

**3.3.3 Total number of books and chapters in edited volumes/books published and papers in national/ international conference proceedings year wise during last five years**

<b>Sr. No.</b>	<b>File Description</b>	<b>Page No.</b>
1	Books by Authors	2

**3.3.3 Total number of books and chapters in edited volumes/books published and papers in national/international conference proceedings year wise during last five years**

Sr. No	Activity Name
1	Books by Author – Dr G.T. Thampi
2	Books by Author – Dr Arti Deshpande
3	Book by Author – Dr Arun Kulkarni
4	Book by Author - Dr Bhushan Jadhav
5	Book by Author- Shilpa Ingoley



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



## Books by Author – Dr G.T. Thampi



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



# PROJECT MANAGEMENT



C. T. Upadhyaya

S. T. Tharopi

Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Borivali (W), Mumbai - 400 103.



**STAREDU  
SOLUTIONS**

[www.staredusolutions.org](http://www.staredusolutions.org)



### Copyright ©StarEdu Solutions

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 written permission of the publisher.

### Limits of Liability:

While the publisher and the author have used their best efforts in preparing this book, StarEdu Solutions and the author make no representation 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.

### Disclaimer:

The contents of this book have been checked for accuracy. Since deviations cannot be precluded entirely, StarEdu Solutions or its author cannot guarantee full agreement. As the book is intended for educational purpose, StarEdu Solutions or its author 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. Due credit for all contributions, images and reference are given in the courseware. In case, we missed on anything, we request you to bring it to our knowledge and we will address the same in next edition of the book. We believe in protecting the intellectual property. If you are copyright owner and if you believe that content constitutes an infringement of copyright, please bring to our knowledge and in case of justified issue, we will address the same in next edition of the book.

This book may contain reference to other resources related to topics. These references are for training purpose only. However, StarEdu Solutions do not recommend or promoting such references, solutions, software and products mentioned in reference or training content.

Images used in the book are for education purpose only and we are not promoting or advertising any product.

All Readers are requested to report errors, inaccuracies and omissions to StarEdu Solutions at [support@staredusolutions.org](mailto:support@staredusolutions.org), so that same can be addressed in the next edition of the the book.

ISBN: 978-93-86765-71-0

Edition: 2020



Dr. G. T. Thampi  
PRINCIPAL

Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



Dr. G. T. Thampi  
PRINCIPAL

Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



# Preface

**Project Management** is a significant concept in the modern times. The business organisations have been established to earn profits by creating deliverables and selling them in the market. These deliverables are developed through projects. Often these projects are carried out without any management and supervision. This leads to increased costs, delayed projects and unachieved aims. Hence, the Project Management is the need of the hour.

The project completion is achieved through the coordinated efforts of the members of project teams led by the project manager. A project leader is responsible for managing all the stages of the project and garnering support from the sponsors and the management, so as to ensure that the goals of the project are achieved. Therefore, the project managers should be familiar with all the project management techniques to control the execution of the projects and deliver the intended results on time. These techniques should be revisited, reorganised and reframed, depending on the ever-changing business scenarios. Keeping this view in mind, the Project Management book has been composed, which serves as a powerful tool for enabling the students to get acquainted with the innovative concepts of project management, its techniques and its applications.

The **Project Management book** is designed to enlighten the students about the importance of project management in business organisations. The book covers all the major aspects of project management, including project life cycle, initiation of projects, project planning and scheduling, project risk management, project execution and monitoring, and project termination. After studying this book, the students will be able to comprehend and apply the various concepts of project management in real-world business scenarios, so as to gain hands-on experience.

The **Project Management book** is developed with an updated content about the different aspects of the project management. It is written in a simple, lucid and easy-to-understand language. In this book, various real-life examples have been provided in all the chapters to enable the students to gain insights into the application of project management concepts. The examples will also guide them on how to deal with the similar real-life situations and take the necessary measures.

The content of the **Project Management book** has been developed after a wide-scale research done on the subject. The book has been designed in such a way so as to enhance the understanding of the students regarding the complex issues, often faced while handling project management related concepts. The book will enable the students to gain a clear understanding and confidence of using project management techniques and lead the multitude of projects in their real lives.



## About the Authors



Dr. K. T. Upadhyaya has over 17 years experience in executing large scale projects initially in Cement, Petrol chemical, Food, and Dairy industry and currently in Software development, Data Analytics and Automation projects. Dr. K. T. Upadhyaya is engaged in training in Project Management, Risk Management, Data warehousing for past 16 years. He has trained over 10,000 participants from companies in different domains and students from top B-schools in India. He also trains on Data Analytics, Supply Chain Analytics, Business Process Management, Process Automation and AI.

Dr. K. T. Upadhyaya has keen interest in relating Indian Mythology, particularly Mahabharata, to management lessons. He is an avid reader of books relating to topics ranging from Data to Fiction and everything in between. Sachin Dev Burman and Sachin Tendulkar are his two favourites in their respective domains. Dr. K. T. Upadhyaya holds a Mechanical Engineering Degree from Sardar Patel College of Engineering, Mumbai and Masters and Ph. D. from BITS Pilani. He is a Certified Project Management Professional from PMI(R), Pennsylvania, USA.



"Dr. G. T. Thampi" is currently The Principal at Thadomal Shahani Engineering College, Bandra, Mumbai. He hold a Ph.D Degree in Technology and has more than 33 years of experience in renowned college. Dr. G.T. Thampi has been a part of some interesting researches and holds interests in Business Process and Re-Engineering in realm of Engineering Education. Apart from his own researches he has been a guide for multiple researches done in technology front. 17 research scholars awarded Ph.D. under his guidance. Dr. Thampi is also a co-author of more than 80 research publications and books

Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



## About the Book

**Project Management** book is designed to familiarise the students with the different aspects of successful project completion. This book provides comprehensive knowledge about the various concepts and strategies for managing projects effectively. It includes various examples to provide an understanding of the application of project management concepts in the real world. The content of the book is designed in a simple and student-friendly manner to help them apply project management techniques in real life.

## Salient Features of the Book

Easy-to-understand language

Comprehensive coverage of all the relevant concepts

Numerous examples to enhance the understanding of students

Variety of diagrams and cases to impart the knowledge about the complexities faced during the management of projects

Several review questions after each chapter to appraise the performance of the students

Index terms at the end of the book to provide the definitions of project-related terms

Dr. G. T. Thampi  
PRINCIPAL

Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



StarEdu Solutions India Pvt. Ltd.

- 450, North Brand Blvd, Suite 600, Glendale, CA 91203
- support@staredusolutions.org
- www.staredusolutions.org

ISBN: 978-93-86765-71-0



# Books by Author – Dr Arti Deshpande




Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.







  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.





# Machine Learning

(Code : 410250)

Semester VIII - Computer Engineering  
(Savitribai Phule Pune University)

Strictly as per the New Credit System Syllabus (2015 Course)  
Savitribai Phule Pune University w.e.f. Academic Year 2018-2019


## Dr. Pallavi N. Halarnkar

Department of Computer Engineering  
Thadomal Shahani Engineering College, Mumbai.  
Maharashtra, India.

## Dr. Arti Deshpande

Department of Computer Engineering  
Thadomal Shahani Engineering College,  
Mumbai.  
Maharashtra, India.



  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



PE66A Price ₹ 150/-



## **Machine Learning**

Dr. Pallavi N. Halarnkar, Dr. Arti Deshpande

(Semester VIII - Computer Engineering) (Savitribai Phule Pune University)

Copyright © by Authors. All rights reserved. No part of this publication may be reproduced, copied, or stored in a retrieval system, distributed or transmitted in any form or by any means, including photocopy, recording, or other electronic or mechanical methods, without the prior written permission of the publisher.

This book is sold subject to the condition that it shall not, by the way of trade or otherwise, be lent, resold, hired out, or otherwise circulated without the publisher's prior written consent in any form of binding or cover other than which it is published and without a similar condition including this condition being imposed on the subsequent purchaser and without limiting the rights under copyright reserved above.

**First Edition : February 2019**

**Second Revised Edition : January 2020 (TechKnowledge Publications)**

This edition is for sale in India, Bangladesh, Bhutan, Maldives, Nepal, Pakistan, Sri Lanka and designated countries in South-East Asia. Sale and purchase of this book outside of these countries is unauthorized by the publisher.


**Printed at :** 37/2, Ashtavinayak Industrial Estate,  
Near Pari Company,  
Narhe, Pune, Maharashtra State, India.  
Pune – 411041

**ISBN : 978-93-89748-20-8**

**Published by**

**TechKnowledge Publications**

**Head Office :** B/5, First floor, Maniratna Complex,  
Taware Colony, Aranyeshwar Corner,  
Pune - 411 009. Maharashtra State, India  
Ph : 91-20-24221234, 91-20-24225678.  
Email : info@techknowledgebooks.com,  
Website : www.techknowledgebooks.com

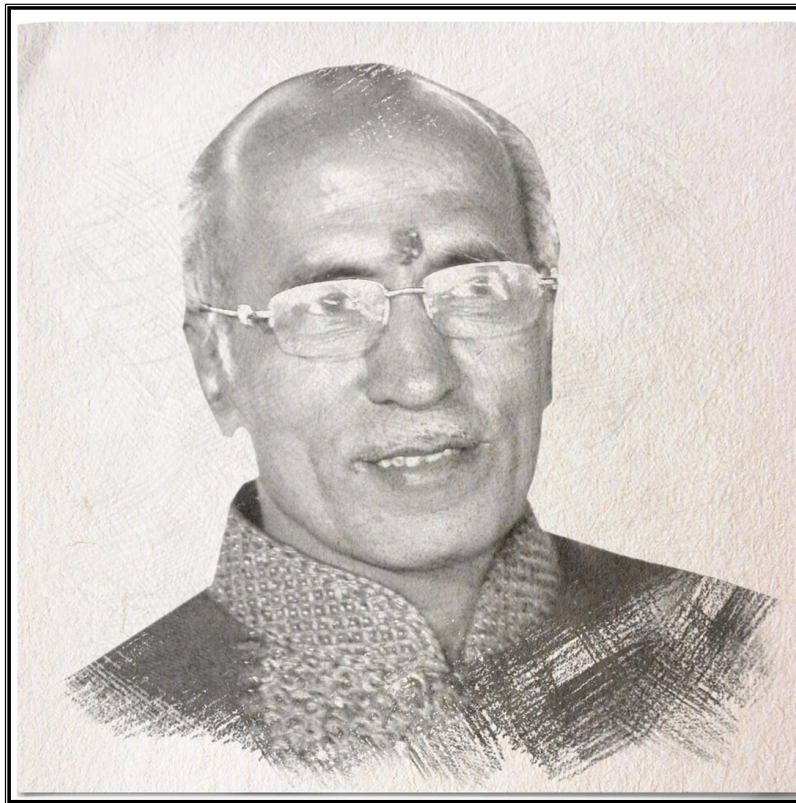
  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



[410250] (FID : PE66) (Book Code : PE66A)

(Book Code : PE66A)

*We dedicate this Publication soulfully and wholeheartedly,  
in loving memory of our beloved founder director,  
**Late Shri. Pradeepji Lalchandji Lunawat,**  
who will always be an inspiration, a positive force and strong support  
behind us.*



***“My work is my prayer to God”***

*- Lt. Shri. Pradeepji L. Lunawat*

Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



***Soulful Tribute and Gratitude for all Your  
Sacrifices, Hardwork and 40 years of Strong Vision...***



## Preface

Dear students,

We are extremely happy to present the book of “**Machine Learning**” for you. We have divided the subject into small chapters so that the topics can be arranged and understood properly. The topics within the chapters have been arranged in a proper sequence to ensure smooth flow of the subject.

We present this book in the loving memory of **Late Shri. Pradeepji Lunawat**, our source of inspiration and a strong foundation of “**TechKnowledge Publications**”. He will always be remembered in our heart and motivate us to achieve our milestone.

We are thankful to Mr. Shital Bhandari, Shri. Arunoday Kumar and Shri. Chandroday Kumar for the encouragement and support that they have extended. We are also thankful to the staff members of TechKnowledge Publications and others for their efforts to make this book as good as it is. We have jointly made every possible efforts to eliminate all the errors in this book. However if you find any, please let us know, because that will help us to improve further.

We are also thankful to my family members and friends for patience and encouragement.

  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.

 **Authors**  


## Syllabus

Savitribai Phule Pune University  
Fourth Year of Computer Engineering (2015 Course)

### 410250 : Machine Learning

<b>Teaching Scheme :</b> <b>TH : 03 Hours/Week</b>	<b>Credit</b> <b>03</b>	<b>Examination Scheme :</b> <b>In-Sem (Paper) : 30 Marks</b> <b>End-Sem (Paper) : 70 Marks</b>
---	----------------------------	--

**Prerequisite Courses :** 207003 - Engineering Mathematics III

**Companion Course:** 410254 - Laboratory Practice III

#### Course Objectives

- To understand human learning aspects and relate it with machine learning concepts.
- To understand nature of the problem and apply machine learning algorithm.
- To find optimized solution for given problem.

#### Course Outcomes

On completion of the course, student will be able to -

- Distinguish different learning based applications.
- Apply different preprocessing methods to prepare training data set for machine learning.
- Design and implement supervised and unsupervised machine learning algorithm.
- Implement different learning models.
- Learn Meta classifiers and deep learning concepts.


### Course Contents

#### Unit I : Introduction to Machine Learning

(08 Hours)

Classic and adaptive machines, Machine learning matters, Beyond machine learning-deep learning and bio inspired adaptive systems, Machine learning and Big data.

Important Elements of Machine Learning - Data formats, Learnability, Statistical learning approaches, Elements of information theory.

  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



(Refer chapter 1)

#### Unit II : Feature Selection

(08 Hours)

Scikit - learn Dataset, Creating training and test sets, managing categorical data, Managing missing features, Data scaling and normalization, Feature selection and Filtering, Principle Component Analysis(PCA) - non negative matrix factorization, Sparse PCA, Kernel PCA. Atom Extraction and Dictionary Learning.

(Refer chapter 2)



### Unit III : Regression

(08 Hours)

**Linear Regression :** Linear models, A bi-dimensional example, Linear Regression and higher dimensionality, Ridge, Lasso and ElasticNet, Robust regression with random sample consensus, Polynomial regression, Isotonic regression.

**Logistic Regression :** Linear classification, Logistic regression, Implementation and Optimizations, Stochastic gradient descent algorithms, Finding the optimal hyper-parameters through grid search, Classification metric, ROC Curve.

(Refer chapter 3)

### Unit IV : Naïve Bayes and Support Vector Machine

(08 Hours)

Bayes' Theorem, Naïve Bayes' Classifiers, Naïve Bayes in Scikit - learn- Bernoulli Naïve Bayes, Multinomial Naïve Bayes, and Gaussian Naïve Bayes.

**Support Vector Machine(SVM) :** Linear Support Vector Machines, Scikit- learn implementation-Linear Classification, Kernel based classification, Non-linear Examples. Controlled Support Vector Machines, Support Vector Regression.

(Refer chapter 4)

### Unit V : Decision Trees and Ensemble Learning

(08 Hours)

**Decision Trees :** Impurity measures, Feature Importance. Decision Tree Classification with Scikit-learn, Ensemble Learning-Random Forest, AdaBoost, Gradient Tree Boosting, Voting Classifier.

**Clustering Fundamentals:** Basics, K-means: Finding optimal number of clusters, DBSCAN, Spectral Clustering. Evaluation methods based on Ground Truth- Homogeneity, Completeness, Adjusted Rand Index.

**Introduction to Meta Classifier :** Concepts of Weak and eager learner, Ensemble methods, Bagging, Boosting, Random Forests.

(Refer chapter 5)

### Unit VI : Clustering Techniques

(08 Hours)


Hierarchical Clustering, Expectation maximization clustering, Agglomerative Clustering-Dendrograms, Agglomerative clustering in Scikit- learn, Connectivity Constraints.

**Introduction to Recommendation Systems :** Naïve User based systems, Content based Systems, Model free collaborative filtering-singular value decomposition, alternating least squares.

**Fundamentals of Deep Networks:** Defining Deep learning, common architectural principles of deep networks, building blocks of deep networks.

(Refer chapter 6)

□□□

  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai-400 050.



eBooks are available on [www.techknowledgebooks.com](http://www.techknowledgebooks.com)

**Your Success is Our Goal**  
**Semester V - Computer Engineering**

Theoretical Computer Science	Dilip Kumar Sultania
Software Engineering	M. A. Ansari
Computer Network	J. S. Katre
Data Warehousing & Mining	Dr. Arti Deshpande, Dr. Pallavi N. Halarankar
<b>Department Level Optional Course - 1</b>	
Internet Programming	Yogesh Prabhakar Pingale, Smita Jawale, Pooja Vartak
Adv. Data Management System	Mahesh Mali

**coming soon.....**

**es**  
**easy-solutions**

Paper Solutions Trusted by lakhs of students from more than 15 years

**Head Office :**  
B/5, First Floor, Maniratra Complex, Taware Colony,  
Aranyeshwar Corner, Pune - 411009, Maharashtra, India.  
Tel. : 91-20-24221234, 91-20-24225678


**Distributors**

Student's Agencies (S) Pvt. Ltd. T. : (020) 2485451, 09132 30777	Vidyaarthi Sales Agencies T. : (020) 23657279, 091697 76110	Shruti Sales Agency T. : (020) 23819359, 096572 82797
---	--	--

**Our Branches : Pune | Mumbai | Kolhapur | Nagpur | Solapur | Nashik**

**For Library Orders Contact - Ved Book Distributors M : 80975 71421 / 92208 77214**

Email : [info@techknowledgebooks.com](mailto:info@techknowledgebooks.com) Website : [www.techknowledgebooks.com](http://www.techknowledgebooks.com)



**DATA WAREHOUSING AND MINING**  
Semester V - Computer Engineering

Strictly as per the New Revised Syllabus (REV- 2019 'C' Scheme)  
of Mumbai University w.e.f. academic year 2021-22

# Data Warehousing and Mining

(Code : CSC504)  
**Semester V - Computer Engineering**

**Dr. Arti Deshpande**  
**Dr. Pallavi N. Halarankar**

**Tech Knowledge Publications**

**Includes :**  
● Solved Latest University Question Papers.


ISBN : 978-93-50654-96-9

Price ₹ 295/-

MO175A

TechKnowledgePublications

Book Code: MO175A  
Price: ₹ 295/-

  
**Dr. G. T. Thampi**  
**PRINCIPAL**  
 Thadomal Shahani Engineering College  
 Bandra (W), Mumbai - 400 050.



# Data Warehousing and Mining

(Code : CSC504)

**Semester V : Computer Engineering (Mumbai University)**

**Strictly as per the New Revised Syllabus (Rev-2019 'C' Scheme)  
of Mumbai University w.e.f. academic year 2021-2022  
(As per Choice Based Credit and Grading System)**

**Dr. Arti Deshpande**

Department of Computer Engineering

Thadomal Shahani Engineering College, Mumbai.

