measure that evaluates the TCO and any additional benefits, such as the mobility of laptops when compared to desktop computers.

*Also refer to terms: Business drivers, Business strategies, Business intelligence, BI component framework.*

**Cardinality of Relation:** In the context of RDBMS, cardinality of a relation is the number of records/tuples in a table/relation. Example: A table “Sample” with a cardinality of four and degree three implies that the table has four records/tuples/rows and three attribute/columns/fields.

*Also refer to terms: Attributes, Entity, Relation, Relationship, RDBMS, ER model, Cardinality of relationship.*

**Cardinality of Relationship:** The cardinality of a relationship defines the type of relationship between two participating entities. In other words, the cardinality of a relationship specifies how many instances of an entity relate to one instance of another entity. There are four types of cardinality relationships. They are (i) one-to-one, (ii) one-to-many, (iii) many-to-one, (iv) many-to-many.

Please note that the cardinality of a relationship differs from the cardinality of a relation. The cardinality of a relation is the number of records/tuples in a relation.

*Also refer to terms: Attributes, Entity, Relation, Relationship, RDBMS, ER model, Cardinality of relation.*

**Casual Users:** In the context of BI, casual users are information consumers. Examples of casual users are: executives, senior management, etc. They work with pre-defined reports produced by power users.

*Also refer to term: Power users.*

**Conceptual Data Model:** In the context of RDBMS, a conceptual data model implies the identification of entities and the relationships between the entities. Example: let us look at a college scenario where a teacher can be a part of only one department but a department can have  $n$  number of teachers belonging to it. In this case, there are two entities: (i) a teacher entity and (ii) a department entity. The relationship or the cardinality of the relationship is  $1:N$  between Department and Teacher entities.

*Also refer to terms: Data model, Entity relationship model, Dimensional data model, Logical data model, Physical data model.*

**Constraints:** The literal meaning is restriction, limitation. In RDBMS, constraints imply the restrictions imposed on the column(s) of a table. Example: a NOT NULL constraint on a column implies that it is mandatory to read in a value in the column. A Primary Key constraint on a column implies that the column can take in only unique and not null values into it. A Foreign Key constraint on a column implies that the column can only take in values existing in the primary key column of its own table or another table to which it refers to. A foreign key column can take in a null value or a duplicate value. A Unique constraint on a column implies that the column cannot have duplicate values. The column however can take in null values.

*Also refer to term: Data quality.*

**Critical Success Factor (CSF):** This is a term commonly used in management to focus attention on activities that must get done flawlessly in order to meet the business goal. The CSFs will take into consideration constraints such as staffing, financial investments, competency development that influences the act of meeting goals positively.

*Also refer to terms: Balanced scorecard, Dashboard.*

**Cube:** The OLAP tools allow the user to turn data stored in relational databases into meaningful, easy to navigate business information by rearranging data in multidimensional format termed as a cube. The dimensions of a cube represent distinct categories for analyzing business data. Categories such as time, geography, or product line breakdowns are typical cube dimensions. Example: The sales amount

of a retail store being analyzed along the time dimension, region dimension, and product category dimension.

*Also refer to term: Dimensional data modeling.*

**Database:** A database is an organized collection of interrelated data. Data in the database: (i) is integrated; (ii) can be shared; (iii) can be concurrently accessed.

*Also refer to terms: Attribute, Entity, Relation, Relationship, RDBMS.*

**Dashboard:** The term has come from the automobile industry. Just like a vehicle dashboard gives a clear idea about the status of the car such as the speed at which the car is being driven, the fuel indicator that indicates the amount of fuel, a corporate dashboard allows one to feel the pulse of the enterprise. A look at the dashboard indicates the “pulse” in key areas of business performance and also alerts decision makers in comparison with thresholds.

*Also refer to terms: KPIs, Balanced scorecard.*

**Data Governance:** In the context of BI component framework, data governance is about proper storage, maintenance, management of enterprise data to ensure quality. Data governance is a quality regime that includes ensuring accuracy, consistency, completeness, and accountability of data. There are policies that govern the use of data in an organization. This is done primarily to secure data from hackers and data from inadvertently leaking out. Data governance also ensures compliance with regulations. It helps to define standards that are required to maintain data quality. The distribution of roles for governance of data is as follows:

- Data ownership
- Data stewardship
- Data custodianship

*Also refer to terms: Enterprise, Data integrity constraint.*

**Data Integrity Constraints:** These are the constraints imposed to ensure data integrity in RDBMS. Example: Entity Integrity Constraint (Primary Key Constraint), Referential Integrity Constraint (Foreign Key Constraint), etc. An entity integrity constraint on a column implies that the column can take in only unique and not null values into it. A Foreign Key constraint on a column implies that the column can only take in values existing in the primary key column of its own table or another table to which it refers to. A foreign key column can take in a null value or a duplicate value. A Unique constraint on a column implies that the column cannot have duplicate values. The column, however, can take in null value.

*Also refer to terms: Data integrity constraints, Data quality.*

**Data Mart:** In the context of BI, a data mart is a focused subset of a data warehouse that deals with a single area of data and is organized for quick analysis. A data mart can be dependent or independent. A dependent data mart is sourced from data from the enterprise-wide data warehouse. An independent data mart is sourced directly from the OLTP systems.