Maharashtra, India

**Dr. Pallavi Halarnkar**



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



 **TechKnowledge**<sup>TM</sup>  
Publications

**M0175A Price ₹ 295/-**



## **Data Warehousing and Mining (CSC504)**

**Dr. Arti Deshpande, Dr. Pallavi Halarnkar**

Semester V : Computer Engineering (Mumbai University)

*Copyright © Authors. All rights reserved. No part of this publication may be reproduced, copied, or stored in a retrieval system, distributed or transmitted in any form or by any means, including photocopy, recording, or other electronic or mechanical methods, without the prior written permission of the publisher.*

*This book is sold subject to the condition that it shall not, by the way of trade or otherwise, be lent, resold, hired out, or otherwise circulated without the publisher's prior written consent in any form of binding or cover other than which it is published and without a similar condition including this condition being imposed on the subsequent purchaser and without limiting the rights under copyright reserved above.*

**First Printed in India** : February 2019

**First Edition** : July 2021

*This edition is for sale in India, Bangladesh, Bhutan, Maldives, Nepal, Pakistan, Sri Lanka and designated countries in South-East Asia. Sale and purchase of this book outside of these countries is unauthorized by the publisher.*

**ISBN** : 978-93-90694-96-9

**Published By**

**TECHKNOWLEDGE PUBLICATIONS**

**Printed @**

37/2, Ashtavinayak Industrial Estate,  
Near Pari Company,  
Narhe, Pune, Maharashtra State, India.  
Pune - 411041

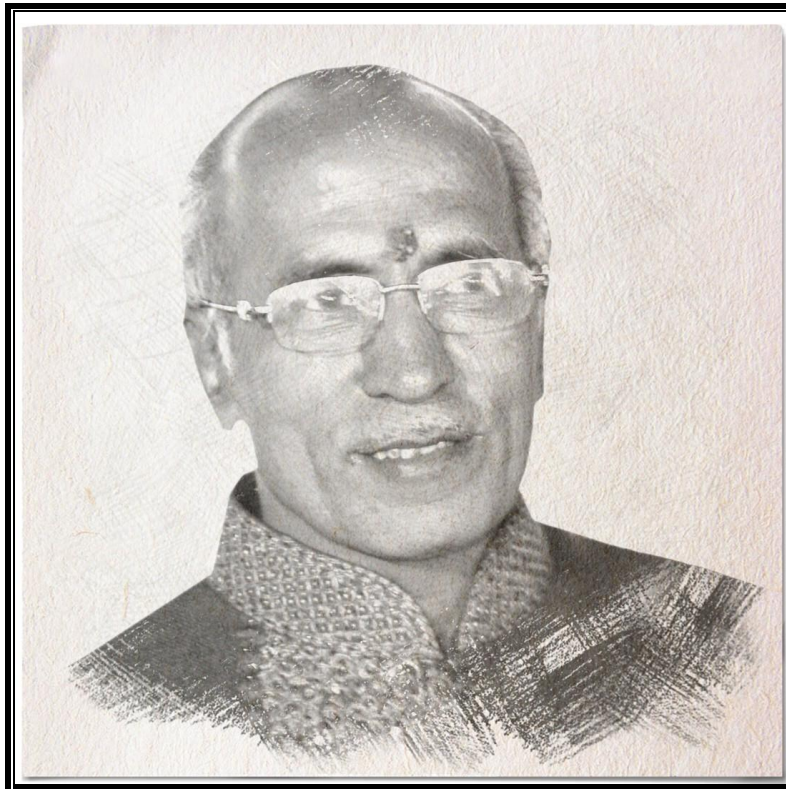
**Head Office**

B/5, First floor, Maniratna Complex, Taware Colony,  
Aranyeshwar Corner, Pune - 411009.  
Maharashtra State, India  
  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.  
**Ph** : 91-20-24221234, 91-20-24225678.  
**Email** : info@techknowledgebooks.com,  
**Website** : www.techknowledgebooks.com

**Subject Code** : CSC504

**Book Code** : MO175A

*We dedicate this Publication soulfully and wholeheartedly,  
in loving memory of our beloved founder director,  
**Late Shri. Pradeepji Lalchandji Lunawat,**  
who will always be an inspiration, a positive force and strong support  
behind us.*



***"My work is my prayer to God"***

*- Lt. **Shri Pradeepji L. Lunawat***

Dr. G. T. Thampi  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



*Soulful Tribute and Gratitude for all Your  
Sacrifices, Hardwork and 40 years of Strong Vision...*

## Preface

My Dear Students,

We are extremely happy to come out with this book on “ **Data Warehousing and Mining**” for you. The topics within the chapters have been arranged in a proper sequence to ensure smooth flow of the subject.

We present this book in the loving memory of Late Shri. Pradeepji Lunawat, our source of inspiration and a strong foundation of “TechKnowledge Publications”. He will always be remembered in our heart and motivate us to achieve our milestone.

We are thankful to Prof. J. S. Katre, Mr. Shital Bhandari, Prof. Arunoday Kumar and Shri. Chandroday Kumar for the encouragement and support that they have extended. We also thankful to Seema Lunawat for technology enhanced reading, E-books support and the staff members of TechKnowledge Publications for their efforts to make this book as good as it is. We have jointly made every possible efforts to eliminate all the errors in this book. However if you find any, please let me know, because that will help me to improve further.

We are thankful to my family members and friends for patience and encouragement.

  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.

 **Authors** 



## SYLLABUS

Mumbai University Third Year of Computer Engineering (2019 Course)		
Subject Code	Subject Name	Credits
CSC504	Data Warehousing and Mining	03

**Prerequisite :** Database Concepts

**Course Objectives :**

1	To identify the significance of Data Warehousing and Mining.
2	To analyze data, choose relevant models and algorithms for respective applications.
3	To study web data mining.
4	To develop research interest towards advances in data mining.

**Course Outcomes :** At the end of the course, the student will be able to

1	Understand data warehouse fundamentals and design data warehouse with dimensional modelling and apply OLAP operations.
2	Understand data mining principles and perform Data preprocessing and Visualization.
3	Identify appropriate data mining algorithms to solve real world problems.
4	Compare and evaluate different data mining techniques like classification, prediction, clustering and association rule mining.
5.	Describe complex information and social networks with respect to web mining

  
**Dr. G. T. Thampi**  
 PRINCIPAL  
 Thadomal Shahani Engineering College  
 100, Kurla Road, Mumbai - 400 050



**Detailed Syllabus :**

Module	Course Module / Contents	Periods
1	<b>Data Warehousing Fundamentals</b> Introduction to Data Warehouse, Data warehouse architecture, Data warehouse versus Data Marts, E-R Modeling versus Dimensional Modeling, Information Package Diagram, Data Warehouse Schemas; Star Schema, Snowflake Schema, Factless Fact Table, Fact Constellation Schema. Update to the dimension tables. Major steps in ETL process, OLTP versus OLAP, OLAP operations: Slice, Dice, Rollup, Drilldown and Pivot. <b>(Refer Chapter 1)</b>	08
2	<b>Introduction to Data Mining, Data Exploration and Data Pre-processing</b> Data Mining Task Primitives, Architecture, KDD process, Issues in Data Mining, Applications of Data Mining, Data Exploration: Types of Attributes, Statistical Description of Data, Data Visualization, Data Preprocessing: Descriptive data summarization, Cleaning, Integration & transformation, Data reduction, Data Discretization and Concept hierarchy generation. <b>(Refer Chapter 2)</b>	08
3	<b>Classification</b> Basic Concepts, Decision Tree Induction, Naïve Bayesian Classification, Accuracy and Error measures, Evaluating the Accuracy of a Classifier: Holdout & Random Subsampling, Cross Validation, Bootstrap. <b>(Refer Chapter 3)</b>	06
4	<b>Clustering</b> Types of data in Cluster analysis, Partitioning Methods (k-Means, k-Medoids), Hierarchical Methods (Agglomerative, Divisive). <b>(Refer Chapter 4)</b>	06
5	<b>Mining Frequent Patterns and Associations</b> Market Basket Analysis, Frequent Item sets, Closed Item sets, and Association Rule, Frequent Pattern Mining, Apriori Algorithm, Association Rule Generation, Improving the Efficiency of Apriori, Mining Frequent Itemsets without candidate generation, Introduction to Mining Multilevel Association Rules and Mining Multidimensional Association Rules. <b>(Refer Chapter 5)</b>	06
6	<b>Web Mining</b> Introduction, Web Content Mining: Crawlers, Harvest System, Virtual Web View, Personalization, Web Structure Mining: Page Rank, Clever, Web Usage Mining <b>(Refer Chapter 6)</b>	05

□□□

Third Year Diploma - Semester VI  
**COMPUTER ENGINEERING GROUP**

**MANAGEMENT**  
Virat V. Giri, Dr. Yogeshwari L. Giri

**MOBILE APPLICATION DEVELOPMENT**  
Virat V. Giri, Sagar Chavan, Ashwini Mane

**PROGRAMMING WITH 'PYTHON'**  
Ravi Majithia

**WEB BASED APPLICATION DEVELOPMENT WITH PHP (Elective)**  
Vijay T. Patil, Yogita N. Jore, Prasad J. Koyande

**NETWORK AND INFORMATION SECURITY (Elective)**  
Shital M. Mate

**DATA WAREHOUSING WITH MINING TECHNIQUES (Elective)**  
Dr. Arti Deshpande, Dr. Pallavi N. Halarnkar



2020 MSBTE T.Y. Diploma Scheme

Strictly as per new revised 'T' Scheme  
w.e.f. academic year 2019-2020

## DATA WAREHOUSING WITH MINING TECHNIQUES

(Code : 22621) (Elective)

Semester VI – Computer Engineering Program Group (CO/CM/CW)

Same Subject, Same Authors with New Publication

Dr. Arti Deshpande      Dr. Pallavi N. Halarnkar

Includes :

- Model Question Papers as per Bloom's Revised Taxonomy.

coming soon.....

**easy-solutions** now with **TechKnowledge Publications**

Paper Solutions Trusted by lakhs of students from more than 15 years

ISBN : 978-93-89684-45-2

Price : ₹ 145/-

MDE57A


Head Office :  
B/5, First Floor, Maniratna Complex, Taware Colony, Aranyeshwar Corner,  
Pune - 411009, Maharashtra, India. Tel. : 91-20-24221234, 91-20-24225678

Distribution Address :  
Shop No. 8-9, Ground Floor, Shaan Baramba Complex, Appa Balwant Chowk, Opp. Farakhana Police Station,  
Pune - 411002, Mobile : 74883 91819 / 74888 91812

Our Branches : Pune | Mumbai | Kolhapur | Nagpur | Solapur | Nashik

Email : info@techknowledgebooks.com | Website : www.techknowledgebooks.com

TechKnowledge Publications

  
**Dr. G. T. Thampi**  
**PRINCIPAL**  
 Thadomal Shahani Engineering College  
 Bandra (W), Mumbai - 400 050.



# Data Warehousing and Mining

(Code - CSC603)

Semester VI - Computer Engineering

(Mumbai University)

Strictly as per the Choice Based Credit and Grading System  
(Revise 2016) of Mumbai University w.e.f. academic year 2018-2019

## Dr. Arti Deshpande

Department of Computer Engineering  
Thadomal Shahani Engineering College,  
Mumbai.  
Maharashtra, India

## Dr. Pallavi Halarnkar

Department of Computer Engineering  
Thadomal Shahani Engineering College,  
Mumbai.  
Maharashtra, India.



(Book Code : ME40A)

Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai-400 050.

## **Data Warehousing and Mining**

Dr. Arti Deshpande, Dr. Pallavi Halarnkar

(Semester VI - Computer Engineering, Mumbai University)

Copyright © by Authors. All rights reserved. No part of this publication may be reproduced, copied, or stored in a retrieval system, distributed or transmitted in any form or by any means, including photocopy, recording, or other electronic or mechanical methods, without the prior written permission of the publisher.

This book is sold subject to the condition that it shall not, by the way of trade or otherwise, be lent, resold, hired out, or otherwise circulated without the publisher's prior written consent in any form of binding or cover other than which it is published and without a similar condition including this condition being imposed on the subsequent purchaser and without limiting the rights under copyright reserved above.

**First Edition** : February 2019

**Second Revised Edition** : January 2020 (**TechKnowledge Publications**)

This edition is for sale in India, Bangladesh, Bhutan, Maldives, Nepal, Pakistan, Sri Lanka and designated countries in South-East Asia. Sale and purchase of this book outside of these countries is unauthorized by the publisher.

**Printed at :** 37/2, Ashtavinayak Industrial Estate,  
Near Pari Company, Narhe, Pune, Maharashtra State, India.  
Pune – 411041.


**ISBN :** 978-93-89748-90-1

**Published by**

**TechKnowledge Publications**

**Head Office :** B/5, First floor, Maniratna Complex, Taware Colony, Aranyeshwar Corner,  
Pune - 411 009. Maharashtra State, India  
Ph : 91-20-24221234, 91-20-24225678.  
Email : info@techknowledgebooks.com,  
Website : www.techknowledgebooks.com

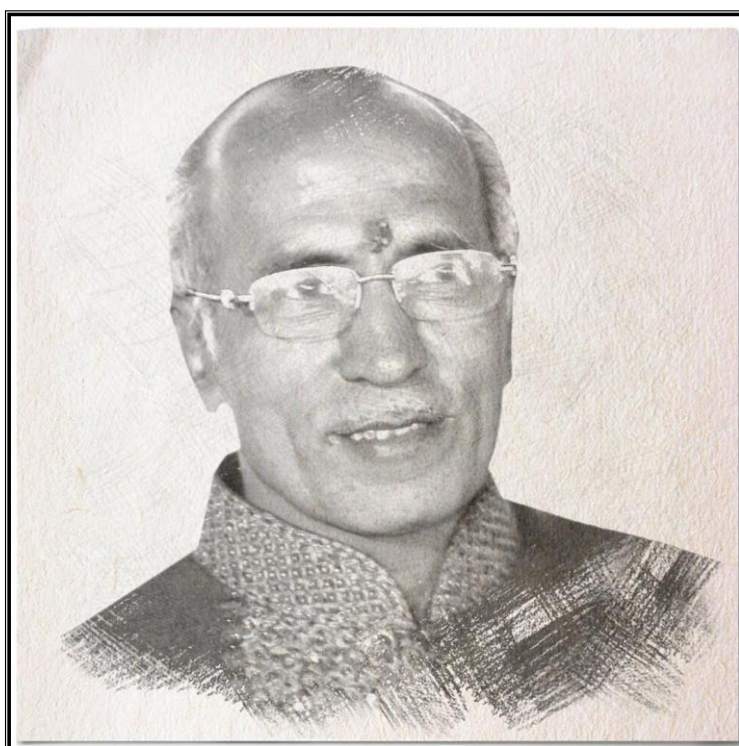
[CSC603] (FID : ME40) ( Book Code : ME40A)

  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



( Book Code : ME40A)

*We dedicate this Publication soulfully and wholeheartedly, in loving memory of our beloved founder director, **Late Shri. Pradeepji Lalchandji Lunawat**, who will always be an inspiration, a positive force and strong support behind us.*



***“My work is my prayer to God”***

*- Lt. Shri. Pradeepji L. Lunawat*

***Soulful Tribute and Gratitude for all Your Sacrifices, Hardwork, and 40 years of Strong Vision...***

Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



( Book Code : ME40A)



## Preface

Dear Students,

We are extremely happy to present the book of **“Data Warehousing and Mining”** for you. We have divided the subject into small chapters so that the topics can be arranged and understood properly. The topics within the chapters have been arranged in a proper sequence to ensure smooth flow of the subject.

We present this book in the loving memory of **Late. Shri. Pradeepji Lunawat**, our source of inspiration and a strong foundation of **“TechKnowledge Publications”**. He will always be remembered in our hearts and motivate us to achieve our new milestone.

We are thankful to Prof. Arunoday Kumar, Mr. Shital Bhandari and Shri. Chandroday Kumar for the encouragement and support that they have extended. We also thankful to the staff members of TechKnowledge Publications for their efforts to make this book as good as it is. We have made every possible efforts to eliminate all the errors in this book. However if you find any, please let us know, because that will help us to improve the book quality further.

We are thankful to my family members and friends for their patience and encouragement.

  
Dr. G. I. T. **Authors**  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



## Syllabus

Course Code	Course Name	Credits
CSC603	Data Warehousing and Mining	4

### Course Objectives :

1. To identify the scope and essentiality of Data Warehousing and Mining.
2. To analyze data, choose relevant models and algorithms for respective applications.
3. To study spatial and web data mining.
4. To develop research interest towards advances in data mining.

### Course Outcomes : On successful completion of course learner will be able to :


1. Understand Data Warehouse fundamentals, Data Mining Principles.
2. Design data warehouse with dimensional modelling and apply OLAP operations.
3. Identify appropriate data mining algorithms to solve real world problems.
4. Compare and evaluate different data mining techniques like classification, prediction, clustering and association rule mining.
5. Describe complex data types with respect to spatial and web mining.
6. Benefit the user experiences towards research and innovation.



**Prerequisite :** Basic database concepts, Concepts of algorithm design and analysis.

Module No.	Topics	Hrs.
1.0	<b>Introduction to Data Warehouse and Dimensional modeling :</b> Introduction to Strategic Information, Need for Strategic Information, Features of Data Warehouse, Data warehouses versus Data Marts, Top-down versus Bottom-up approach. Data warehouse architecture, metadata, E-R modelling versus Dimensional Modelling, Information Package Diagram, STAR schema, STAR schema keys, Snowflake Schema, Fact Constellation Schema, Factless Fact tables, Update to the dimension tables, Aggregate fact tables. <b>(Refer chapter 1)</b>	8
2.0	<b>ETL Process and OLAP :</b> Major steps in ETL process, Data extraction : Techniques, Data transformation : Basic tasks, Major transformation types, Data Loading : Applying Data, OLTP Vs OLAP, OLAP definition, Dimensional Analysis, Hypercubes, OLAP operations : Drill down, Roll up, Slice, Dice and Rotation, OLAP models : MOLAP, ROLAP. <b>(Refer chapter 2)</b>	8

( Book Code : ME40A)

Module No.	Topics	Hrs.
3.0	<b>Introduction to Data Mining, Data Exploration and Preprocessing :</b> Data Mining Task Primitives, Architecture, Techniques, KDD process, Issues in Data Mining, Applications of Data Mining, Data Exploration : Types of Attributes, Statistical Description of Data, Data Visualization, Data Preprocessing : Cleaning, Integration, Reduction : Attribute subset selection, Histograms, Clustering and Sampling, Data Transformation and Data Discretization : Normalization, Binning, Concept hierarchy generation, Concept Description : Attribute oriented Induction for Data Characterization. <b>(Refer chapter 3)</b>	10
4.0	<b>Classification, Prediction and Clustering :</b> Basic Concepts, Decision Tree using Information Gain, Induction : Attribute Selection Measures, Tree pruning, Bayesian Classification : Naive Bayes, Classifier Rule - Based Classification : Using IF-THEN Rules for classification, Prediction : Simple linear regression, Multiple linear regression Model Evaluation and Selection : Accuracy and Error measures, Holdout, Random Sampling, Cross Validation, Bootstrap, Clustering : Distance Measures, Partitioning Methods (k-Means, k-Medoids), Hierarchical Methods (Agglomerative, Divisive). <b>(Refer chapter 4)</b>	12
5.0	<b>Mining Frequent Patterns and Association Rules :</b> Market Basket Analysis, Frequent Item sets, Closed Item sets and Association Rule, Frequent Pattern Mining, Efficient and Scalable Frequent Item set Mining Methods : Apriori Algorithm, Association Rule Generation, Improving the Efficiency of Apriori, FP growth, Mining frequent Itemsets using Vertical Data Format, Introduction to Mining Multilevel Association Rules and Multidimensional Association Rules. <b>(Refer chapter 5)</b>	8
6.0	<b>Spatial and Web Mining :</b> Spatial Data, Spatial Vs. Classical Data Mining, Spatial Data Structures, Mining Spatial Association and Co-location Patterns, Spatial Clustering Techniques : CLARANS Extension, Web Mining : Web Content Mining, Web Structure Mining, Web Usage mining, Applications of Web Mining. <b>(Refer chapter 6)</b>	6
	<b>Total</b>	<b>52</b>

  
**Dr. G. T. Thampi**  
**PRINCIPAL**  
 Thadomal Shahani Engineering College  
 Bandra (W), Mumbai - 400 050.

( Book Code : ME40A)

*Your Success is Our Goal*

**Semester VI - Computer Engg. / CSE**

- **Theory of Computation**  
Dilip Kumar Sultania
- **Microprocessor and Interfacing**  
Harish G. Narula, Khushboo Shah
- **Cryptography and Network Security**  
Pravin Goyal
- **Data Mining (Elective-II)**  
Arti Deshpande, Pallavi Halarankar
- **Advance Java Programming (Elective-III)**  
Ravi Majethia
- **Image Processing (Elective-II)**  
Dhananjay K. Thekkedath
- **IOT & Applications (Open Elective-II)**  
Pravin Goyal

ebooks are available on [www.techknowledgebooks.com](http://www.techknowledgebooks.com)

**coming soon.....**

**easy-solutions**

Paper Solutions Trusted by lakhs of students from more than 15 years

Head Office :  
B/S, First Floor, Marimtas Complex, Tawane Colony,  
Acaryashwer Corner, Pune - 411009, Maharashtra, India  
Tel. : 91-20-24221234, 91-20-24225678

Our Distributors

Student's Agencies (I) Pvt. Ltd.  
T. : (022) 40496161, 91872 90777

Vidarthi Sales Agencies  
T. : (022) 23867279, 98197 76110

Our Branches : Pune | Mumbai | Kolhapur | Nagpur | Solapur | Nashik

Email : [info@techknowledgebooks.com](mailto:info@techknowledgebooks.com) Website : [www.techknowledgebooks.com](http://www.techknowledgebooks.com)

ISBN : 978-93-80694-65-1

Price : 295/-

GERIA

TechKnowledge Publications

**DATA MINING**

Strictly as per the New Revised Syllabus of  
**Gujarat Technological University**  
w.e.f. academic year 2020-2021

**Data Mining**

(Code : 3160714) (Professional Elective - II)


Semester VI - Computer Engineering /  
Computer Science and Engineering

**Dr. Arti Deshpande**  
**Dr. Pallavi Halarankar**

**Tech Knowledge Publications**

Includes :  
Solved Latest University Question Papers

New Edition  
GTU

  
**Dr. G. T. Thampi**  
**PRINCIPAL**  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



# Data Mining

(Code : 3160714) (Professional Elective-II)

Semester VI - Computer Engineering/Computer Science and Engineering,  
(Gujarat Technological University)

**Strictly as per the New Revised Syllabus of  
Gujarat Technological University w.e.f. academic year 2020-2021**

## Dr. Arti Deshpande

Assistant Professor,  
Department of Computer Engineering  
Thadomal Shahani Engineering College,  
Mumbai.  
Maharashtra, India

## Dr. Pallavi N. Halarnkar

Associate Professor,  
Department of Computer Engineering  
Thadomal Shahani Engineering College,  
Mumbai.  
Maharashtra, India.



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



# Data Mining (3160714)

Dr. Arti Deshpande, Dr. Pallavi N. Halarnkar

Semester VI – Computer Engineering/Computer Science and Engineering, (Gujarat Technological University)

*Copyright © by Authors. All rights reserved. No part of this publication may be reproduced, copied, or stored in a retrieval system, distributed or transmitted in any form or by any means, including photocopy, recording, or other electronic or mechanical methods, without the prior written permission of the publisher.*

*This book is sold subject to the condition that it shall not, by the way of trade or otherwise, be lent, resold, hired out, or otherwise circulated without the publisher's prior written consent in any form of binding or cover other than which it is published and without a similar condition including this condition being imposed on the subsequent purchaser and without limiting the rights under copyright reserved above.*

**First Printed In India** : April 2010 (Mumbai University)

**First Edition** : February 2021

*This edition is for sale in India, Bangladesh, Bhutan, Maldives, Nepal, Pakistan, Sri Lanka and designated countries in South-East Asia. Sale and purchase of this book outside of these countries is unauthorized by the publisher.*

**ISBN** : 978-93-90694-05-1

**Published By**

**TechKnowledge Publications**

## Printed @

37/2, Ashtavinayak Industrial Estate,  
Near Pari Company,  
Narhe, Pune, Maharashtra State, India.  
Pune - 411041

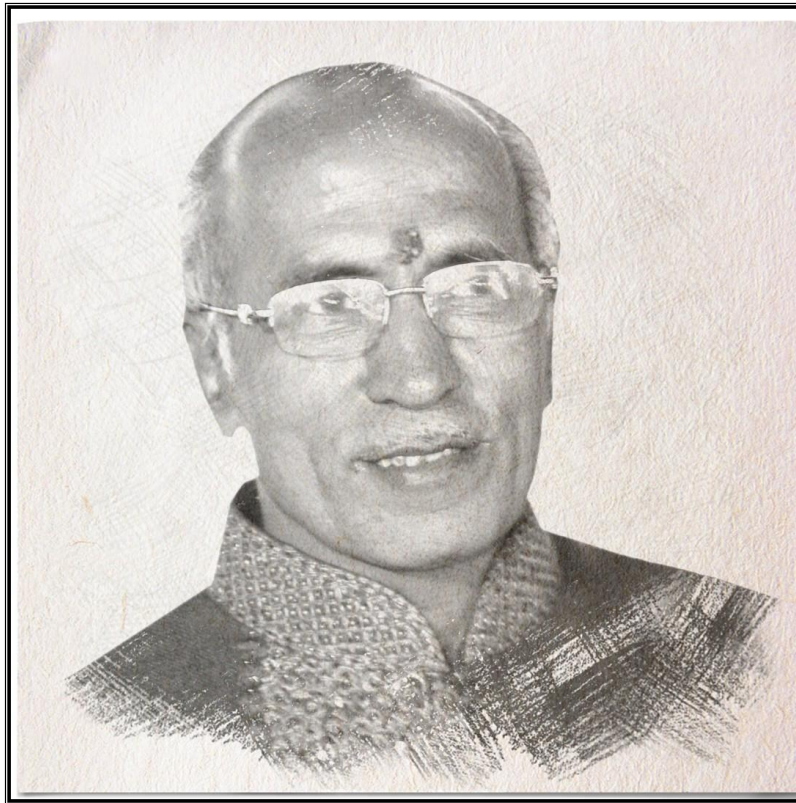
## Head Office

B/5, First floor, Mahatma Complex,  
Taware Colony, Aranyeshwar Corner,  
Pune - 411 009, Maharashtra State, India  
**Ph** : 91-20-24221234, 91-20-24225678.  
**Email** : info@techknowledgebooks.com,  
**Website** : www.techknowledgebooks.com

**Subject Code** : 3160714

**Book Code** : GE91A

*We dedicate this Publication soulfully and wholeheartedly,  
in loving memory of our beloved founder director,  
**Late Shri. Pradeepji Lalchandji Lunawat,**  
who will always be an inspiration, a positive force and strong support  
behind us.*



***“My work is my prayer to God”***

*- Lt. Shri. **Pradeepji L. Lunawat***

Dr. G. T. Thampi  
Thadomal Shamani Engineering College  
Bandra (W), Mumbai - 400 050.



***Soulful Tribute and Gratitude for all Your  
Sacrifices, Hardwork, and 40 years of Strong Vision...***





**Dr. G. T. Thampi**  
**PRINCIPAL**  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai-400 050.



## Preface

My Dear Students,

We are extremely happy to come out with this book on **Data Mining** for you. The topics within the chapters have been arranged in a proper sequence to ensure smooth flow of the subject.

We present this book in the loving memory of **Late Shri. Pradeepji Lunawat**, our source of inspiration and a strong foundation of **“TechKnowledge Publications”**. He will always be remembered in our heart and motivate us to achieve our milestone.

We are thankful to Shri. J. S. Katre, Shri. Shital Bhandari, Shri. Arunoday Kumar and Shri. Chandroday Kumar for the encouragement and support that they have extended. We are also thankful to Seema Lunawat for technology enhanced reading, E-books support and the staff members of TechKnowledge Publications for their efforts to make this book as good as it is. We have jointly made every possible efforts to eliminate all the errors in this book. However if you find any, please let us know, because that will help us to improve further.

We are also thankful to our family members and friends for their patience and encouragement.

  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.

 - Authors

# SYLLABUS

## Gujarat Technological University Sixth Semester of Computer Engineering / Computer Science and Engineering Data Mining (Code : 3160714)

### Teaching and Examination Scheme

Teaching scheme			Credits	Examination Marks				Total Marks	
L	T	P	C	Theory Marks		Practical Marks			
				ESE (E)	PA (M)		ESE (V)		PA (I)
3	0	2	4	70	30		30	20	150

Sr. No.	Content	Total Hours
1.	<b>Introduction to data mining (DM) :</b> Motivation for Data Mining - Data Mining-Definition and Functionalities – Classification of DM Systems - DM task primitives - Integration of a Data Mining system with a Database or a Data Warehouse - Issues in DM – KDD Process <b>(Refer Chapter 1)</b>	3
2.	<b>Data Pre-processing :</b> Data summarization, data cleaning, data integration and transformation, data reduction, data discretization and concept hierarchy generation, feature extraction, feature transformation, feature selection, introduction to Dimensionality Reduction, CUR decomposition <b>(Refer Chapter 2)</b>	4
3.	<b>Concept Description, Mining Frequent Patterns, Associations and Correlations :</b> What is concept description? - Data Generalization and summarization-based characterization - Attribute relevance - class comparisons, Basic concept, efficient and scalable frequent item-set mining methods, mining various kind of association rules, from association mining to correlation analysis, Advanced Association Rule Techniques, Measuring the Quality of Rules. <b>(Refer Chapter 3)</b>	10
4.	<b>Classification and Prediction :</b> Classification vs. prediction, Issues regarding classification and prediction, Statistical-Based Algorithms, Distance-Based Algorithms, Decision Tree-Based Algorithms, Neural Network-Based Algorithms, Rule-Based Algorithms, Combining Techniques, accuracy and error measures, evaluation of the accuracy of a classifier or predictor. Neural Network Prediction methods: Linear and nonlinear regression, Logistic Regression Introduction of tools such as DB Miner / WEKA / DTREG DM Tools <b>(Refer Chapter 4)</b>	10
5.	<b>Cluster Analysis :</b> Clustering: Problem Definition, Clustering Overview, Evaluation of Clustering Algorithms, Partitioning Clustering -K-Means Algorithm, K-Means Additional issues, PAM Algorithm; Hierarchical Clustering – Agglomerative Methods and divisive methods, Basic Agglomerative Hierarchical Clustering, Strengths and Weakness; Outlier Detection, Clustering high dimensional data, clustering Graph and Network data. <b>(Refer Chapter 5)</b>	10
8.	<b>Web mining and other Data Mining :</b> Web Mining: Introduction to Web Mining, Web content mining, Web usage mining, Web Structure mining, Web log structure and issues regarding web logs, Spatial Data Mining, Temporal Mining, And Multimedia Mining. Applications of Distributed and parallel Data Mining. <b>(Refer Chapter 6)</b>	5

□□□

**Your Success is Our Goal**

**Semester VI - Computer Engineering**

**SOFTWARE ENGINEERING**  
Dr. Pathan Mohd Shafi, Rashmi B. Kale, Arti A. Bhise, Prasad B. Chaudhari, M. A. Ansari


**SYSTEM PROGRAMMING & COMPLIER CONSTRUCTION**  
Shweta A. Loonkar

**DATA WAREHOUSING & MINING**  
Dr. Arti Deshpande, Dr. Pallavi Halarankar

**CRYPTOGRAPHY & SYSTEM SECURITY**  
Pravin Goyal

**ADVANCED DATABASE MANAGEMENT SYSTEM (Department Level Elective)**  
Mahesh Mali

**ADVANCE COMPUTER NETWORK (Department Level Elective)**  
J. S. Katre



**2020 MU**

Strictly as per the New Revised Syllabus (Rev - 2016) of **Mumbai University** w.e.f. academic year 2018-2019 (As per Choice Based Credit and Grading System)

## Data Warehousing and Mining

**Semester VI - Computer Engineering**

**Same Subject, Same Authors with New Publication**

**Dr. Arti Deshpande      Dr. Pallavi Halarankar**

With Solved Latest University Question Paper of May 2019.

**TechKnowledge Publications**

coming soon.....

**easy-solutions** now with **TechKnowledge Publications**

Paper Solutions Trusted by lakhs of students from more than 15 years

**2020 MU**

**DATA WAREHOUSING AND MINING**

Semester VI - Computer Engineering

Dr. Arti Deshpande  
Dr. Pallavi Halarankar

Book Code ME40A  
Price ₹ 325/-

Head Office :  
B-5, First Floor, Maniratra Complex, Taware Colony,  
Aranyeshwar Corner, Pune - 411009, Maharashtra, India.  
Tel. : 91-20-24221234, 91-20-24225978

**Distributors**

Student's Agencies (S. Pvt. Ltd.)      Vidyarthi Sales Agencies      Bharat Sales Agency  
T. : (020) 4440164, 041872 80777      T. : (020) 2380729, 043037 78110      T. : (020) 2381326, 040572 82787

**Our Branches : Pune | Mumbai | Kolhapur | Nagpur | Solapur | Nashik**

For Library Orders Contact - Ved Book Distributors      M : 80875 71421 / 82208 77214

Email : info@techknowledgebooks.com      Website : www.techknowledgebooks.com


ISBN : 978-83-93748-80-1

Price ₹ 325/-

ME40A

USE QR CODE

TechKnowledgePublications

  
**Dr. G. T. Thampi**  
**PRINCIPAL**  
 Thadomal Shahani Engineering College  
 Bandra (W), Mumbai - 400 050.



# Data Warehousing and Mining

(Code - CSC603)

Semester VI - Computer Engineering

(Mumbai University)

Strictly as per the Choice Based Credit and Grading System  
(Revise 2016) of Mumbai University w.e.f. academic year 2018-2019

## Dr. Arti Deshpande

Department of Computer Engineering  
Thadomal Shahani Engineering College,  
Mumbai.  
Maharashtra, India

## Dr. Pallavi Halarnkar

Department of Computer Engineering  
Thadomal Shahani Engineering College,  
Mumbai.  
Maharashtra, India.



Dr. G. T. Thampi  
PRINCIPAL

Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



ME40A Price ₹ 325/-



( Book Code : ME40A)

## **Data Warehousing and Mining**

Dr. Arti Deshpande, Dr. Pallavi Halarnkar

(Semester VI - Computer Engineering, Mumbai University)

Copyright © by Authors. All rights reserved. No part of this publication may be reproduced, copied, or stored in a retrieval system, distributed or transmitted in any form or by any means, including photocopy, recording, or other electronic or mechanical methods, without the prior written permission of the publisher.