*Also refer to terms: Data warehouse, Database, Data mining.*

**Data Mining:** In the context of BI, data mining is the process of processing large volumes of data stored in the data warehouse, searching for patterns and relationships within that data.

*Also refer to terms: Data warehouse, Database, Data marts.*

**Data Model:** It is a conceptual data tool to describe data, data relationships, data semantics, and consistency constraints. There are two popular data models:

- **Object Based Logical Model:** ER Model.
- **Record Based Logical Model:**
  - **Hierarchical Data Model:** Example: IMS.
  - **Network Model:** Example: IDMS.
  - **Relational Data Model:** Relational data model uses a collection of tables (relations) to represent data and the relationships among those data. Example: Oracle, Sybase.

*Also refer to terms: Conceptual model, Logical model, Physical model.*

**Data Quality:** It is about accuracy, consistency, completeness, and timeliness of data. Example: Let us consider a list of participants attending a training program. The participants have been asked to provide their name, email id, and address for further communication. All the participants clearly provide the asked details. The data here is accurate. A couple of participants, however, do not provide their address. The data here is incomplete. This whole process took two days time but the training coordinator required this information in a day's time as he wanted the information to reach the printer who would then print out participation certificates for the participants. As the information was delayed, the certificates could not be handed over to the participants. This is about the timeliness of data.

*Also refer to terms: Constraints, Data governance, Data integrity constraint.*

**Data Warehouse:** In terms of BI, data warehouse is a repository which stores integrated information from multiple disparate sources for efficient querying and analysis. The key characteristics of data warehouse are: (i) subject-centric, (ii) integrated, (iii) non-volatile, (iv) time-invariant. There are two approaches to building a data warehouse: (i) the top-down approach given by Bill Inmon and (ii) the bottom-up approach given by Ralph Kimball.

*Also refer to terms: Data marts, Database, Data mining.*

**DDL (Data Definition Language):** In the context of RDBMS, it allows a user to define and alter the structure and the organization of the data to be stored and the relationships among the stored data items. A few common DDL statements are: Create table < table name > ..., Drop Table < table name > ...., Create Index ...., Drop Index ...., etc.

*Also refer to terms: Database, Data model, Entity relation, DML.*

**Dimension Table:** In the context of multidimensional modeling, a dimension table contains attributes that describe fact records on the fact table. Some of these attributes provide descriptive information; others are used to specify how fact table data should be summarized to provide useful information to the analyst. Refer to Chapter 7 for more details.

*Also refer to terms: Dimensional model, Star schema, Snowflake schema, Fact.*

**Dimensional Model:** A data model considered apt for OLAP. In the context of dimensional modeling, it is also known as Star schema because in dimensional modeling there is a large central fact table with many dimensional tables surrounding it.

*Also refer to terms: Dimensional table, Star schema, Snowflake schema, Fact.*

**DML (Data Manipulation Language):** In the context of RDBMS, DML statements let a user or application program update the database by allowing adding new data, deleting the existing data, and modifying the existing data. A few common DML statements are: Insert into table ..., delete from table ..., update table ... , select \* from table....

*Also refer to terms: Database, Data model, Entity relation, DDL.*

**DSS (Decision Support System):** An IT application system that supports the decision making process of business managers and leaders. Example: Customers at a car showroom use an application to provide details such as model of the car, power engine, diesel or petrol, automatic transmission or manual transmission, mileage etc, and the application can provide them with the range of cars (to select from) with the specified features. This application supports the decision making process of the customers.

*Also refer to terms: EIS, ERP, BI.*

**Enterprise:** A Company or Organization. A typical enterprise will have functions such as Human Resources, Sales, Marketing, Finance, Production, Information Technology (IT), etc.

**Executive Information System (EIS):** A system that supports the decision making process of strategic managers.

*Also refer to terms: DSS, ERP, BI.*

**Entity:** It is an object or concept about which business user wants to store information. Example: A project idea is an entity. A business user can store information about the project such as project code, project manager, project location from where the project will be executed, the number of project team members, the project start date, the project end date, etc.

*Also refer to terms: Attributes, ER (Entity Relationship) model, Relation, Relationship.*

**ER (Entity Relationship) Model:** Modeling the databases using ER diagrams is called ER Modeling. It is one of the many ways to represent business findings in pictorial format. ER model helps to identify all entities and the relationships that exist between the entities.

*Also refer to terms: Attributes, Entities, Relation, Relationship.*

**Enterprise Resource Planning (ERP):** This refers to a class of packaged software application that automates business processes of several important functions of a business enterprise. ERP provides unified data store for functions such as Finance, HR, Purchase, Sales, Marketing, and so on. ERP can support with some pre-defined reports and interfaces needed for ad hoc reporting or the analytical needs of the enterprise.

*Also refer to terms: DSS, ERP, EIS.*

**Extraction:** In building a data warehouse, extractions refers to the process of collecting/obtaining data from multiple disparate data sources. The data sources are varied. Example: Data could come from .txt file/.xls file/.csv file/any relational database. Data sources could also be in different geographical locations.