This book is sold subject to the condition that it shall not, by the way of trade or otherwise, be lent, resold, hired out, or otherwise circulated without the publisher's prior written consent in any form of binding or cover other than which it is published and without a similar condition including this condition being imposed on the subsequent purchaser and without limiting the rights under copyright reserved above.

**First Edition** : February 2019

**Second Revised Edition** : January 2020 (**TechKnowledge Publications**)

This edition is for sale in India, Bangladesh, Bhutan, Maldives, Nepal, Pakistan, Sri Lanka and designated countries in South-East Asia. Sale and purchase of this book outside of these countries is unauthorized by the publisher.

**Printed at** : 37/2, Ashtavinayak Industrial Estate,  
Near Pari Company, Narhe, Pune, Maharashtra State, India.  
Pune – 411041.


**ISBN** : 978-93-89748-90-1

**Published by**

**TechKnowledge Publications**

**Head Office** : B/5, First floor, Maniratna Complex, Taware Colony, Aranyeshwar Corner,  
Pune - 411 009. Maharashtra State, India  
Ph : 91-20-24221234, 91-20-24225678.  
Email : info@techknowledgebooks.com,  
Website : www.techknowledgebooks.com

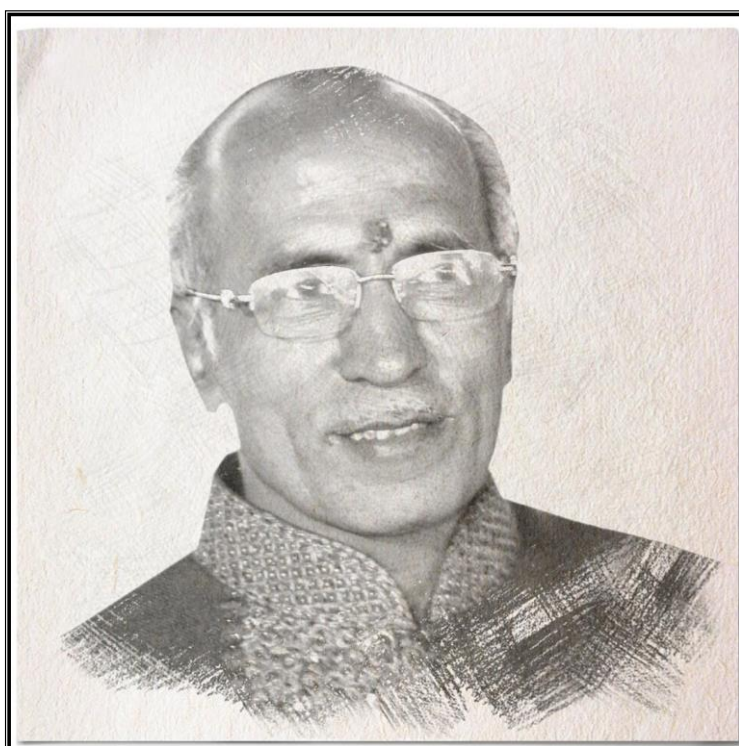
[CSC603] (FID : ME40) ( Book Code : ME40A)

  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



( Book Code : ME40A)

*We dedicate this Publication soulfully and wholeheartedly, in loving memory of our beloved founder director, **Late Shri. Pradeepji Lalchandji Lunawat**, who will always be an inspiration, a positive force and strong support behind us.*



***“My work is my prayer to God”***

*- Lt. Shri. Pradeepji L. Lunawat*

***Soulful Tribute and Gratitude for all Your Sacrifices, Hardwork, and 40 years of Strong Vision...***

Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



( Book Code : ME40A)



## **Preface**

Dear Students,

We are extremely happy to present the book of **“Data Warehousing and Mining”** for you. We have divided the subject into small chapters so that the topics can be arranged and understood properly. The topics within the chapters have been arranged in a proper sequence to ensure smooth flow of the subject.

We present this book in the loving memory of **Late. Shri. Pradeepji Lunawat**, our source of inspiration and a strong foundation of **“TechKnowledge Publications”**. He will always be remembered in our hearts and motivate us to achieve our new milestone.

We are thankful to Prof. Arunoday Kumar, Mr. Shital Bhandari and Shri. Chandroday Kumar for the encouragement and support that they have extended. We also thankful to the staff members of TechKnowledge Publications for their efforts to make this book as good as it is. We have made every possible efforts to eliminate all the errors in this book. However if you find any, please let us know, because that will help us to improve the book quality further.

We are thankful to my family members and friends for their patience and encouragement.

  
**Dr. G. I. T. Authors**  
**PRINCIPAL**  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



( Book Code : ME40A)

## Syllabus

Course Code	Course Name	Credits
CSC603	Data Warehousing and Mining	4

### Course Objectives :

1. To identify the scope and essentiality of Data Warehousing and Mining.
2. To analyze data, choose relevant models and algorithms for respective applications.
3. To study spatial and web data mining.
4. To develop research interest towards advances in data mining.

### Course Outcomes : On successful completion of course learner will be able to :


1. Understand Data Warehouse fundamentals, Data Mining Principles.
2. Design data warehouse with dimensional modelling and apply OLAP operations.
3. Identify appropriate data mining algorithms to solve real world problems.
4. Compare and evaluate different data mining techniques like classification, prediction, clustering and association rule mining.
5. Describe complex data types with respect to spatial and web mining.
6. Benefit the user experiences towards research and innovation.



**Prerequisite :** Basic database concepts, Concepts of algorithm design and analysis.

Module No.	Topics	Hrs.
1.0	<b>Introduction to Data Warehouse and Dimensional modeling :</b> Introduction to Strategic Information, Need for Strategic Information, Features of Data Warehouse, Data warehouses versus Data Marts, Top-down versus Bottom-up approach. Data warehouse architecture, metadata, E-R modelling versus Dimensional Modelling, Information Package Diagram, STAR schema, STAR schema keys, Snowflake Schema, Fact Constellation Schema, Factless Fact tables, Update to the dimension tables, Aggregate fact tables. <b>(Refer chapter 1)</b>	8
2.0	<b>ETL Process and OLAP :</b> Major steps in ETL process, Data extraction : Techniques, Data transformation : Basic tasks, Major transformation types, Data Loading : Applying Data, OLTP Vs OLAP, OLAP definition, Dimensional Analysis, Hypercubes, OLAP operations : Drill down, Roll up, Slice, Dice and Rotation, OLAP models : MOLAP, ROLAP. <b>(Refer chapter 2)</b>	8

( Book Code : ME40A)

Module No.	Topics	Hrs.
3.0	<b>Introduction to Data Mining, Data Exploration and Preprocessing :</b> Data Mining Task Primitives, Architecture, Techniques, KDD process, Issues in Data Mining, Applications of Data Mining, Data Exploration : Types of Attributes, Statistical Description of Data, Data Visualization, Data Preprocessing : Cleaning, Integration, Reduction : Attribute subset selection, Histograms, Clustering and Sampling, Data Transformation and Data Discretization : Normalization, Binning, Concept hierarchy generation, Concept Description : Attribute oriented Induction for Data Characterization. <b>(Refer chapter 3)</b>	10
4.0	<b>Classification, Prediction and Clustering :</b> Basic Concepts, Decision Tree using Information Gain, Induction : Attribute Selection Measures, Tree pruning, Bayesian Classification : Naive Bayes, Classifier Rule - Based Classification : Using IF-THEN Rules for classification, Prediction : Simple linear regression, Multiple linear regression Model Evaluation and Selection : Accuracy and Error measures, Holdout, Random Sampling, Cross Validation, Bootstrap, Clustering : Distance Measures, Partitioning Methods (k-Means, k-Medoids), Hierarchical Methods (Agglomerative, Divisive). <b>(Refer chapter 4)</b>	12
5.0	<b>Mining Frequent Patterns and Association Rules :</b> Market Basket Analysis, Frequent Item sets, Closed Item sets and Association Rule, Frequent Pattern Mining, Efficient and Scalable Frequent Item set Mining Methods : Apriori Algorithm, Association Rule Generation, Improving the Efficiency of Apriori, FP growth, Mining frequent Itemsets using Vertical Data Format, Introduction to Mining Multilevel Association Rules and Multidimensional Association Rules. <b>(Refer chapter 5)</b>	8
6.0	<b>Spatial and Web Mining :</b> Spatial Data, Spatial Vs. Classical Data Mining, Spatial Data Structures, Mining Spatial Association and Co-location Patterns, Spatial Clustering Techniques : CLARANS Extension, Web Mining : Web Content Mining, Web Structure Mining, Web Usage mining, Applications of Web Mining. <b>(Refer chapter 6)</b>	6
	<b>Total</b>	<b>52</b>

  
**Dr. G. T. Thampi**  
**PRINCIPAL**  
 Thadomal Shahani Engineering College  
 Bandra (W), Mumbai - 400 050.

( Book Code : ME40A)



# Data Mining and Warehousing

## Elective I

(Code : 410244(D))

Semester VII - Computer Engineering

(Savitribai Phule Pune University)

Strictly as per the New Credit System Syllabus (2015 Course)  
Savitribai Phule Pune University w.e.f. academic year 2018-2019

### Dr. Arti Deshpande

Assistant Professor,  
Department of Computer Engineering  
Thadomal Shahani Engineering College, Mumbai.  
Maharashtra, India.

### Dr. Pallavi N. Halarnkar

Associate Professor,  
Department of Computer Engineering  
Thadomal Shahani Engineering College, Mumbai.  
Maharashtra, India.



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



 **TechKnowledge**<sup>TM</sup>  
Publications

P078A Price ₹ 175/-



## **Data Mining and Warehousing**

Dr. Arti Deshpande, Dr. Pallavi N. Halarnkar

(Semester VII - Computer Engineering) (Savitribai Phule Pune University)

Copyright © by Authors. All rights reserved. No part of this publication may be reproduced, copied, or stored in a retrieval system, distributed or transmitted in any form or by any means, including photocopy, recording, or other electronic or mechanical methods, without the prior written permission of the publisher.

This book is sold subject to the condition that it shall not, by the way of trade or otherwise, be lent, resold, hired out, or otherwise circulated without the publisher's prior written consent in any form of binding or cover other than which it is published and without a similar condition including this condition being imposed on the subsequent purchaser and without limiting the rights under copyright reserved above.

**First Edition** : July 2018

**Second Revised Edition** : July 2019 (TechKnowledge Publications)


This edition is for sale in India, Bangladesh, Bhutan, Maldives, Nepal, Pakistan, Sri Lanka and designated countries in South-East Asia. Sale and purchase of this book outside of these countries is unauthorized by the publisher.

**Printed at ::** 37/2, Ashtvinayak Industrial Estate, Near Pari Company,  
Narhe, Pune, Maharashtra State India, Pune - 411041

**ISBN** 978-93-89299-36-6

**Published by**  
**TechKnowledge Publications**

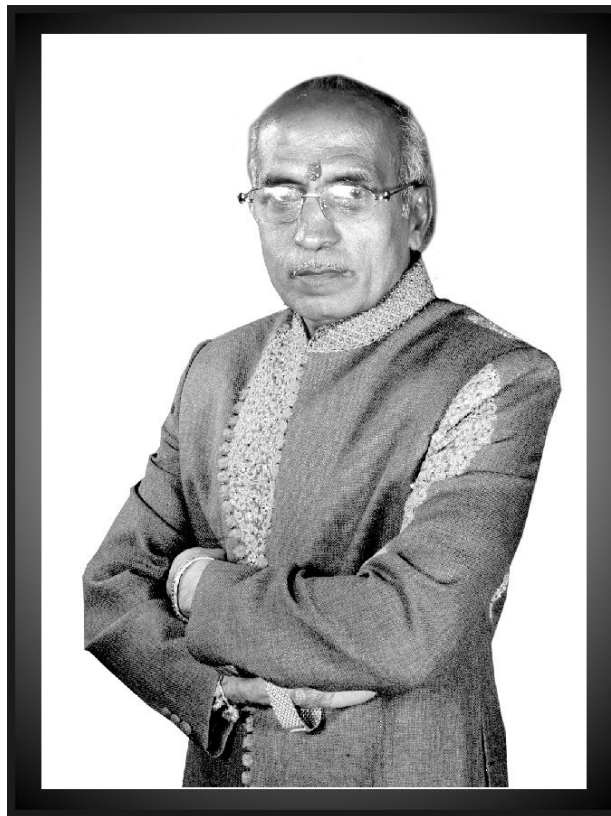
**Head Office :** B/5, First floor, Maniratna Complex,  
Taware Colony, Aranyeshwar Corner,  
Pune - 411 009. Maharashtra State,  
India. Ph : 91-20-24221234, 91-20-24225678.

  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



[410244(D)] (FID : PO78) (Book Code : PO78A)

*We dedicate this Publication soulfully and wholeheartedly,  
in loving memory of our beloved founder director  
**Late. Shri. Pradeepsheth Lalchandji Lunawat,**  
who will always be an inspiration, a positive force and strong support  
behind us.*



*Lt. Shri. Pradeepji L. Lunawat*

Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra, Mumbai - 400 050



*Soulful Tribute and Gratitude for all Your  
Sacrifices, Hardwork and 40 years of Strong Vision.....*



## Preface



Dear Students,

We are extremely happy to present the book of “**Data Mining and Warehousing**” for you. We have divided the subject into small chapters so that the topics can be arranged and understood properly. The topics within the chapters have been arranged in a proper sequence to ensure smooth flow of the subject.

We present this book in the loving memory of **Late. Shri. Pradeepji Lunawat**, our source of inspiration and a strong foundation of “**TechKnowledge Publications**”. He will always be remembered in our heart and motivate us to achieve our milestone.

We are thankful to Shri. J. S. Katre, Shri. Shital Bhandari, Shri. Arunoday Kumar and Shri. Chandroday Kumar for the encouragement and support that they have extended. We are also thankful to the staff members of TechKnowledge Publications and others for their efforts to make this book as good as it is.

We have jointly made every possible effort to eliminate all the errors in this book. However if you find any, please let us know, because that will help us to improve further.

  
**Dr. Arti Deshpande**  
  
**Dr. Pallavi Halarnkar**  
Dr. G. T. Halarnkar  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.

□□□

## Syllabus

### Savitribai Phule Pune University Fourth Year of Computer Engineering (2015 Course) Elective I

#### 410244(D) : Data Mining and Warehousing

<b>Teaching Scheme :</b> TH : 03 Hours/Week	<b>Credit</b> 03	<b>Examination Scheme :</b> In-Sem (Paper) : 30 Marks End-Sem (Paper) : 70 Marks
--	---------------------	--

#### Pre-requisites Courses

310242-Database Management Systems, 310244 - Information Systems and Engineering Economics

**Companion Course :** 410247- Laboratory Practice II

#### Course Objectives

- To understand the fundamentals of Data Mining.
- To identify the appropriateness and need of mining the data.
- To learn the preprocessing, mining and post processing of the data.
- To understand various methods, techniques and algorithms in data mining.

#### Course Outcomes

On completion of the course the student should be able to :

- Apply basic, intermediate and advanced techniques to mine the data.
- Analyze the output generated by the process of data mining.
- Explore the hidden patterns in the data.
- Optimize the mining process by choosing best data mining technique.

### Course Contents

#### Unit I : Introduction

(08 Hours)

Data Mining, Data Mining Task Primitives, Data : Data, Information and Knowledge; Attribute Types : Nominal, Binary, Ordinal and Numeric attributes, Discrete versus Continuous Attributes; Introduction to Data Preprocessing, Data Cleaning : Missing values, Noisy data; Data integration : Correlation analysis; transformation : Min-max normalization, z-score normalization and decimal scaling; data reduction : Data Cube Aggregation, Attribute Subset Selection, sampling; and Data Discretization : Binning, Histogram Analysis

Dr. G. T. Thampi  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.

(Refer chapter 1)

#### Unit II : Data Warehouse

(08 Hours)

Data Warehouse, Operational Database Systems and Data Warehouses (OLTP Vs OLAP), A Multidimensional Data Model: Data Cubes, Stars, Snowflakes, and Fact Constellations Schemas; OLAP Operations in the Multidimensional Data Model, Concept Hierarchies, Data Warehouse Architecture, The Process of Data Warehouse Design, A three-tier data warehousing architecture, Types of OLAP Servers : ROLAP versus MOLAP versus HOLAP.

(Refer chapter 2)

### **Unit III : Measuring Data Similarity and Dissimilarity**

**(08 Hours)**

Measuring Data Similarity and Dissimilarity, Proximity Measures for Nominal Attributes and Binary Attributes, interval scaled; Dissimilarity of Numeric Data : Minkowski Distance, Euclidean distance and Manhattan distance; Proximity Measures for Categorical, Ordinal Attributes, Ratio scaled variables; Dissimilarity for Attributes of Mixed Types, Cosine Similarity.

**(Refer chapter 3)**

### **Unit IV : Association Rules Mining**

**(08 Hours)**

Market basket Analysis, Frequent item set, Closed item set, Association Rules, a-priori Algorithm, Generating Association Rules from Frequent Item sets, Improving the Efficiency of a-priori, Mining Frequent Item sets without Candidate Generation : FP Growth Algorithm; Mining Various Kinds of Association Rules : Mining multilevel association rules, constraint based association rule mining, Meta rule-Guided Mining of Association Rules.

**(Refer chapter 4)**

### **Unit V : Classification**

**(08 Hours)**

Introduction to : Classification and Regression for Predictive Analysis, Decision Tree Induction, Rule-Based Classification : using IF-THEN Rules for Classification, Rule Induction Using a Sequential Covering Algorithm. Bayesian Belief Networks, Training Bayesian Belief Networks, Classification Using Frequent Patterns, Associative Classification, Lazy Learners-k-Nearest-Neighbor Classifiers, Case-Based Reasoning.

**(Refer chapter 5)**

### **Unit VI : Multiclass Classification**

**(08 Hours)**

Multiclass Classification, Semi-Supervised Classification, Reinforcement learning, Systematic Learning, Wholistic learning and multi-perspective learning. Metrics for Evaluating Classifier Performance : Accuracy, Error Rate, precision, Recall, Sensitivity, Specificity; Evaluating the Accuracy of a Classifier : Holdout Method, Random Sub sampling and Cross-Validation.

**(Refer chapter 6)**

□□□



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



"Your Success is Our Goal"

**es<sup>®</sup>**  
**easy-solutions**

now with

**TechKnowledge<sup>™</sup>**  
**Publications**

**Study, Learn & Excel in Examinations**

**Includes:**

- University Paper Solved by Eminent Faculty Members.
- Unit-wise distribution of solved questions.
- Statistical Analysis of each unit.
- Up-to-date and Quality contents.
- In-line with the books of TechKnowledge Publications.

**Head Office :**  
B/5, First Floor, Maniratra Complex, Taware Colony, Aranyeshwar Corner,  
Pune - 411009, Maharashtra, India. Tel. : 91-20-24221234, 91-20-24225678

**Our Distributors**

**Student's Agencies (I) Pvt. Ltd.**  
T. : (022) 40406101, © 91072 90777

**Vidyarthi Sales Agencies**  
T. : (022) 23867279, © 96197 76110

**Our Branches : Pune | Mumbai | Kolhapur | Nagpur | Solapur | Nashik**

Email : info@techknowledgebooks.com | Website : www.techknowledgebooks.com

ISBN : 978-93-89503-32-6

Price ₹ 175/-

GO46A

TechKnowledgePublications

DATA MINING AND BUSINESS INTELLIGENCE

Dr. Arti Deshpande

Price ₹ 175/-

TechKnowledgePublications



2019  
GTU

Strictly as per the New Revised syllabus of  
Gujarat Technological University  
w. e. f. academic year 2016-2017

# Data Mining and Business Intelligence

(Departmental Elective II)

Semester VII - Information Technology / Computer Engineering &  
Computer Science Engineering

Same Subject, Same Authors with New Publication

Arti Deshpande

Dr. Pallavi N. Halarnkar

Chapterwise Solved University Question Papers Upto Dec. 2018.

**TechKnowledge<sup>™</sup>**  
**Publications**

*Handwritten signature*

Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



# Data Mining and Business Intelligence

(Code - 2170715)

**Semester VII - Information Technology / Computer Engineering & Computer Science Engineering (Departmental Elective-II)**  
(Gujarat Technological University)

**Strictly as per the New Revised Syllabus of  
Gujarat Technological University w.e.f. academic year 2016-2017**

## **Mrs. Arti Deshpande**

ME (Comp. Engg.)

Thadomal Shahani Engineering College , Mumbai.

## **Dr. Pallavi N. Halarnkar**

ME (Comp. Engg.)

Mukesh Patel School of Technology, Management and Engineering, Mumbai.



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



**G046A Price ₹ 175/-**



(Book Code : G046A)



## **Data Mining and Business Intelligence**

Mrs. Arti Deshpande, Dr. Pallavi N. Halarnkar

(Semester VII - Information Technology/ Computer Engineering &  
Computer Science Engineering : Departmental Elective-II, GTU)

Copyright © by Authors. All rights reserved. No part of this publication may be reproduced, copied, or stored in a retrieval system, distributed or transmitted in any form or by any means, including photocopy, recording, or other electronic or mechanical methods, without the prior written permission of the publisher.

This book is sold subject to the condition that it shall not, by the way of trade or otherwise, be lent, resold, hired out, or otherwise circulated without the publisher's prior written consent in any form of binding or cover other than which it is published and without a similar condition including this condition being imposed on the subsequent purchaser and without limiting the rights under copyright reserved above.

**First Printed In India** : April 2010 (Mumbai University)  
**First Edition** : July 2016  
**Second Revised Edition** : June 2017  
**Third Revised Edition** : June 2018  
**Fourth Revised Edition** : October 2019 (TechKnowledge Publications)

This edition is for sale in India, Bangladesh, Bhutan, Maldives, Nepal, Pakistan, Sri Lanka and designated countries in South-East Asia. Sale and purchase of this book outside of these countries is unauthorized by the publisher.

**Printed at :** 37/2, Ashtvinayak Industrial Estate, Near Pari Company,  
Narhe, Pune, Maharashtra State India,  
Pune – 411041

**ISBN** 978-93-89503-32-6

**Published by**  
**TechKnowledge Publications**

**Head Office :** B/5, First floor, Maniratna Complex, Taware Colony, Aranyeshwar Colony,  
Pune - 411 009. Maharashtra State, India  
Ph : 91-20-24221234, 91-20-24225678.



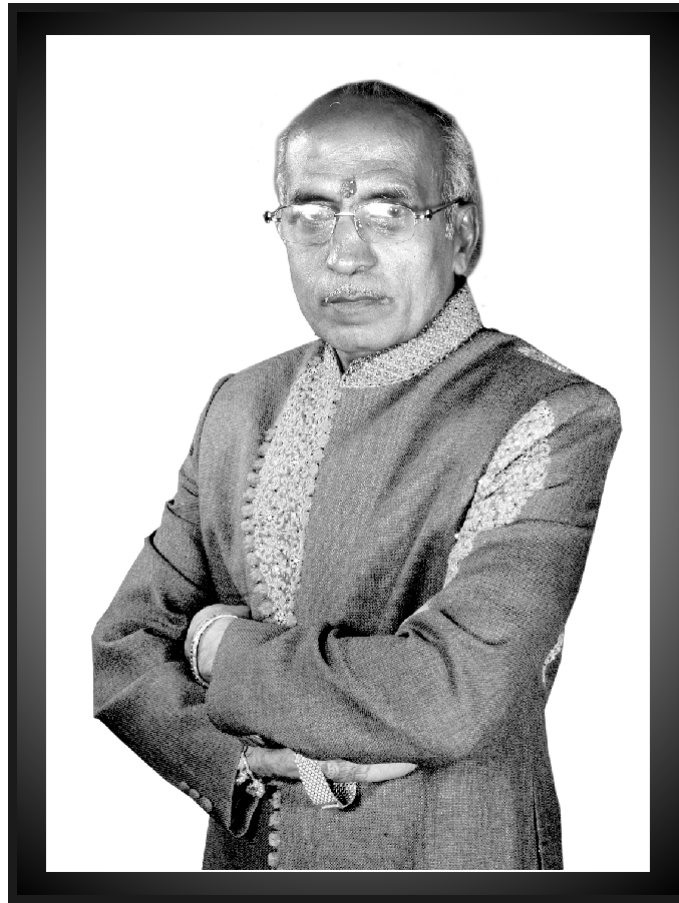
Dr. G. T. Thampi  
Principal  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



[2K16GTUCOMP29] (FID : GO46) (Book Code : GO46A)

(Book Code : GO46A)

*We dedicate this Publication soulfully and wholeheartedly,  
in loving memory of our beloved founder director,  
**Late Shri. Pradeepsheth Lalchandji Lunawat,**  
who will always be an inspiration, a positive force and strong support  
behind us.*



***Lt. Shri. Pradeepji L. Lunawat***

Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



***Soulful Tribute and Gratitude for all Your  
Sacrifices, Hardwork and 40 years of Strong Vision...***

## Preface

My Dear Students,

We are extremely happy to come out with this book on **“Data Mining & Business Intelligence”** for you. The topics within the chapters have been arranged in a proper sequence to ensure smooth flow of the subject.

We present this book in the loving memory of **Late Shri. Pradeepji Lunawat**, our source of inspiration and a strong foundation of **“TechKnowledge Publications”**. He will always be remembered in our heart and motivate us to achieve our milestone.

We are thankful to Shri. J. S. Katre, Mr. Shital Bhandari, Shri. Arunoday Kumar and Shri. Chandroday Kumar for the encouragement and support that they have extended. We are also thankful to the staff members of TechKnowledge Publications and others for their efforts to make this book as good as it is. We have jointly made every possible efforts to eliminate all the errors in this book. However if you find any, please let us know, because that will help us to improve further.

We are also thankful to my family members and friends for patience and encouragement.

  
Dr. G. T. Thakur  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 058  
  
**Arti Deshpande**  
**Pallavi N. Halankar**



# Syllabus

## Unit 1 : Overview and concepts Data Warehousing and Business Intelligence :

Why reporting and Analysing data, Raw data to valuable information-Lifecycle of Data - What is Business Intelligence - BI and DW in today's perspective - What is data warehousing - The building Blocks : Defining Features - Data warehouses and data marts - Overview of the components - Metadata in the data warehouse - Need for data warehousing - Basic elements of data warehousing - trends in data warehousing. (Refer Chapter 1)

## Unit 2 : The Architecture of BI and DW :

BI and DW architectures and its types - Relation between BI and DW - OLAP (Online analytical processing) definitions - Difference between OLAP and OLTP - Dimensional analysis - What are cubes? Drill-down and roll-up - slice and dice or rotation - OLAP models - ROLAP versus MOLAP - defining schemas : Stars, snowflakes and fact constellations. (Refer Chapter 2)

## Unit 3 : Introduction to data mining (DM) :

Motivation for Data Mining - Data Mining-Definition and Functionalities – Classification of DM Systems - DM task primitives - Integration of a Data Mining system with a Database or a Data Warehouse - Issues in DM – KDD Process. (Refer Chapter 3)

## Unit 4 : Data Pre-processing :

Why to pre-process data? - Data cleaning: Missing Values, Noisy Data - Data Integration and transformation - Data Reduction : Data cube aggregation, Dimensionality reduction - Data Compression - Numerosity Reduction - Data Mining Primitives - Languages and System Architectures : Task relevant data - Kind of Knowledge to be mined - Discretization and Concept Hierarchy. (Refer Chapter 4)

## Unit 5 : Concept Description and Association Rule Mining :

What is concept description? - Data Generalization and summarization-based characterization - Attribute relevance - class comparisons Association Rule Mining: Market basket analysis – basic concepts - Finding frequent item sets: Apriori algorithm - generating rules – Improved Apriori algorithm – Incremental ARM – Associative Classification – Rule Mining. (Refer Chapter 5)

## Unit 6 : Classification and Prediction :

What is classification and prediction? – Issues regarding Classification and prediction :

**Classification methods :** Decision tree, Bayesian Classification, Rule based, CART, Neural Network

**Prediction methods :** Linear and nonlinear regression, Logistic Regression

Introduction of tools such as DB Miner /WEKA/DTREG DM Tools.

(Refer Chapter 6)



### Unit 7 : Data Mining for Business Intelligence Applications :

Data mining for business Applications like Balanced Scorecard, Fraud Detection, Click stream Mining, Market Segmentation, retail industry, telecommunications industry, banking & finance and CRM etc.

**Data Analytics Life Cycle :** Introduction to Big data Business Analytics - State of the practice in analytics role of data scientists

**Key roles for successful analytic project :** Main phases of life cycle - Developing core deliverables for stakeholders. **(Refer Chapter 7)**


### Unit 8 : Advance topics :

Introduction and basic concepts of following topics.

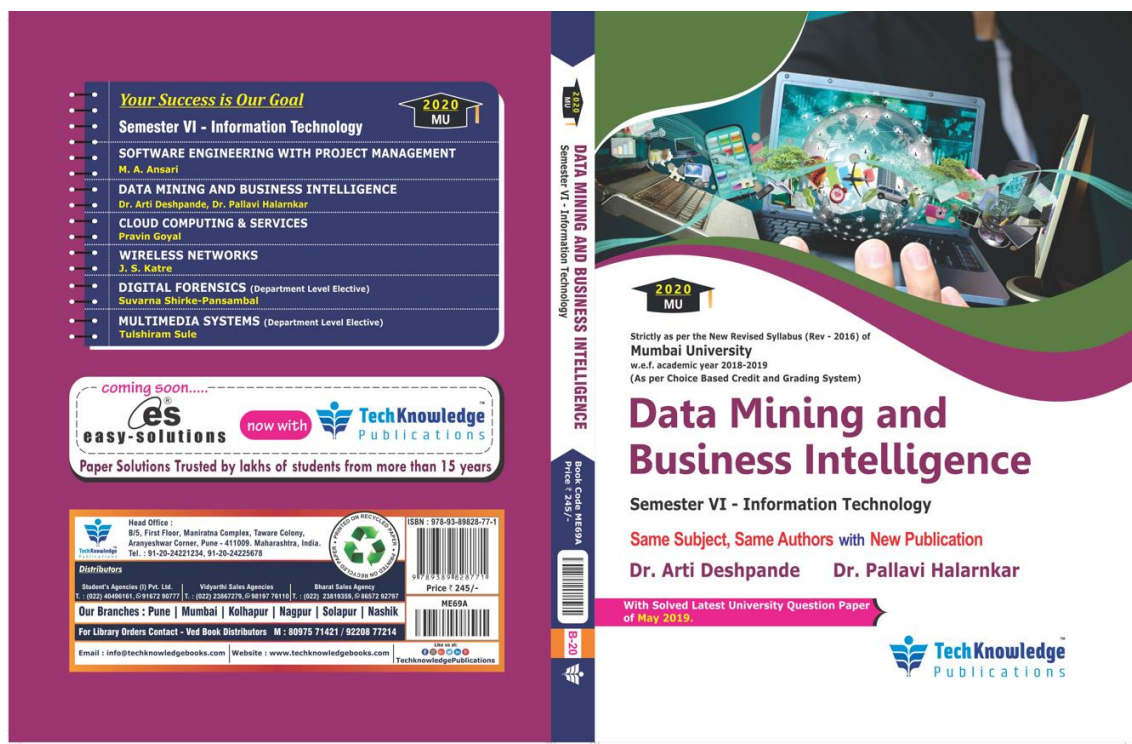
Clustering, Spatial mining, web mining, text mining,

**Big Data :** Introduction to big data : distributed file system – Big Data and its importance, Four Vs, Drivers for Big data, Big data analytics, Big data applications. Algorithms using map reduce, Matrix-Vector Multiplication by Map Reduce. Introduction to Hadoop architecture: Hadoop Architecture, Hadoop Storage: HDFS, Common Hadoop Shell commands , Anatomy of File Write and Read., NameNode, Secondary NameNode, and DataNode, Hadoop MapReduce paradigm, Map and Reduce tasks, Job, Task trackers – Cluster Setup – SSH & Hadoop Configuration – HDFS Administering – Monitoring & Maintenance. **(Refer Chapter 8)**

□□□

  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.





  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.





# Data Mining and Business Intelligence

(Code - ITC602)

Semester VI - Information Technology

(Mumbai University)

Strictly as per the Choice Based Credit and Grading System  
(Revise 2016) of Mumbai University w.e.f. academic year 2018-2019

**Dr. Arti Deshpande**

Department of Computer Engineering  
Thadomal Shahani Engineering College,  
Mumbai.  
Maharashtra, India.

**Dr. Pallavi Halarnkar**

Ph.D (Computer Engineering)



Dr. G. T. Thampi  
PRINCIPAL

Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



**TechKnowledge**  
Publications

ME69A Price ₹ 245/-



(Book Code : ME69A)

## **Data Mining and Business Intelligence**

Dr. Arti Deshpande, Dr. Pallavi Halarnkar

(Semester VI - Information Technology, Mumbai University)

Copyright © by Authors. All rights reserved. No part of this publication may be reproduced, copied, or stored in a retrieval system, distributed or transmitted in any form or by any means, including photocopy, recording, or other electronic or mechanical methods, without the prior written permission of the publisher.

This book is sold subject to the condition that it shall not, by the way of trade or otherwise, be lent, resold, hired out, or otherwise circulated without the publisher's prior written consent in any form of binding or cover other than which it is published and without a similar condition including this condition being imposed on the subsequent purchaser and without limiting the rights under copyright reserved above.

**First Edition** : February 2019

**Second Revised Edition** : February 2020 (**TechKnowledge Publications**)


This edition is for sale in India, Bangladesh, Bhutan, Maldives, Nepal, Pakistan, Sri Lanka and designated countries in South-East Asia. Sale and purchase of this book outside of these countries is unauthorized by the publisher.

**Printed at :** 37/2, Ashtavinayak Industrial Estate,  
Near Pari Company, Narhe, Pune, Maharashtra State, India.  
Pune – 411041.

**ISBN :** 978-93-89828-77-1

**Published by : TechKnowledge Publications**

**Head Office :** B/5, First floor, Maniratna Complex, Taware Colony, Aranyeshwar Corner,  
Pune - 411 009. Maharashtra State, India  
Ph : 91-20-24221234, 91-20-24225678.  
Email : info@techknowledgebooks.com,  
Website : www.techknowledgebooks.com

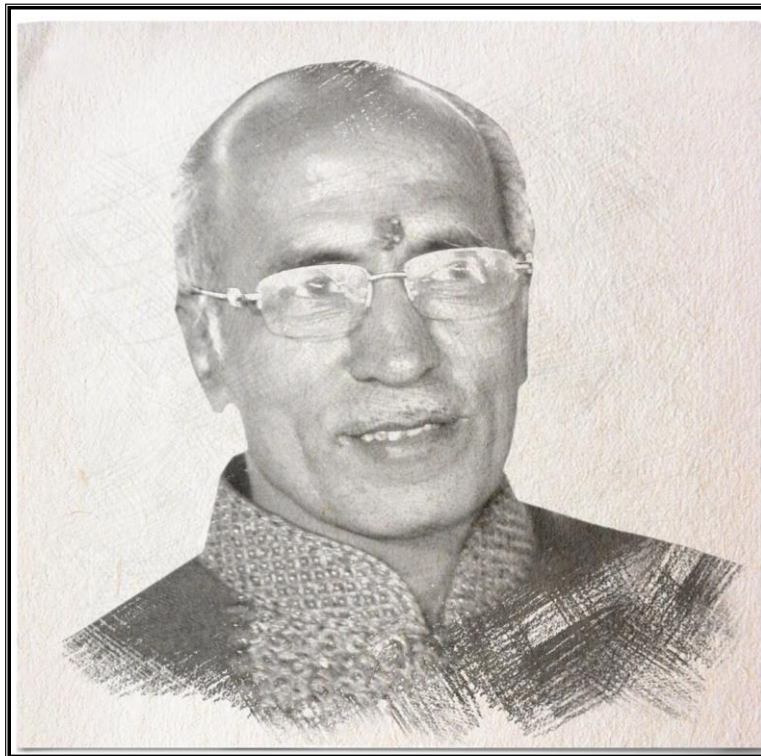
  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



[ITC602] (FID : ME69) (Book Code : ME69A)

(Book Code : ME69A)

*We dedicate this Publication soulfully and  
wholeheartedly,  
in loving memory of our beloved founder director,  
**Late Shri. Pradeepji Lalchandji Lunawat,**  
who will always be an inspiration, a positive force and  
strong support behind us.*



*“My work is my prayer to God”*

*- Lt. Shri. Pradeepji L. Lunawat*

*Soulful Tribute and Gratitude for all Your  
Sacrifices, Hardwork, and 40 years of  
Strong Vision...*





**Dr. G. T. Thampi**  
**PRINCIPAL**  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



## **Preface**




Dear students,

We are extremely happy to present the book on **“Data Mining and Business Intelligence”** for you. We have divided the subject into small chapters so that the topics can be arranged and understood properly. The topics within the chapters have been arranged in a proper sequence to ensure smooth flow of the subject.