*Also refer to terms: Data marts, Database, Data warehouse, Data mining, Transformation, Loading.*

**Fact:** In the context of multidimensional modeling, a fact is a measure. It is usually a numeric value that can be aggregated. Example: SalesAmount, UnitQuantity, etc.

*Also refer to terms: Dimensional table, Dimensional model, Star schema, Snowflake schema.*

**Fact Constellation:** The constellation schema is shaped like a constellation of stars (i.e. star schemas). This is more complex than star or snowflake schema variations, as it contains multiple fact tables. This allows the dimension tables to be shared among the various fact tables. It is also called “galaxy schema”. The main disadvantage of the fact constellation is more complicated design because multiple aggregations must be taken into consideration. Refer to Chapter 7 for more details.

*Also refer to terms: Star schema, Snowflake schema, Dimensional model, Dimensional table, Fact.*

**Fact Table:** In the context of multidimensional modeling, a fact is a measure. It is usually a numeric value that can be aggregated. Example: SalesAmount, UnitQuantity, etc. It is central to a Star or Snowflake schema, and captures the data that measures the organization’s business operations. It usually contains large number of rows.

*Also refer to terms: Star schema, Snowflake schema, Dimensional model, Dimensional table, Fact.*

**Grains:** In the context of multidimensional modeling, given the hierarchy as Year → Quarter → Month → Week → Day, Day is said to be the grain. It refers to level at which data is summarized.

*Also refer to terms: Star schema, Snowflake schema, Dimensional model, Dimensional table, Fact, Hierarchies.*

**Hierarchies:** In the context of multidimensional modeling, an example of a hierarchy is Year → Quarter → Month → Week → Day. Year is said to be at the highest level in the hierarchy and Day is said to be at the lowest level in the hierarchy.

*Also refer to terms: Star schema, Snowflake schema, Dimensional model, Dimensional table, Grains.*

**HOLAP (Hybrid On-Line Analytical Processing):** In the context of OLAP (on-line analytical processing), this model combines the features of ROLAP (Relational On-Line Analytical Processing) and MOLAP (Multidimensional On-Line Analytical) processing. This model has the scalability feature of relational tables and the multiple perspectives of a cube.

*Also refer to terms: OLTP, OLAP, ROLAP, MOLAP.*

**Key Performance Indicators (KPIs):** KPIs are important business health indicators used to steer the team member actions to meet business goals. KPIs influence the way business team members do their jobs, approach their daily work, and deal with alerts. KPIs help people focus on the “big picture”.

*Also refer to terms: balanced scorecard, measures, metrics and KRIs.*

**Key Result Indicators (KRIs):** These tell us how we have done. Example: A feedback of 4.5 on a scale of 5 (with 5 being the highest) for measuring the effectiveness of the learning program implies that the learning program was effective.

*Also refer to terms: Balanced scorecard, Measures, Metrics, KPIs.*

**Loading:** In building a data warehouse, loading refers to the process of loading cleansed, corrected, and transformed data into the data warehouse or data mart. It is the third step in building a data warehouse. The first step is extracting the data from multiple disparate sources and the second is transforming the data to have it in alignment with a universal data warehouse format.

*Also refer to terms: Data warehouse, Data mart, Data mining, Extraction, Transformation.*

**Metadata:** It is the information about the data. In other words, it is data about data. This is the layer of the data warehouse which stores the information like the name and type of data sources, data about the transformation process, date and time of extraction, target databases, date and time of data loading, etc. There are four categories of metadata: application metadata, business metadata, process metadata, and technical metadata.

*Also refer to terms: Database, Data warehouse, Data mart.*

**Metrics:** Metrics refer to a **system of measures**/facts based on standard unit of measurement with a business context. In business management terms, metrics are used to track business performance in numeric terms. Examples include employee attrition rate, product defect rate, frequency of goods returned, etc.

*Also refer to terms: Balanced scorecard, Measures, KPIs, KRIs.*

**MOLAP (Multidimensional On-Line Analytical Processing):** In the context of OLAP (On-Line Analytical Processing), this data model helps view the data in the form of a cube. MOLAP requires pre-computation and storage of information in the cube. The advantages of using MOLAP are: (i) Fast query performance due to optimized storage; (ii) multidimensional indexing and caching; (iii) smaller on-disk size of data compared to relational databases.

*Also refer to terms: OLTP, OLAP, ROLAP, MOLAP.*

**Normalization:** In the context of RDBMS, it is a process followed to organize data such that the redundancy of data is kept to the minimum possible level. It is a refinement process. It helps in removing anomalies present in insert, update and delete operations. Few normal forms are 1NF, 2NF, 3NF, BCNF, etc.

*Also refer to terms: ER model, Database, OLTP, OLAP, Dimensional model.*

**Object Based Logical Data Model:** ER model is a widely known object-based logical model. It is used to describe data at the conceptual and the view level. The ER model is based on the perception of the real world that consists of a collection of basic objects called entities, and of relationships among these objects.

*Also refer to terms: Database, ER model, Dimensional model, Entity, Attributes.*

**ODS (Operational Data Store):** In the context of data warehouse, ODS is a database to store data extracted and integrated from multiple sources for additional operations on the data. The data may be passed back and forth to operational systems for updates and to the data warehouse for reporting. In other words, it is a repository where clean operational data of the enterprise is placed. It helps to answer ad hoc queries for operational decision making.