We present this book in the loving memory of **Late. Shri. Pradeepji Lunawat**, our source of inspiration and a strong foundation of **“TechKnowledge Publications”**. He will always be remembered in our hearts and motivate us to achieve our new milestone.

We are thankful to Mr. Arunoday Kumar, Mr. Shital Bhandari and Mr. Chandroday Kumar for the encouragement and support that they have extended. We also thankful to the staff members of TechKnowledge Publications for their efforts to make this book as good as it is. We have made every possible efforts to eliminate all the errors in this book. However if you find any, please let us know, because that will help us to improve the book quality further.

We are also thankful to our family members and friends for their patience and encouragement.

 -  **Authors**  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050. 

(Book Code : ME69A)

## Books by Author – Dr Arun Kulkarni



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.

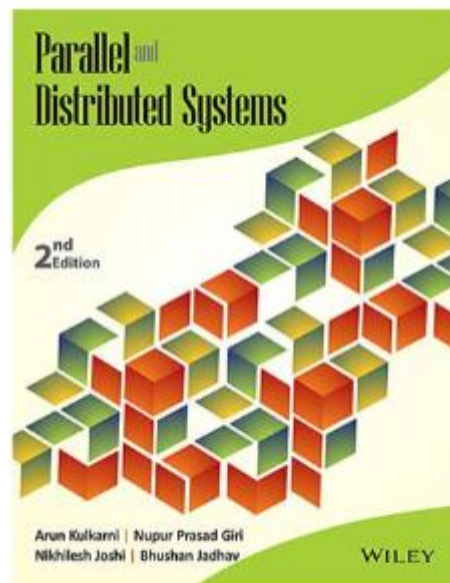




# Parallel and Distributed Systems

by [Bhushan Jadhav](#) [Arun Kulkarni](#), [Nupur Prasad Giri](#), [Nikhilesh Joshi](#) (Author)

ISBN: 978-8126558674



Preface

Acknowledgement

About the Authors

Chapter 1 Introduction to Parallel Computing

1.1 Introduction


1.2 Computing


1.3 Parallel Architecture

1.4 Classification Based on Architectural Schemes

1.5 Classification Based on Memory Access

1.6 Classification Based on Interconnections between PEs and Memory Modules

  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



## 1.7 Classification Based on Characteristic Nature of Processing Elements

### 1.8 Performance Metrics

### 1.9 Parallel Programming Models

### 1.10 Serial and Parallel Algorithms

### 1.11 Parallelism

## Chapter 2 Pipelining

### 2.1 Introduction

### 2.2 Pipeline Performance

### 2.3 Types of Pipeline

### 2.4 Pipeline Stage Design

### 2.5 Pipeline Hazards

### 2.6 Instruction Scheduling

## Chapter 3 Synchronous Parallel Processing

### 3.1 Introduction

### 3.2 SIMD Architecture and Its Programming Principle

### 3.3 Single Instruction Multiple Data (SIMD) Parallel Algorithms

### 3.4 Data Mapping and Memory in Array Processor

### 3.5 Case Studies of SIMD Parallel Processors

## Chapter 4 Introduction to Distributed Systems

### 4.1 Introduction

### 4.2 Definition

### 4.3 Goals of the Distributed System

### 4.4 Issues Related to the Distributed System

### 4.5 Types of Distributed System


- 4.6 Distributed System Models
- 4.7 Hardware Concept
- 4.8 Software Concept
- 4.9 Models of Middleware
- 4.10 Services Offered by Middleware System
- 4.11 Client–Server Model

## Chapter 5 Communication

- 5.1 Introduction
- 5.2 Layered Protocols
- 5.3 Remote Procedure Call
- 5.4 Remote Object Invocation
- 5.5 Remote Method Invocation
- 5.6 Message-Oriented Communication
- 5.7 Stream-Oriented Communication

## Chapter 6 Resource and Process Management

- 6.1 Resource Management in Distributed System
- 6.2 Desirable Features of Global Scheduling Algorithm
- 6.3 Scheduling in the Distributed System
- 6.4 Taxonomy of the Distributed Scheduling
- 6.5 Task Assignment Approach
- 6.6 Load Balancing Approach
- 6.7 Issues in Designing Load Balancing Algorithm
- 6.8 Load Sharing Approach
- 6.9 Introduction to Process Management

  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



## Chapter 7 Synchronization

### 7.1 Introduction: Clock Synchronization

### 7.2 Physical Clock

### 7.3 Logical Clock

### 7.4 Election Algorithms

### 7.5 Mutual Exclusion

### 7.6 Centralized Algorithm

### 7.7 Distributed Mutual Exclusion

## Chapter 8 Replication, Consistency and Distributed File System

### 8.1 Introduction

### 8.2 Replication and Consistency

### 8.3 Replication Management

### 8.4 Distributed File Systems

### 8.5 Case Studies

## Summary

## Multiple Choice Questions

## Short Answer Questions

## Long Answer Questions

## Answers

### Practical No. 1

### Practical No. 2

### Practical No. 3

### Practical No. 4

### Practical No. 5

Practical No. 6

Practical No. 7

Practical No. 8

Practical No. 9

Practical No. 10

Practical No. 11

Practical No. 12

Index



**Dr. G. T. Thampi**  
**PRINCIPAL**  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



## Books by Author – Dr Bhushan Jadhav



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.

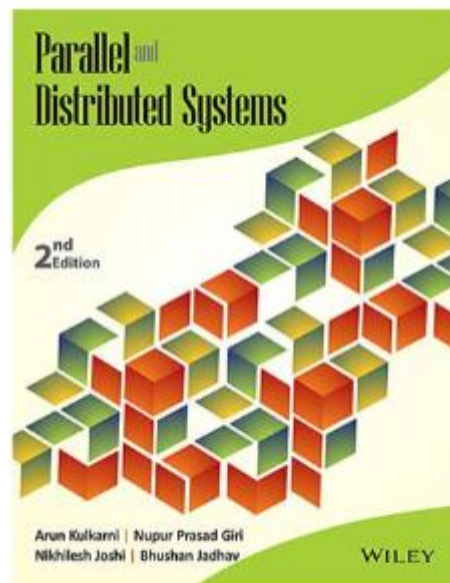




# Parallel and Distributed Systems

by [Bhushan Jadhav](#) [Arun Kulkarni](#), [Nupur Prasad Giri](#), [Nikhilesh Joshi](#) (Author)

ISBN: 978-8126558674



Preface

Acknowledgement

About the Authors

Chapter 1 Introduction to Parallel Computing

1.1 Introduction


1.2 Computing


1.3 Parallel Architecture

1.4 Classification Based on Architectural Schemes

1.5 Classification Based on Memory Access

1.6 Classification Based on Interconnections between PEs and Memory Modules

  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



## 1.7 Classification Based on Characteristic Nature of Processing Elements

### 1.8 Performance Metrics

### 1.9 Parallel Programming Models

### 1.10 Serial and Parallel Algorithms

### 1.11 Parallelism

## Chapter 2 Pipelining

### 2.1 Introduction

### 2.2 Pipeline Performance

### 2.3 Types of Pipeline

### 2.4 Pipeline Stage Design

### 2.5 Pipeline Hazards

### 2.6 Instruction Scheduling

## Chapter 3 Synchronous Parallel Processing

### 3.1 Introduction

### 3.2 SIMD Architecture and Its Programming Principle

### 3.3 Single Instruction Multiple Data (SIMD) Parallel Algorithms

### 3.4 Data Mapping and Memory in Array Processor

### 3.5 Case Studies of SIMD Parallel Processors

## Chapter 4 Introduction to Distributed Systems

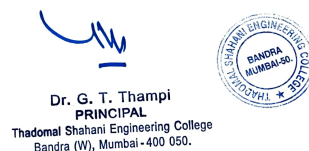
### 4.1 Introduction

### 4.2 Definition

### 4.3 Goals of the Distributed System

### 4.4 Issues Related to the Distributed System

### 4.5 Types of Distributed System




- 4.6 Distributed System Models
- 4.7 Hardware Concept
- 4.8 Software Concept
- 4.9 Models of Middleware
- 4.10 Services Offered by Middleware System
- 4.11 Client–Server Model

## Chapter 5 Communication

- 5.1 Introduction
- 5.2 Layered Protocols
- 5.3 Remote Procedure Call
- 5.4 Remote Object Invocation
- 5.5 Remote Method Invocation
- 5.6 Message-Oriented Communication
- 5.7 Stream-Oriented Communication

## Chapter 6 Resource and Process Management

- 6.1 Resource Management in Distributed System
- 6.2 Desirable Features of Global Scheduling Algorithm
- 6.3 Scheduling in the Distributed System
- 6.4 Taxonomy of the Distributed Scheduling
- 6.5 Task Assignment Approach
- 6.6 Load Balancing Approach
- 6.7 Issues in Designing Load Balancing Algorithm
- 6.8 Load Sharing Approach
- 6.9 Introduction to Process Management

  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



## Chapter 7 Synchronization

### 7.1 Introduction: Clock Synchronization

### 7.2 Physical Clock

### 7.3 Logical Clock

### 7.4 Election Algorithms

### 7.5 Mutual Exclusion

### 7.6 Centralized Algorithm

### 7.7 Distributed Mutual Exclusion

## Chapter 8 Replication, Consistency and Distributed File System

### 8.1 Introduction

### 8.2 Replication and Consistency

### 8.3 Replication Management

### 8.4 Distributed File Systems

### 8.5 Case Studies

## Summary

## Multiple Choice Questions

## Short Answer Questions

## Long Answer Questions

## Answers

### Practical No. 1

### Practical No. 2

### Practical No. 3

### Practical No. 4

### Practical No. 5



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.

Practical No. 6

Practical No. 7

Practical No. 8

Practical No. 9

Practical No. 10

Practical No. 11

Practical No. 12

Index



**Dr. G. T. Thampi**  
**PRINCIPAL**  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.





# DISTRIBUTED COMPUTING

Includes Labs



Nupur Giri  
Lata Ragha  
Bhushan Jadhav

Dr. G. T. Thampi  
PRINCIPAL

Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



 **STAREDU  
SOLUTIONS**

[www.staredusolutions.org](http://www.staredusolutions.org)



# DISTRIBUTED COMPUTING

Nupur Prasad Giri

Lata Ragha

Bhushan Jadhav



Dr. G. T. Thampi  
PRINCIPAL

Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



# Author Profile



**Dr. Nupur Prasad Giri** is Professor and Head of Department, Computer Engineering at Vivekanand Education Society Institute of Technology, Mumbai. She has more than 25 years of industry and teaching experiences. Her doctoral research was in the multidisciplinary field of Multi Agent System and Mobile Computing, typically Cellular Networks. Her contributions in the field of Artificial Intelligence, Mobile and Distributed Computing have been published in many journals and International conferences. She has filed many patents.

She is also a committee member of many international conferences. She is recognized PhD guide in University of Mumbai. She has represented India as an Expert in Worldskills-2011, London as well as in Worldskills-2013, Germany in the category of Web Design. She has also been awarded Microsoft's "AI for Earth" Grant in the year 2018.



**Dr. Lata L. Ragha** is Professor and Head, Department of Computer Engineering at Fr. C. Rodrigues Institute of Technology, Vashi, Navi-Mumbai. She has received her Ph. D. degree from Jadavpur University, Kolkata in 2011. Her research interests include in the areas of Networking, Security, Internet Routing, and Data Mining. She has more than 100 research publications in International Journals and conferences. She is having 32 years of teaching experience. She is a member of Board of Studies, Mumbai University. She is associated with two colleges as member of Department Advisory Board. Two members have received their PhD degree under her guidance and two more have submitted their thesis under her guidance.

Dr. G. T. Thampi  
PRINCIPAL

Thadomal Shahani Engineering College  
Bandra (W), Mumbai-400 050.



**Dr. Bhushan A Jadhav** is an Assistant Professor at Thadomal Sahani Engineering College, Bandra, Mumbai. He has more than 11 years of teaching experience and has completed his Ph.D in area of Cloud Computing & Big Data Analytics. His area of interest and research includes cutting edge technologies such as DevOps, Internet of Things (IOT), Cloud Computing, Big Data Analytics, Python Programming, R Programming and Advance Information Security. He has published more than 8 research papers in National and International journals and conferences He has published six books with various publishers. He is certified trainer for Star Certification's Python Programming, Cloud Computing and DevOps Courses.



# Contents

<b>1. INTRODUCTION TO DISTRIBUTED SYSTEMS</b>	<b>1</b>
1.1. Introduction	2
1.2. Definition of Distributed Systems	3
1.3. Significant Consequences of Distributed Systems	5
1.4. Goals of Distributed Systems	5
1.5. Issues of Distributed Systems	10
1.6. Types of Distributed Systems	13
1.6.1. Distributed Computing system	13
1.6.2. Distributed information systems	17
1.6.3. Distributed pervasive system:	20
1.7. Distributed System Models	23
1.7.1. Architectural Model:	23
1.7.2. Interaction Model:	25
1.7.3. Fault Models	27
1.7.4. Security Model	29
1.8. Hardware Concept	30
1.8.1. Multiprocessor System	30
1.8.2. Multicomputer System	32

1.9. Software Concept . . . . .	33
1.9.1. Distributed Operating System . . . . .	33
1.9.2. Network Operating System . . . . .	34
1.9.3. Middleware . . . . .	35
1.10. Services offered by Middleware . . . . .	36
1.11. Models of Middleware. . . . .	36
1.11.1. Remote Procedure Call (RPC) . . . . .	36
1.11.2. Message Oriented Middleware (MOM). . . . .	37
1.11.3. Distributed Object Technology . . . . .	37
1.12. Client-Server Model . . . . .	38
1.12.1. Client . . . . .	38
1.12.2. Server. . . . .	39
1.12.3. Network . . . . .	39
Summary . . . . .	40
Review Questions . . . . .	41
<b>2. COMMUNICATION . . . . .</b>	<b>42</b>
2.1. Introduction. . . . .	42
2.2. Layered Protocols . . . . .	43
2.2.1. Seven Layers of OSI . . . . .	49
2.3. Remote Procedure call. . . . .	50
2.3.1. The RPC Model . . . . .	51
2.3.2. Implementing RPC Mechanism . . . . .	53
2.3.3. Parameter Passing in RPC . . . . .	53
2.3.4. Extended RPC Models. . . . .	56
2.3.5. Example of RPC. . . . .	56
2.4. Remote Object Invocation . . . . .	57
2.4.1. Types of Object and Binding . . . . .	58
2.5. Remote Method Invocation . . . . .	58
2.5.1. Architecture of Remote Method Invocation . . . . .	60
2.5.2. Remote Method Invocation Process . . . . .	62
2.6. Message-Oriented Communication . . . . .	



2.6.1. Types of Communication .....	63
2.6.2. Message-Oriented Transient Communication .....	65
2.6.3. Message-Oriented Persistent Communication .....	69
2.7. Group Communication .....	71
2.7.1. One – to – many communication .....	71
2.7.2. Many – to – one Communication .....	73
2.7.3. Many –to – many .....	74
2.8. Stream-Oriented Communication .....	77
Summary .....	79
Review Questions .....	80
<b>3. SYNCHRONIZATION .....</b>	<b>81</b>
3.1 Introduction: Clock Synchronization .....	82
3.2 Physical clock .....	83
3.2.1 Christian's Algorithm .....	83
3.2.2 Berkeley Algorithm .....	85
3.2.3 Network Time Protocol .....	87
3.3 Logical clock .....	88
3.1.1 Lamport's Scalar Clock .....	89
3.3.2 Vector Timestamp Ordering .....	91
3.4 Election algorithms .....	93
3.4.1 Bully Algorithm .....	93
3.4.2 Ring Algorithm .....	97
3.5 Mutual exclusion .....	99
3.6 Centralized algorithm .....	101
3.7 Distributed Mutual exclusion .....	104
3.7.1 Non-Token Based Algorithms .....	104
3.7.2 Token-Based Algorithms .....	113
3.7.3 Singhal's Heuristic Algorithm .....	118
3.7.4 Raymond's Tree-Based Algorithm .....	121
Summary .....	127
Review Questions .....	128

Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai-400 050.