*Also refer to terms: OLTP, OLAP, Database, Data warehouse, Data marts, Staging.*

**OLAP (On-Line Analytical Processing):** In OLAP, data is held in dimensional form rather than relational form. OLAP's life blood is multidimensional data. OLAP tools are based on multidimensional data model. The multidimensional data model views data in the form of a data cube. Example: The sales figure of a retail outlet analyzed along the product categories, time, and region dimension.

*Also refer to terms: OLTP, ROLAP, MOLAP, HOLAP.*

**OLTP (On-Line Transactional Processing):** OLTP systems refer to a class of systems that manage transaction-oriented applications. These applications are mainly concerned with the entry, storage, update, and retrieval of data. Example: Airline/railway ticket reservation, point of sale system (POS) at a supermarket store.

*Also refer to terms: OLAP, ROLAP, MOLAP, HOLAP.*

**Physical Data Model:** Physical data model is a representation of how the model will be built in the database. A physical database model will exhibit all the table structures, including the column names, the columns data type, the column constraints, primary key, foreign key, and the relationships between tables. Example: A project table.

| Column Name      | Data type and length | Constraints |
|------------------|----------------------|-------------|
| ProjectCode      | Varchar(5)           | Primary Key |
| ProjectName      | Varchar(30)          | Not Null    |
| ProjectLocation  | Varchar(30)          | Not Null    |
| ProjectStartDate | Date                 |             |
| ProjectEndDate   | Date                 |             |

*Also refer to terms: Conceptual data model, Logical data model, Data model, ER model, Dimensional model.*

**Pivot/Cross Tab:** In data processing, a pivot table is a data summarization tool found in data visualization programs such as spreadsheets (Micorsoft Excel, Google Docs, etc.). Pivot is also called as rotate. In order to provide an alternative representation or perspective of the data, the pivot operation rotates the data axes in view.

*Also refer to terms: OLTP, OLAP, MOLAP, ROLAP, HOLAP.*

**Power Users:** In the context of BI, there are essentially two kinds of users: casual users or consumers of information and power users. They are the producers of information. They rely on ad hoc query and other explorative mechanisms. Examples of power users are developers, analysts, etc.

*Also refer to terms: Casual users.*

**Query:** In general, a query is a form of questioning, a line of enquiry. In the context of RDBMS, a query is essentially a request that a user makes on the database to retrieve data. Database queries are performed using a specific language termed Structured Query Language (SQL).

*Also refer to terms: Database, DDL, DML.*

**RDBMS:** It is a class of system software that manages digital data and is based on relational theory concepts. The relational model uses a collection of tables (relations), each of which is assigned a unique name, to represent both data and the relationships among those data. It is a type of DBMS that stores data in the form of related tables.

*Also refer to terms: Database.*

**Real Time Data Warehouse:** Real time data warehouse is known to house real time business data. If such a data warehouse is queried, it will reflect the state of the business at the time the query was run.

*Also refer to terms: Database, Data warehouse, Data marts.*

**Relation/Table:** A data structure that is based on the relational model and stores the data in the form of rows and columns. Example: Employee Table as shown below.

| EmployeeID | EmployeeName | EmployeeDesignation    |
|------------|--------------|------------------------|
| 101        | Alex         | Senior Project Manager |
| 103        | Felix        | Software Engineer      |

*Also refer to terms: Entity, Attributes, Relationships, ER model.*

**Relationships:** It illustrates how two entities share information in the database structure. Example: Let us consider two entities: Employee and Project. An employee can work on only one project whereas a project can have several employees allocated to it. The relationship between Employee and Project is 1:M.

*Also refer to terms: Entity, Attributes, ER model.*

**ROLAP:** In ROLAP, data is stored in a relational database. ROLAP differs significantly from MOLAP in that it does not require the pre-computation and storage of information. Instead, ROLAP tools access the data in a relational database and generate SQL queries to calculate information at the appropriate level when an end user requests it.

*Also refer to terms: OLTP, OLAP, MOLAP, HOLAP.*

**Schema:** In a relational database schema refers to a collection of database tables, the fields in each table, and the relationships between fields and tables. Schemas are generally stored in a data dictionary.

*Also refer to terms: Database, Relation, Attributes, Entity, Data constraints.*

**Semi-Structured Data:** Semi-structured data does not conform to any data model, i.e. it is difficult to determine the meaning of data. Also, the data cannot be stored in rows and columns as in a database. Semi-structured data, however, has tags and markers which help to group the data and describe how the data is stored, giving some metadata, but it is not sufficient for management and automation of data. Example: XML (eXtensible Markup Language).

*Also refer to terms: Unstructured data, Structured data.*

**Snowflake Schema:** In the context of multidimensional data model, Snowflake schema is a complex data warehouse schema. It has a single, central fact table surrounded by normalized dimension hierarchies. In the Snowflake schema, dimensions are present in a normalized form in multiple related tables. A Snowflake structure materializes when the dimensions of a star schema are detailed and highly structured, having several levels of relationship, and the child tables have multiple parent tables. Refer to Chapter 7 for more details.