<b>4. RESOURCE AND PROCESS MANAGEMENT</b>	<b>129</b>
4.1 Resource Management in Distributed system.	130
4.2 Desirable Features of global scheduling algorithm	130
4.3 Scheduling in the Distributed system	132
4.4 Taxonomy of the Distributed scheduling	134
4.4.1 Hierarchical Classification.	134
4.4.2 Dynamic Scheduling Algorithm	137
4.5 Task assignment approach	139
4.6 Load balancing approach	142
4.6.1 Benefits of Load Balancing	142
4.6.2 Static Load Balancing	143
4.6.3 Dynamic Load Balancing.	145
4.6.4 Comparison Between Static and Dynamic Load Balancing Algorithms	146
4.6.5 Static versus Dynamic Algorithms.	147
4.6.6 Deterministic versus Probabilistic Algorithms	147
4.6.7 Centralized versus Distributed Algorithms.	148
4.7 Issues in Designing Load balancing algorithm	148
4.7.1 Load Estimation Policy	148
4.7.2 Process Transfer Policy.	149
4.7.3 Location Policies	151
4.7.4 State Information Exchange Policy	152
4.7.5 Priority Assignment Policy.	153
4.8 Load sharing approach	153
4.8.1 Load Estimation Policy	153
4.9 Process Management	154
4.9.1 Threads	155
4.10 Virtualization	158
4.11 Clients	161
4.11.1 Thin-Client Network Computing.	161
4.11.2 Client-Side Software for Distribution Transparency	163
4.12 Servers	164



4.12.1 General Design Issues . . . . .	164
4.12.2 Cluster of Servers . . . . .	166
4.12.3 Distributed Servers . . . . .	167
4.13 Code Migration . . . . .	170
4.13.1 Approaches to Code Migration . . . . .	170
4.13.2 Models for Code Migration . . . . .	171
4.13.3 Migration and Local Resources. . . . .	172
4.13.4 Migration in Heterogeneous Systems . . . . .	175
Summary . . . . .	177
Review Questions . . . . .	177
<b>5. CONSISTENCY, REPLICATION AND FAULT TOLERANCE . . . . .</b>	<b>179</b>
5.1. Introduction . . . . .	180
5.2. Consistency . . . . .	180
5.2.1 Consistency Models and their Evolution . . . . .	181
<b>5.2.1 Data-Centric Consistency Models . . . . .</b>	<b>183</b>
5.2.2 Client-Centric Consistency Models. . . . .	191
5.3. Replication . . . . .	197
5.3.1 Replica Location . . . . .	197
5.3.2 Replication Models . . . . .	198
5.3.3 Replica Consistency. . . . .	198
5.4. Fault Tolerance. . . . .	199
5.4.1 Failure Models. . . . .	200
5.4.2 Failure Detection . . . . .	202
5.4.3 Failure Masking . . . . .	203
5.4.4 Resilience by Process Groups: . . . . .	204
5.4.5 Reliable Communication: Client - Server . . . . .	212
5.4.6 Reliable Group Communication: . . . . .	215
5.4.7 Atomic Multicast . . . . .	219
5.4.8 Recovery . . . . .	223
Summary . . . . .	226
Review Questions . . . . .	226

Dr. G. T. Thampi  
**PRINCIPAL**  
 Thadomal Shahani Engineering College  
 Bandra (W), Mumbai - 400 050.





<b>6. DISTRIBUTED FILE SYSTEMS AND NAME SERVICES</b>	<b>227</b>
6.1. Introduction to Distributed File systems	228
6.1.1 Features of a Good DFS	228
6.1.2 File Services	229
6.1.3 Naming	230
6.1.4 Naming Schemes	231
6.1.5 File Sharing Semantics	231
6.1.6 Caching	233
6.1.7 File Replication	235
6.2. Case Study on Distributed File Systems	236
6.2.1 NFS	236
6.2.2 Andrew File System	243
6.3. Introduction to Name Services and Domain Name System	247
6.3.1 Name Service	248
6.3.2 The Domain Name System	251
6.4. Directory services	253
6.5. Case study: The Global Name Service	254
6.6. X.500 Directory Service	257
6.7. Google Case Study	259
6.7.1 Components of Google Infrastructure	260
Summary	276
Review Questions	276
<b>LAB</b>	<b>277</b>
<b>INDEX</b>	<b>340</b>

Dr. G. T. Thamp  
PRINCIPAL

Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



# Information Management

SEM VII

IT

AU

# NoSQL

# hadoop



**TECHNICAL<sup>™</sup>**  
**PUBLICATIONS**

*An Up-Thrust for Knowledge*

**Asha Bharambe**  
**Bhushan Jadhav**  
**Anjali Yeole**

**WILEY**



# Information Management

**Asha Bharambe**

Assistant Professor

Vivekanand Education Society's Institute of Technology  
Chembur (E), Mumbai

**Bhushan Jadhav**

Assistant Professor

Thadomal Shahani Engineering College  
Bandra (W), Mumbai

**Anjali Yeole**

Assistant Professor

Vivekanand Education Society's Institute of Technology  
Chembur (E), Mumbai



Dr. G. T. Thampi  
PRINCIPAL

Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



**WILEY**

*This book is dedicated to my parents Mr. Yadeo Barhate and Mrs. Usha Barhate,  
my in-laws Mr. Laxman Bharambe and Mrs. Mangala Bharambe,  
my husband Mr. Aniket Bharambe and my son Atharva.*

*—Asha Bharambe*

*This book is dedicated to my parents Mr. Ashok Jadhav and Mrs. Shanta Jadhav,  
my wife Sonali Jadhav and my son Atharva.*

*—Bhushan Jadhav*

*This book is dedicated to my parents Mr. Devidas Pawar and Mrs. Alka Pawar,  
my in-laws Mr. Madhukar Yeole and Late Kusum Yeole,  
my husband Mr. Shrikant Yeole and my daughters Shrutika and Sanvee.*

*—Anjali Yeole*



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.



# About the Authors



**Asha Bharambe** is currently working as Assistant Professor at Vivekanand Education Society's Institute of Technology, Chembur (E), Mumbai. She is also pursuing PhD in Information Technology from Mumbai University. She has received her undergraduate degree in Computer Science from SNDT Women's University and postgraduate in Computer Engineering from Mumbai University. Prof. Bharambe has taught both graduate and undergraduate students and has a teaching experience of more than 15 years. She has published many research papers in national and international conferences on

Data Mining, Big Data and Artificial Intelligence. Her areas of interest are Data Mining, Big Data and Natural Language Processing.



**Bhushan Jadhav** is currently working as Assistant Professor at Thadomal Shahani Engineering College, Bandra (W), Mumbai. He is also pursuing PhD in Computer Science Engineering from Thadomal Shahani Engineering College, Mumbai University. He received his undergraduate and postgraduate education from Mumbai University. Prof. Jadhav has taught both postgraduate and undergraduate students and has teaching experience of more than 8 years. He has published many research papers in national and international conferences on Virtualization and Cloud Computing. His book *Parallel and Distributed*

*Systems* has received very good response at Mumbai University. His areas of interest are Database Management, Parallel Computing, Distributed Computing, Virtualization and Cloud Computing.



**Anjali Yeole** is currently working as Assistant Professor at Vivekanand Education Society's Institute of Technology, Chembur (E), Mumbai. She is also pursuing PhD in Computer Science Engineering from Thadomal Shahani Engineering College, Mumbai University. She received her undergraduate in Computer Science from SNDT University and postgraduate education in Computer Engineering from Veermata Jijabai Technological Institute, Mumbai University. Prof. Yeole has taught both postgraduate and undergraduate students and has teaching experience of more than 12 years. She has published

many research papers in national and international conferences on Security, Cloud Computing and Web Applications. Her areas of interest are Parallel Computing, Distributed Computing, Virtualization and Cloud Computing, System Security, Computer Network, and Web Technology.

# Contents

<b>Preface</b>	<b>vii</b>
<b>About the Authors</b>	<b>ix</b>
<b>Syllabus</b>	<b>xi</b>
<b>Chapter 1 Database Design and Modelling</b>	<b>1</b>
Learning Objectives	1
1.1 Introduction	1
1.2 Database Design	1
1.3 Database Modelling	2
1.3.1 ER Model	3
1.3.2 ER to Relational Data Model	6
1.3.3 Normalisation	14
1.3.4 Extended Entity Relationship Diagram	16
1.3.5 Case Study: Library Management System	17
1.4 Business Rules	19
1.4.1 Identifying Business Rules	20
1.5 Java Database Connectivity (JDBC)	20
1.5.1 Types of JDBC Drivers	21
1.6 Accessing Database Using JDBC	23
1.6.1 Registering JDBC Driver	23
1.6.2 Creating Database Connection	23
1.6.3 Executing Queries	24
1.6.4 Processing the Results	27
1.6.5 Closing the Database Connection	27
1.7 Stored Procedures	29
Summary	30
Multiple Choice Questions	31
Solved Questions	31

  
 Dr. G. T. Thampi  
 PRINCIPAL  
 Thadomal Shahani Engineering College  
 Bandra (W), Mumbai - 400 050.





## Chapter 2 Introduction to Big Data

### Learning Objectives

- 2.1 Introduction
  - 2.2 Need for Big Data
  - 2.3 Characteristics of Big Data
    - 2.3.1 *High-Level Hadoop Architecture*
    - 2.3.2 *Characteristics of Hadoop*
  - 2.4 Components and Ecosystem
  - 2.5 Running Hadoop
  - 2.6 HDFS
    - 2.6.1 *Preparing HDFS Writes*
    - 2.6.2 *Reading Data from HDFS*
  - 2.7 MapReduce
    - 2.7.1 *MapReduce Framework*
    - 2.7.2 *MapReduce Flow*
    - 2.7.3 *Execution of a Job*
    - 2.7.4 *MapReduce Program Components*
  - 2.8 YARN
    - 2.8.1 *YARN MapReduce Application Execution Flow*
    - 2.8.2 *Advantages of YARN*
  - 2.9 NoSQL
    - 2.9.1 *Types of NoSQL Databases*
  - 2.10 Hive
    - 2.10.1 *Hive Architecture*
    - 2.10.2 *Hive Applications*
    - 2.10.3 *Difference with RDBMS*
- Summary  
Multiple Choice Questions  
Solved Questions  
Review Questions  
Answers to Multiple Choice Questions<sup>92</sup>

  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.





## Chapter 3 Data Security and Privacy 63

Learning Objectives	63
3.1 Introduction	63
3.2 Program Security	64
3.2.1 Fixing Faults	64
3.2.2 Unexpected Behavior	65
3.2.3 Types of Flaws	65
3.3 Overview: Covert Channels – Programs that Leak Information	77
3.3.1 Storage Channels	79
3.3.2 Timing Channels	79
3.3.3 Shared Resource Matrix	79
3.3.4 Information Flow Method	79
3.3.5 Controls Against Program Threats	80
3.3.6 Good Design	81
3.4 Operating System	81
3.4.1 History of Protection in Operating Systems	82
3.4.2 Protected Objects	82
3.4.3 Security Methods of Operating Systems	83
3.5 Firewalls	92
3.5.1 Design Goals of Firewalls	92
3.5.2 Techniques Used by Firewalls for Access Control and Enforcement of Security Policy	93
3.5.3 Capabilities of a Firewall	93
3.5.4 Limitations of Firewalls	93
3.5.5 Types of Firewalls	93
3.5.6 Example Firewall Configurations	96
3.6 Network Security	98
3.6.1 Introduction to Network	98
3.6.2 TCP/IP Vulnerability	99
3.6.3 Security Measures	104
3.6.4 Different Way of Attacking on Network	105
3.6.5 Protocols for Security	110
3.7 Intrusion Detection System (IDS)	113
3.7.1 Goals of an Ideal IDS	114
3.7.2 Strengths and Weaknesses of IDS	114

3.7.3	<i>Based on IDS Operation</i>	115
3.7.4	<i>Based on Where It is Placed</i>	116
3.7.5	<i>Stealth Mode</i>	118
3.8	Data Privacy Principles	119
3.8.1	<i>Types of Privacy</i>	120
3.8.2	<i>Data Privacy Principles</i>	121
3.9	Data Privacy Laws and Compliance	122
3.9.1	<i>Data Protection Laws</i>	122
3.9.2	<i>Compliance</i>	123
	Summary	124
	Multiple Choice Questions	124
	Solved Questions	126
	Review Questions	128
	Answers to Multiple Choice Questions	130

## Chapter 4 Information Governance

131

### Learning Objectives

4.1	Introduction	131
4.1.1	<i>Need for Master Data Management</i>	131
4.1.2	<i>Definition of Master Data Management</i>	132
4.1.3	<i>Characteristics and Benefits of Master Data Management</i>	133
4.1.4	<i>Master Data Management Versus Data Warehouse</i>	133
4.2	Stages of Master Data Management Implementation	134
4.3	Master Data Management Architectural Dimensions	135
4.3.1	<i>Design and Deployment Dimension</i>	137
4.3.2	<i>Use Pattern Dimension</i>	138
4.3.3	<i>Information Scope or Data Domain Dimension</i>	139
4.3.4	<i>Master Data Management Reference Architecture</i>	140
4.4	Risk Management	141
4.4.1	<i>Privacy, Regulatory Requirements and Compliance</i>	143
4.4.2	<i>Implications of Data Security and Privacy Regulations on Master Data Management</i>	144
4.5	Data Governance	150
4.5.1	<i>Goals of Data Governance</i>	150
4.6	Data Synchronization	151
4.7	Data Quality Management	152



Summary	154
Multiple Choice Questions	155
Solved Questions	156
Review Questions	166
Answers to Multiple Choice Questions	166

## **Chapter 5 Information Architecture** **167**

Learning Objectives	167
5.1 Introduction	167
5.2 Principles of Information Architecture and Framework	167
5.2.1 <i>Definitions of Information Architecture</i>	168
5.2.2 <i>Components of Information Architecture</i>	170
5.3 Organization System	171
5.3.1 <i>Classification of Organization System</i>	171
5.3.2 <i>Organization Schemes</i>	172
5.3.3 <i>Organization Structure</i>	176
5.4 Navigation Systems	180
5.4.1 <i>Types of Navigation Systems</i>	181
5.5 Labeling Systems	185
5.5.1 <i>Sources of Labeling Systems</i>	188
5.6 Searching Systems	189
5.7 Conceptual Design	189
5.8 Granularity of Contents	191
Summary	191
Multiple Choice Questions	193
Solved Questions	194
Review Questions	203
Answers to Multiple Choice Questions	204

## **Chapter 6 Information Lifecycle Management** **205**

Learning Objectives	205
6.1 Introduction	205
6.2 Data Retention Policies	205
6.2.1 <i>Purpose</i>	206

6.2.2	<i>Scope</i>	24
6.2.3	<i>Policy Contents</i>	24
6.2.4	<i>Managing the Data Retention Policy</i>	24
6.3	<i>Data Retention in Telecommunication Industry</i>	2
6.3.1	<i>Internet Service Provider (ISP) License</i>	2
6.3.2	<i>Unified Access Service License (UASL)</i>	2
6.3.3	<i>Sample Retention Records</i>	2
6.3.4	<i>Laws Related to Data Retention Policy in India</i>	2
6.4	<i>Confidential and Sensitive Data Handling</i>	2
6.4.1	<i>Handling of Sensitive Data</i>	2
6.4.2	<i>Access Decisions</i>	2
6.4.3	<i>Types of Disclosures</i>	2
6.4.4	<i>Handling Data</i>	2
6.4.5	<i>Law Provision in India Defining Sensitive Data and Its Handling</i>	2
6.5	<i>Lifecycle Management Costs</i>	2
6.6	<i>Archive Data Using Hadoop</i>	2
6.7	<i>Testing and Delivering Big Data Applications for Performance and Functionality</i>	2
6.8	<i>Challenges with Data Administration</i>	2
	<i>Summary</i>	
	<i>Multiple Choice Questions</i>	
	<i>Solved Questions</i>	
	<i>Review Questions</i>	
	<i>Answers to Multiple Choice Questions</i>	



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai-400 050.



# INTERNET OF EVERYTHING

Includes Labs and Cases



## INTERNET

# INTERNET OF EVERYTHING

---

Bhushan Jadhav

---

---

Anjali Yeole

---

---

Gopal Pardesi

---

---

Vaishali Khairnar

---


---

Dhananjay Kalbande


---








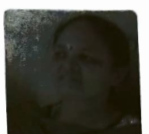
**Dr. Bhushan Jadhav** is an Assistant Professor at Thadomal Shahani Engineering College, Bandra, Mumbai. He has more than 11 years of teaching experience and has completed his Ph.D in area of Cloud Computing & Big Data Analytics. His area of interest and research includes cutting edge technologies such as DevOps, Internet of Things (IoT), Cloud Computing, Big Data Analytics, Python Programming, R Programming and Advance Information Security. He has published more than 8 research papers in National and International journals and conferences He has published six books with various publishers. He is certified trainer for Star Certification's Python Programming, Cloud Computing and DevOps Courses.




**Dr. Anjali Yeole** is currently working as Assistant Professor at Vivekanand Education Society's Institute of Technology, Chembur (E), Mumbai. She has received her PhD in Computer Science Engineering from Thadomal Shahani Engineering College under Mumbai University, India. Her research area is "Internet of Things". She received her undergraduate in Computer science from SNDT university and postgraduate education in Computer engineering from VJTI, Mumbai University. She has taught both Masters and undergraduate students and has teaching experience of more than 17 years. She has published many research papers in national and international conferences and journals on IoT, security, cloud computing and web applications. She has authored a book "Information management" Her areas of interest are IoT, Parallel Computing, Distributed Computing, Virtualization and Cloud Computing, System Security, Computer Network and Web Technology.



**Dr. Gopal Pardesi** is currently working as Associate Professor at Thadomal Shahani Engineering College, Bandra, Mumbai. He has more than 28 years of teaching experience. He has completed his Ph.D in the field of Information and Communication Technologies. His areas of interest are Wireless Networks, Project Management, Microprocessors and Microcontrollers, Embedded Systems and Machine learning.



**Dr. Vaishali Khairnar** is currently associated as Head of Department Information Technology Department at Terna College of Engineering, Nerul, Navi Mumbai. She has done her Doctoral work from Nirma University, Ahmedabad, Gujrat on Vehicular Ad-Hoc networks. She has more than 19 years of collective experience in industry, teaching as well as research. She has more than 50 papers published in national and international journals and conferences to her credit. Her area of interest includes wireless networks, IOT, Open source tools, Ad-Hoc Sensor networks, Internet Programming, Mobile Application and Development & Storage Network Management and Retrieval. She has authored four books published in India, UK and Singapore. She has been a resource person for several workshops and short term training programmes. She is paper reviewer on different editor bodies at national/ international journals and conferences. She is active member of CSI- Mumbai Chapter and ACM.