*Also refer to terms: Star schema, Fact constellation.*

**Staging Area:** A data staging area can be defined as an intermediate storage area that falls in between the operational/transactional sources of data and the data warehouse (DW) or data mart (DM).

*Also refer to terms: Data warehouse, Data marts, Extraction, Transformation, Loading, ODS.*

**Star Schema:** It is the simplest data warehouse schema. It resembles a star. The center of the star consists of one or more fact tables and the points radiating from the center are the dimensions table. Refer to Chapter 7 for more details.

*Also refer to terms: Snowflake schema, Fact constellation.*

**Structured Data:** Data coming from databases such as Access, OLTP systems, SQL as well spreadsheets such as Excel, etc. are all in the structured format. Structured data conforms to a data model. Here, the data is organized in fixed/variable length fields of a record/file. Example: A project table with its schema definition.

| Column Name      | Data type and length | Constraints |
|------------------|----------------------|-------------|
| ProjectCode      | Varchar(5)           | Primary Key |
| ProjectName      | Varchar(30)          | Not Null    |
| ProjectLocation  | Varchar(30)          | Not Null    |
| ProjectStartDate | Date                 |             |
| ProjectEndDate   | Date                 |             |

Few records in the project table:

| ProjectCode | ProjectName    | ProjectLocation | ProjectstartDate | ProjectEndDate |
|-------------|----------------|-----------------|------------------|----------------|
| P1001       | FinanceProject | Mexico          | 22 July 2011     | 22 Oct. 2011   |
| P1002       | HealthProject  | New Jersey      | 22 Aug 2011      | 22 Nov. 2011   |

*Also refer to terms: Unstructured data, Semi-structured data.*

**Surrogate Key:** Surrogate key is a substitution for the natural primary key. It is simply a unique identifier or number for each row that can be used for the primary key to the table. The only requirement for a surrogate primary key is that it is unique for each row in the table. Surrogate keys are always integer or numeric.

*Also refer to terms: Relation, RDBMS, Data integrity constraint.*

**Transformation:** In building a data warehouse, transformation refers to the process of converting the data from host or legacy format to data warehouse format. These transformations may include operations like making fields into uniform length, converting text to upper case, changing data format to one standard, and so on.

*Also refer to terms: Data marts, Database, Data warehouse, Data mining, Extraction, Loading.*

**Unstructured data:** It refers to information that either does not have a pre-defined data model and/or does not fit well into relational tables. Examples: video files, audio files, chat conversations, wiki, blogs, PowerPoint presentations. Almost 80% of the data generated in enterprise is unstructured data. The challenge remains extracting useful information from unstructured data.

*Also refer to terms: Semi-structured data, Structured data.*



# Index

## A

Alignment, 260  
Analysis  
channel analysis, 302, 304  
distribution channel analysis, 300–301  
funnel analysis, 298, 302–303  
conversion funnel, 299  
conversion rate optimization, 300  
performance analysis, 301–305  
tools, 146  
Analytical applications, 88  
Analytics, 77–78  
in healthcare, 337  
in retail, 335  
in telecom, 334  
Anatomy of recommendation systems, 342  
Anatomy of social media analytics, 339  
ANOVA (Analysis of Variance), 325  
Apple iPad, 372  
Apple iPhone, 372  
Association rule mining, 347  
Attribute, 207

## B

Balanced scorecard, 282  
as a management system, 285–287  
as strategy map  
initiative, 284  
measurements, 284  
objectives, 284  
target, 284  
four perspectives of

customer perspective, 284  
financial perspective, 283  
internal business process perspective, 284  
learning and growth perspective, 284  
Baldrige Criteria for Performance Excellence, 6  
Bespoke IT applications, 14–17  
bespoke application development, 15  
Best practices in BI/DW, 141–142  
business initiative, 142  
data stewardship, 142  
BI component framework, 117–118  
administration and operation layer, 120  
BI and DW operations, 120  
BI architecture, 120–121  
business applications, 123  
data resource administration, 121  
data governance, 121  
metadata management, 121–122  
application metadata, 122–123  
business metadata, 122–123  
process metadata, 122–123  
technical metadata, 122–123  
business layer, 118  
business drivers, 119  
business goals, 119  
business requirements, 118  
business strategies, 119  
business value, 119  
return on asset (ROA), 119  
return on investment (ROI), 119  
total cost of ownership (TCO), 119  
total value of ownership (TVO), 119

- development, 120
  - program management, 120
  - BI professional
    - business intelligence, business analytics, and performance management, 145
    - content management and enterprise information management, 145
    - data mining and predictive analytics, 145
    - data modeling and metadata management, 144
    - data quality and data governance, 144
    - master data management, data integration, and data warehousing, 144
  - BI project team roles
    - business manager, 137
    - business requirements analyst, 138
    - business specialist, 137
    - database administrator (DBA), 140–141
    - decision support analyst, 138–139
    - designer, 139–140
    - ETL specialist, 140
    - project manager, 137–138
  - BI roles and responsibilities
    - administrator, 137
    - metadata manager
      - metadata, 136–137
    - program team roles, 135
      - data architect, 136
      - ETL architect, 136
      - program manager, 135–136
      - technical architect, 136
  - Big data, 91
  - Boxplot, 315
  - Business analytics, 112
  - Business decisions, 268
  - Business intelligence (BI), 104, 126
    - and business analytics, 112
    - business solutions
      - behavior analysis, 133
      - customer analytics, 133
      - marketplace analysis, 133
      - performance analysis, 133
      - productivity analysis, 133
      - sales channel analysis, 133
      - supply chain analytics, 133
    - casual users, 129
      - information consumers, 130
    - cloud computing, 373–375
    - 360 degree perspective, 105
    - in ERP, 378–379
    - fact-based decision making, 104
    - for management, 127
  - and mobility, 369–370
  - mobility timeline, 370–373
  - on the move, 370
  - operational, 127–128
  - operational decisions, 106
  - for process improvement, 128
  - power users, 129
    - information producers, 130
  - single version of truth, 104
  - solutions
    - ad hoc reporting, 108
    - decision support system (DSS), 109
    - executive information system (EIS), 109
  - strategic level, 106
  - tactical decisions, 106
  - technology solutions
    - data mining, 133
    - DSS, 131
    - EIS, 131
    - managed query and reporting, 132
    - OLAP, 131
  - to improve customer experience, 128
  - unstructured data, 110
  - value chain, 111
- Business metrics and KPIs, 267–268
- Business process/model innovation, 12
- Business units, 2
- C**
- Cardinality of relationship, 207
  - Cloud computing, 373–375
  - Compliance analytics, 332
  - Conformed dimensions, 254
  - Contingency table, 316
  - Conventional (slow) method, 253
  - Core business functions, 3
    - accounting, 4
    - product development, 4
    - product/service delivery, 4
    - quality, 4
    - sales and marketing, 4
  - Core business processes, 5
  - Correlation, 318, 324
  - Correlation analysis, 323
  - Covariance, 318
  - Customer behavior analytics, 333
  - Customer relationship management (CRM), 381
    - social, 381
    - sugar, 381
    - traditional, 381
  - Customer segmentation, 333

**D**

- Dashboards, 290, 295  
comparative measures, 293  
evaluation mechanisms, 294  
non-quantitative data, 293  
quantitative data, 293  
scorecards, 296–297  
types of  
customer support dashboards, 292  
divisional dashboards, 292–293  
enterprise performance dashboards, 291–292
- Data, 311
- Data architecture and design, 181
- Data lake, 94
- Database management, 181
- Data mining, 91
- Data governance, 181
- Data integration, 167  
data interchange, 170  
data warehousing, 168–170  
federated databases, 167–168  
instance integration, 165–167  
memory-mapped data structure, 170  
modeling techniques  
ER modeling, 170–171  
dimensional modeling, 171–174  
object brokering, 170  
schema integration, 164–165
- Data integrity  
check constraint, 176  
foreign key, 176  
not null, 176  
primary key, 176
- Data mart(s), 254  
independent data marts, 155
- Data mining, 254
- Data model  
conceptual data model, 207–208  
logical data model, 207–215  
physical data model, 207, 215–219
- Data profiling, 182  
software, 183, 186  
Datiris Profiler, 186  
IBM Infosphere Information Analyzer, 186  
Oracle Warehouse Builder, 186–187  
SSIS Data Profiling Task, 186  
Talend Data Profiler, 186  
Trillium Enterprise Data Quality, 186
- Data quality, 175–176, 180, 182  
completeness, 178  
consistency, 178  
correctness/accuracy, 178  
metadata, 179  
timeliness, 178–179
- Data quality profiling, 184
- Data science, 93
- Data security, 181
- Data warehouse, 74, 109, 152–154  
ad hoc querying, 154  
data access tools, 159  
data presentation area  
dependent data marts, 158  
data security, 157  
data sources, 159  
data staging area, 158  
fact-based decision making, 157  
flexible to change, 157  
information  
accessibility, 157  
consistency, 157  
credibility, 157
- integrated, 154
- lack of  
information credibility, 153  
information sharing, 153
- non-volatile, 155
- on clouds, 377
- operational source systems, 158
- reports take a longer time to be prepared, 154
- subject-oriented, 154
- time-variant, 154
- Data warehousing and business intelligence, 182
- Decision support, 12
- Decision tree, 357, 360, 361
- Decision support system (DSS), 370
- De-normalization, 253
- Departmental IT applications, 11
- Descriptive statistics, 311
- Dimensional model(s), 251  
Fact Constellation schema, 243–246  
Snowflake schema, 239–243  
Star schema, 237–239
- Dimensional modeling, 222, 225  
dimensions, 224  
facts, 224  
multidimensional view, 223
- Dimensional modeling life cycle, 246–247  
dimensions, 250  
facts, 250  
grain, 249–250

granularity, 249  
 requirements gathering  
   source-driven requirements gathering, 247–248  
   user-driven requirements gathering, 248–249  
**Dimension table**  
   dimension hierarchies, 229–230  
   types of  
     degenerate dimension, 230  
     junk dimension, 230, 235–237  
     rapidly changing dimension, 230, 233–234  
     role-playing dimension, 230, 234–235  
     slowly changing dimension, 230–231  
       Type-I SCDS, 232  
       Type-II SCDS, 232–233  
       Type-III SCDS, 233  
**Direct (fast) method**, 253  
**Document type descriptors (DTDs)**, 50  
**DSS (Decision Support System)**, 109  
**Drill-down reports**, 74  
**Drill-through reports**, 74

**E**

**EIS (executive information system)**, 109

**Email**, 44

**Enterprise IT applications**

  IT applications, 16

**Enterprise reporting**

  characteristics

    alerts, 281

    anywhere/anytime/any-device access, 281

    personalization, 281

    reports repository, 282

    role-based delivery, 281

    security, 281

    single version of truth, 281

  better predictability, 282

  enhanced collaboration, 282

  objective communication, 282

**Enterprise resource planning (ERP) systems**, 78, 109,  
 377–378

**Entity**, 207

**ER modeling**

  and dimensional modeling, 174

  entity relationship model, 191

    dimensional model, 192

**ERP plus BI**, 379–380

**ERP software**, 77

**ETL tools**, 146

**eXtensible Markup Language (XML)**, 40, 50

**eXtensible Stylesheet Language**, 50

**Extract, transform, load (ETL)**, 159  
   data extraction, 160  
   data loading, 161–164  
   data mapping, 159  
   data staging, 159–160  
   data transformation, 161

**F**

**Fact-based decision making**  
   and KPIs, 264–266

**Fact-based systems**, 264

**Factless fact table**, 253

**Fact table, types of**

  additive facts, 225–226

  non-additive facts, 226–227

  factless facts (event-based fact tables), 227

  semi-additive facts, 226

**Financial analytics**, 335

**F-test**, 325

**G**

**Google Docs**, 375

**Granularity**, 253

**Groupware**, 11

**H**

**Heterogeneous sources**, 46

**Histogram**, 314

**HTML**, 37

**Human capital analytics**, 332

**Hybrid on-line analytical processing (HOLAP)**, 66, 68

**I**

**Implementation layer**, 123–124

  data warehouse, 124–125

  data warehousing, 123

  information services, 124

**Indexing and searching**, 35

**Indicators**, 297–298

**Inferential statistics**, 311

**Information users**

  IT application users, 17

**Informed decision**, 88

**Inmon's top-down approach**, 156

**Internet-ready applications**, 14

**Inter quartile range (IQR)**, 314

**Interval data**, 320

**IT analytics**, 333

**K**

Key Performance Indicators (KPIs), 264–267  
balanced scorecard (BSC), 266  
cause and effect modeling, 266  
dashboards, 297  
economic value add, 266  
GQMM, 266  
MBNQA, 266  
scorecards, 297  
Six Sigma, 266  
TQM, 266  
*k*-means clustering, 355

**M**

Machine learning, 94  
Malcolm Baldrige Performance Excellence Program, 6  
Management information system (MIS), 107  
Management reporting, 88  
Market basket analysis, 348  
Marketing research, 88  
Market survey, 88  
Measure, 311  
Measurement, Analysis, and Knowledge Management, 6–8  
Measurement system, 284–285  
Measurement system terminology  
  data, 258  
  index, 259  
  indicator, 259  
  measure  
    unit of measure (UOM), 258  
  metric, 259  
Metadata management, 181  
Metric, 261–262, 312  
  supply chain associated with the  
    business activity areas, 263  
    business application, 263  
    business value, 263  
    metrics delivery, 263  
Metric data  
  application, 259  
  quantum, 259  
  stratum, 259  
  subject, 259  
Mining data, 35  
Mobile BI, 373  
  exceptions and alerts, 372  
  pull reporting, 372  
  push reporting, 372  
  device security, 373

transmission security, 373  
authorization, authentication, and network security, 373  
Mobile operating systems (MOS), 372  
Multidimensional modeling, 227–229  
Multidimensional on-line analytical processing (MOLAP), 66, 73  
  dice, 74–75  
  drill-across, 75, 77  
  drill-down, 74, 76  
  drill-through, 75, 77  
  pivot, 75–76  
  roll-up or drill-up, 74–76  
  slice, 74–75

**N**

Nominal data, 320  
Non-parametric statistical tests, 319  
Normalization (entity relationship), 219  
  ER diagram, 221  
Normalization levels  
  normal form  
    1NF, 209  
    2NF, 209  
    3NF, 209  
Null hypothesis, 321

**O**

Object exchange model (OEM), 48  
OLTP database, 253  
On-line analytical processing (OLAP), 62–63, 66, 68–70, 89  
  BI architecture, 73–74  
  Fact Constellation, 70  
  Galaxy schema, 70  
  one-dimensional data, 63–64  
  snowflake model, 70  
  Snowflake schema, 70  
  Star schema, 70  
  three-dimensional data, 65  
  two-dimensional data, 64–65  
On-line transaction processing (OLTP) system, 12, 60, 89  
  aggregation, 74  
  concurrency control (locking), 74  
  delete, 60  
  entity relationship, 70  
  fast query processing, 61  
  insert, 60  
  and OLAP, 68–70

recovery mechanisms (logging), 74  
security, 61  
summarization, 74  
update, 60  
Operational analytics, 335  
Operational data store (ODS), 155, 253  
Open source markup language, 50  
Open source tools, 147  
Ordinal data, 320

## P

Parametric Statistical tests, 319  
Pattern, 312  
Pearson correlation coefficient, 318  
Percentile rank, 314  
Performance management, 265  
Performance measurement, 264  
Popular BI tools, 146  
Population, 316  
p-value, 322

## R

Ratio data, 320  
Ralph Kimball's approach to building  
a data warehouse, 156  
RDBMS, 146  
Referential integrity constraint, 177  
Regression, 324  
Regression to mean (RTM), 324  
Relational on-line analytical processing (ROLAP), 66–67,  
73  
Report delivery formats  
eBOOK, 281  
FTP, 281  
link to reports, 281  
printed reports, 280  
secure soft copy, 280  
Reporting, 276  
Reporting/Ad Hoc querying tools/Visualization, 147  
Reporting perspectives  
function level, 274  
role-based, 274  
standard/ad hoc, 274  
strategic/operational, 274  
summary/detail, 274  
Report layouts, 277  
chart reports, 279  
gauge reports, 280  
list reports, 278–279

matrix reports, 277–278  
tabular reports, 277  
Report standardization and presentation practices  
content standardization, 275  
data standardization, 275  
metrics standardization, 275  
presentation standardization, 275  
reporting tools' standardization, 275  
Roll-up reports, 74  
Rotate, 76

## S

Sales and marketing analytics, 333  
Salesforce, 381  
Sample, 316  
Scorecards *vs.* dashboards, 296–297  
Semi-structured data, 32, 43, 45  
graph-based data models, 47  
indexing, 50  
mining tools, 50  
schemas, 47  
structured data, 51–52  
web pages, 44  
XML, 47, 50  
Single sign-on, 13  
SMART test, 262–263  
Snowflake schema, 254  
Snowflaking, 241  
Social CRM and BI, 380–381  
Standard deviation, 314  
Star model, 71–72  
Star schema, 254  
Statistical data, 88  
Structured data, 32  
characteristics of, 33–34  
OLTP systems, 34  
scalability, 34  
security, 34  
storage, 34  
update and delete, 34  
Subscriber analytics, 335  
Support units, 2  
Surrogate key, 253

## T

TCP/IP packets, 44  
Time series analysis, 327  
t-test, 322

**U**

Unstructured data, 32, 36  
bitmap objects, 37  
BLOBs, 40  
classification/taxonomy, 38–39, 41  
content addressable storage (CAS), 39  
indexing, 38  
retrieve information, 39  
scalability, 39  
searching, 39  
security, 39  
storage space, 39  
tags, 41  
tags/metadata, 38  
text mining, 41  
textual objects, 37  
update and delete, 39  
XML, 40  
Unstructured Information Management Architecture (UIMA), 42–43

**V**

Variance, 314

**W**

Workforce  
compensation analytics, 332  
planning analytics, 332  
sentiment analytics, 332  
talent development analytics, 332  
utilization analytic, 332

**X**

XML, 44

**Z**

Zipped files, 44  
Zoho, 381  
Z-Test, 322



## SALIENT FEATURES OF THE BOOK

- Single source of introductory knowledge on Business Intelligence (BI) which can be taught in one semester.
- Provides a good start for first time learners typically from the engineering and management discipline.
- Covers the complete life cycle of BI/ Analytics project: covering operational/transactional data sources, data transformation, data mart/warehouse design-build, analytical reporting and dashboards.
- Provides a holistic coverage beginning with an enterprise context, developing deeper understanding through the use of tools, touching a few domains where BI is embraced and discussing the problems that BI can help solve.
- Explains concepts with the help of illustrations, application in real-life scenarios and provides opportunities to test understanding.
- In addition, the book also has the following features:
  - ◆ Industrial application case studies.
  - ◆ Crossword puzzles/do it yourself exercises/assignments to help with self-assessment
  - ◆ Glossary
  - ◆ References/web links/bibliography

## NEW TO THIS EDITION

New topics related to *Big Data, Statistics, Lab with R, Lab with Advanced Excel* and Industry examples of applications of analytics.

follow us on



[facebook.com/wileyindia](http://facebook.com/wileyindia)



[twitter.com/wileyindiapl](http://twitter.com/wileyindiapl)



[linkedin.com/in/wileyindia](http://linkedin.com/in/wileyindia)



[google.com/+wileyindia](http://google.com/+wileyindia)

### CD:

To ensure that concepts can be practiced for deeper understanding at low cost or no cost, the book is accompanied with STEP-BY-STEP Hands-On manual on:

- ◆ Advanced MS Excel
- ◆ Introduction to R Programming



**READER LEVEL**  
Undergraduate/Graduate

**SHELVING CATEGORY**  
Engineering

# WILEY

## Wiley India Pvt. Ltd.

4435-36/7, Ansari Road, Daryaganj  
New Delhi-110 002  
Customer Care +91 11 43630000  
Fax +91 11 23275895  
[csupport@wiley.com](mailto:csupport@wiley.com)  
[www.wileyindia.com](http://www.wileyindia.com)

ISBN 978-81-265-6379-1