**Dr. Dhananjay Kalbande** is currently a Professor in Computer Engineering and Dean (Industry Relations), Sardar Patel Institute of Technology, Andheri (West), Mumbai, India. He was Head of the Department from April 2012 to Oct 2019. He completed his B.E. in Computer Technology from Nagpur University in 1997 and Master of Engineering in Information Technology in May 2005, from Vivekanand Education Society's Institute of Technology (VESIT), Mumbai University, Mumbai, India. He has obtained a Ph.D in Technology from University of Mumbai, Mumbai in 2011. He has been awarded a Post-Doctorate (PDF) from Tata Institute of Social Sciences (T.I.S.S.) in 2016. He has also been honoured as a Senior Research Fellow (SRF) on the NCW-TISS Project, funded by National Commission for Women, Govt. of India at T.I.S.S., from July 23, 2016 to Oct 10, 2016. He was a Research Fellow on the CliX Project at T.I.S.S., funded by Tata Trust and M.I.T.(U.S.A.) from Feb 20, 2017 to May 19, 2017. He has over 19+ Years of experience in teaching & research. His research interests include Soft computing (Neural Networks, Fuzzy Logic) Computer Network, Human Machine Interaction Decision making and business Intelligence. Mobile application development for social cause. ICT for semi-rural development for social cause. He has authored four books namely Graphical User Interface (Pareen Publications), MIS (Pareen Publications), Human Machine Interaction and Digital Forensic with Wiley India Pvt. Ltd. He has delivered and conducted workshops, seminars, Tutorials and Expert Talks on NS2 Neural Network, VB.Net and ADO.Net, Transfer Learning. Skinzy is his brainchild which has been turned into reality into Healthcare Start-up operated from Mumbai. Skinzy's flagship product "DermaPhoto" is an AI based mobile application that can detect skin diseases. He has patented 6 innovative ideas of research work.



# Contents

<b>1. INTRODUCTION</b>	<b>1</b>
1.1. Introduction to IoT and IoE	2
1.2. History of IoE	4
1.2.1. Telemetry (1845 to 1990s)	5
1.2.2. M2M systems (1990s to 2010)	5
1.2.3. Internet of Things (2010 to present)	5
1.2.4. Internet of Everything (IoE) 2014 till present	6
1.2.5. Difference between IoT and M2M	6
1.3. Objects in IoE	8
1.3.1. Sensors	10
1.3.2. Data	13
1.3.3. Actuator	13
1.3.4. People/ Process	13
1.4. Identifiers in the IoE	13
1.4.1. Thing Identifier	14
1.4.2. Communication Identifier	16
1.4.3. User Identifier	17

1.5. IoE-enabling Technologies .....	17
1.5.1. Sensors .....	17
1.5.2. Wireless Sensor Network .....	18
1.5.3. Cloud Computing .....	19
1.5.4. Big Data Analytics .....	20
1.5.5. Communication Protocols .....	21
1.5.6. Messaging Protocol .....	26
1.5.7. Embedded Computing Boards .....	28
1.5.8. Web services and REST-based web services .....	30
Summary .....	30
Important Questions .....	31
<b>2. RFID TECHNOLOGIES .....</b>	<b>33</b>
2.1. Introduction .....	34
2.2. RFID .....	34
2.3. IOT and RFID Technology .....	35
2.4. RFID System .....	37
2.5. Components of RFID System .....	38
2.5.1. RFID Antenna .....	38
2.5.2. RFID Reader .....	40
2.5.3. RFID Transponder .....	42
2.6. RFID Tag .....	43
2.7. Principle of RFID .....	46
2.8. RFID Middleware .....	47
2.9. Issues .....	49
Summary .....	51
Important Questions .....	52
<b>3. RFID APPLICATIONS .....</b>	<b>53</b>
3.1. Introduction .....	54
3.2. Applications of RFID System .....	54
3.3. Challenges of RFID Technology .....	60
3.4. Hardware Issues and Protocols .....	61

Dr. G. T. Thampi  
PRINCIPAL

Thadomal Shahani Engineering College  
• Bandra (W), Mumbai - 400 050.



3.5. RFID Anti-Collision Protocols for Tag Identification . . . . .	62
3.6. Pure Aloha . . . . .	64
3.7. Slotted Aloha . . . . .	65
3.8. Framed Slotted Aloha and Dynamic Framed Slotted Aloha . . . . .	66
3.9. Tree-Based Protocols . . . . .	67
3.10. Query Tree (QT) Protocol . . . . .	67
3.11. Smart Trend Traversal (STT) Protocol . . . . .	68
3.12. Query Window Tree Protocol . . . . .	69
3.13. Binary Search (BS) Protocol . . . . .	70
3.14. Bitwise Arbitration (BAT) Anti-collision Protocols . . . . .	72
Summary . . . . .	76
Important Questions . . . . .	78
<b>4. WIRELESS SENSOR NETWORKS . . . . .</b>	<b>79</b>
4.1. Introduction . . . . .	80
4.1.1. Wireless Sensor Networks (WSNs) . . . . .	80
4.1.2. Advantages of Wireless Sensor Networks . . . . .	81
4.1.3. Limitations of Wireless Sensor Networks . . . . .	81
4.1.4. Comparison of Wireless Sensor Networks and Adhoc Networks . . . . .	82
4.1.5. Wireless Sensor Network Applications . . . . .	82
4.2. Node . . . . .	83
4.2.1. Components of a Sensor Node . . . . .	84
4.2.2. WSN Network Topologies . . . . .	86
4.3. Architecture and Communication . . . . .	87
4.3.1. Application Layer . . . . .	87
4.3.2. Transport Layer . . . . .	88
4.3.3. Network Layer . . . . .	88
4.3.4. Data Link Layer . . . . .	88
4.3.5. Physical Layer . . . . .	88
4.4. Types of Wireless Sensor Networks . . . . .	89
4.4.1. Terrestrial WSNs . . . . .	89
4.4.2. Underground WSNs . . . . .	89

4.4.3. Underwater WSNs .....	90
4.4.4. Mobile WSNs .....	90
4.4.5. Multimedia WSNs .....	91
4.5. Networking Nodes .....	91
4.5.1. Medium Access Control Layer .....	92
4.6. Securing Communication .....	92
4.7. Standards and Fora .....	93
4.8. Networking and Internet .....	94
4.8.1. IP Addressing .....	94
4.9. Wireless Protocols .....	96
4.9.1. MQTT .....	96
4.9.2. CoAP .....	98
4.10. REST Transferring Data .....	101
4.10.1. Introduction to RESTful web services .....	102
4.10.2. RESTful Methods .....	103
4.10.3. Requirements of REST .....	104
4.10.4. Advantages of REST .....	105
4.10.5. Disadvantages of REST .....	105
Summary .....	106
Important Questions .....	108
<b>5. MOBILITY AND SETTING .....</b>	<b>109</b>
5.1. Introduction .....	110
5.2. Localization .....	110
5.2.1. Localization Techniques .....	115
5.3. Mobility Management .....	117
5.3.1. Mobile IP .....	118
5.3.2. Mobility Management at Different Layers .....	120
5.4. Localization and Handover Management in RFID .....	124
5.4.1. RFID-enabled Localization .....	125
5.4.2. RFID-enabled Movement Detection .....	126
5.4.3. RFID Collision Problem .....	127





5.5. Technology Consideration . . . . .	128
5.5.1. Path Loss Model . . . . .	128
5.5.2. Antenna Radiation Pattern . . . . .	129
5.5.3. Multiple Tags-to-reader Collisions . . . . .	130
5.5.4. Multiple Readers-to-tag Collision . . . . .	133
5.5.5. Reader-to-reader Collision . . . . .	134
5.5.6. Interface for Specific Material. . . . .	135
5.6. Performance Evaluation. . . . .	135
5.6.1. Simulation Setup . . . . .	135
5.6.2. Performances Result . . . . .	136
5.7. Identification of IoT. . . . .	136
5.7.1. IPv6 . . . . .	136
5.7.2. Why does IPv6 suit IoT? . . . . .	138
5.7.3. Uniform Resource Identifier (URI) . . . . .	138
Summary. . . . .	139
Important Questions . . . . .	140
<b>6. DATA ANALYTICS FOR IOE . . . . .</b>	<b>141</b>
6.1. Introduction . . . . .	142
6.2. Apache Hadoop . . . . .	144
6.2.1. Architecture of Hadoop. . . . .	147
6.3. MapReduce Programming Model. . . . .	149
6.3.1. Using Hadoop MapReduce for Batch Data Analysis . . . . .	152
6.4. Apache Oozie. . . . .	153
6.4.1. Features of Oozie. . . . .	156
6.5. Apache Spark. . . . .	156
6.5.1. Spark Clusters . . . . .	158
6.6. Apache Storm. . . . .	159
6.7. Using Apache Storm for Real Time Data Analysis . . . . .	161
6.8. Case Study on Structural Health Monitoring . . . . .	163
6.8.1. Different Sensors Used in Healthcare Monitoring System . . . . .	163
6.8.2. Internet of Things Devices for Healthcare . . . . .	169
6.8.3. An IoT-Aware platform for Structural Health Monitoring. . . . .	171



6.9. Tools for IoT .....	187
6.9.1. CHEF .....	187
6.9.2. Introduction to Puppet .....	188
6.10. IoT System Network Management .....	189
6.10.1. NETCONFIG .....	189
6.10.2. YANG .....	190
6.10.3. NETCONF- YANG Case Studies .....	191
6.11. IoT Code Generator .....	192
Summary .....	193
Important Questions .....	194
<b>INTRODUCTION TO LAB 'INTERNET OF EVERYTHING' .....</b>	<b>195</b>
Experiment 1: Basics of Cooja Simulator .....	196
Experiment 2: Sensor Activation through Button .....	197
Experiment 3: Simple UDP RPL Broadcast with Sky mote .....	204
Experiment 4: 6LoWPAN Protocol with Packet Analyzer .....	210
Experiment 5: Collect View .....	212
<b>INDEX .....</b>	<b>214</b>

# CLOUD COMPUTING AND SERVICES



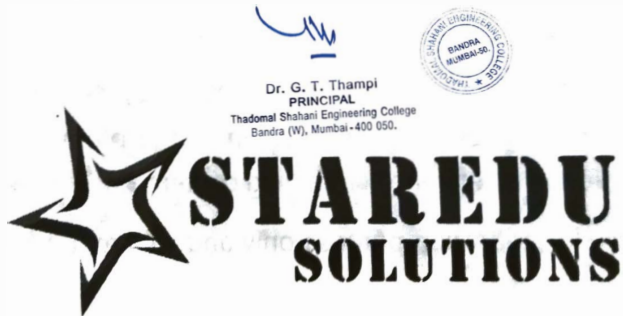
Bhushan A. Jadhav  
Dr. Deven Shah  
Arup Vitthal



[www.staredusolutions.org](http://www.staredusolutions.org)



# CLOUD COMPUTING AND SERVICES



---

Bhushan A. Jadhav | Dr. Deven Shah | Arup Vithal

# Contents

<b>1. INTRODUCTION TO CLOUD COMPUTING</b>	<b>1</b>
➤ Introduction to Cloud Computing	2
→ Definition of cloud Computing	2
→ Characteristics of Cloud Computing	3
➤ Cloud and other similar configurations	4
→ Peer to Peer Architecture	4
→ Client Server Architecture	4
→ Grid Computing	5
➤ Components of Cloud Computing	6
→ Client – The End User	6
→ Cloud Network – The Link	6
→ Cloud Application Programming Interfaces (APIs) – The Gateway	6
→ Large Scale Hardware – Cloud Storage and VM Access	7
➤ Cloud Types	8
→ NIST Model	8
→ Cloud Cube Model	9
➤ Cloud Deployment Models	10
→ Public Cloud	10
→ Private Cloud	11
→ Hybrid Cloud	12
→ Community Cloud	12
→ Comparison between various Cloud Deployment Models	12

➤ Cloud Computing Services	
→ "Infrastructure as a Service (IaaS)"	
→ "Platform as a Service (PaaS)"	
→ "Software as a Service (SaaS)"	
➤ NIST Architecture of Cloud Computing	
→ Cloud Consumer	
→ Cloud Provider	
→ Cloud Auditor	
→ Cloud Broker	
→ Cloud Carrier	
➤ Advantages of Cloud computing	
➤ Cloud Computing Challenges	

## 2. VIRTUALIZATION

➤ Introduction	
➤ Characteristics of Virtualization	
➤ Taxonomy of Virtualization	
→ Full Virtualization	
→ Paravirtualization	
→ Hardware-Assisted Virtualization	
→ Operating System Virtualization	
→ Application Server Virtualization	
→ Application Virtualization	
→ Network Virtualization	
→ Storage Virtualization	
→ Service Virtualization	
➤ Understanding the importance of Type I & Type II Hypervisors	
→ Hosted Structure (Type II Hypervisor)	
→ Bare-metal Structure (Type I Hypervisor)	
➤ Implementation Levels of Virtualization	
➤ Resource Virtualization	
→ CPU Virtualization	

Dr. G. T. Thampi  
PRINCIPAL

Thadomal Shahani Engineering College  
Bandra (W), Mumbai-400 050.



→ Memory Virtualization	49
→ Device and I/O Virtualization	50
➤ Virtualization Vs Cloud Computing	51
➤ Pros and Cons of Virtualization	54
➤ Technology Examples	55
→ KVM Architecture	56
→ Xen Architecture	57
→ VMware	57
→ Hyper-V	59
→ Comparison Table between VMware, Hyper V, KVM and Xen	60
<b>3. CLOUD COMPUTING SERVICES</b>	<b>65</b>
➤ Introduction	66
➤ Exploring Cloud Computing Services	66
➤ SPI Model	67
→ Software as a Service	68
→ Platform as a Service	69
→ Infrastructure as a Service	72
➤ Anything as a Service or Everything as a Service (XaaS)	74
→ Security as a Service (SECaaS)	74
→ Identity Management as a Service (IdMaaS)	75
→ Database as a Service	75
→ Collaboration as a Service	77
→ Compliance as a Service (CaaS)	77
→ Monitoring as a Service	78
→ Communication as a Service	78
→ Disaster Recovery as a Service	79
→ Backup as a Service	79
→ Storage as a Service	80
→ Network as a Service	80
→ Analytics as a Service (AaaS)	81

<b>4. CLOUD IMPLEMENTATION, PROGRAMMING AND MOBILE CLOUD COMPUTING</b>	<b>87</b>
➤ Introduction . . . . .	88
➤ OpenStack Cloud Architecture . . . . .	88
→ Components of OpenStack . . . . .	89
→ Open Stack Architecture . . . . .	90
→ Features and Benefits of OpenStack . . . . .	94
→ Modes of Operations of OpenStack . . . . .	94
➤ Programming Support for Google Apps Engine . . . . .	95
→ GFS . . . . .	95
→ Bigtables . . . . .	98
→ Chubby . . . . .	101
→ Google APIs . . . . .	104
➤ Mobile Cloud Computing . . . . .	104
→ Definition . . . . .	105
→ Basic Components of Mobile Cloud Computing . . . . .	106
→ Additional Components of Mobile Cloud Computing . . . . .	108
→ Architecture of Mobile Cloud Computing . . . . .	112
→ Benefits of Mobile Cloud Computing . . . . .	114
→ Challenges of Mobile Cloud Computing . . . . .	116
<b>5. EXPLORING THE COMPONENTS OF AMAZON WEB SERVICES</b>	<b>123</b>
➤ Amazon Web Services . . . . .	124
➤ AWS Cloud Computing Platform . . . . .	126
→ Elastic Compute Cloud (EC2) . . . . .	126
→ Simple Storage Service (S3) . . . . .	131
→ Elastic Block Storage (EBS) . . . . .	134
→ Amazon Virtual Private Cloud (Amazon VPC) . . . . .	137
→ Elastic Load Balancing (ELB) . . . . .	140
<b>6. CLOUD BACKUP &amp; SOLUTIONS</b>	<b>147</b>
➤ Cloud Backup Solutions and their Features . . . . .	148
→ Where is the data stored in the cloud? . . . . .	148
→ How is data stored? . . . . .	148

Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.




→	Difference between Cloud Sync and Cloud Backup . . . . .	149
→	Benefits and Risks. . . . .	150
➤	Cloud Data Management Interface (CDMI) . . . . .	151
➤	Cloud Storage Gateways (CSGs). . . . .	153
→	Use of CSG . . . . .	154
→	Advantages of Using a CSG . . . . .	156
➤	Comparison between different Cloud Platforms: Amazon Web Services and OpenStack . . .	157
→	AWS Services . . . . .	157
→	OpenStack Services . . . . .	163
→	Comparing Amazon Web Services and Open Stack . . . . .	171

**LAB**

**180**

**INDEX**

**243**

  
**Dr. G. T. Thampi**  
**PRINCIPAL**  
 Thadomal Shahani Engineering College  
 Bandra (W), Mumbai - 400 050.





## Books by Author – Shilpa Ingoley



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai - 400 050.





Sem  
**4**



Information Technology



# Computer Network & Network Design

- Manoj S. Kavedia

- Shilpa Ingoley

☆ With All Latest Solved University Q. Papers

Course Code  
(ITC402)  
(Compulsory  
Subject)



e-books  
(PDF download)



Download App



SCAN TO VISIT



ISBN  
978-93-90376-29-2



M4-27A



Price ₹ 365/-

with  
**MCQ's**



**FREE  
DOWNLOAD**

Sample chapter from our Android App



**TECH-NEO  
PUBLICATIONS**

*Where Authors Inspire Innovation*

A Sachin Shah Venture



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai-400 050.





## About the Authors



### Er. Manoj S. Kavedia

Assistant Professor,  
Electronics and Telecommunication Department.  
Thadomal Shahani Engineering College, Bandra, Mumbai  
(Total 24 years of Teaching Experience)



### Shilpa Ingoley

Assistant Professor,  
M.E. Computer Engineering (Mumbai University) Computer Department  
Thadomal Shahani Engineering College, Bandra, Mumbai  
(Total 21 years of Teaching Experience)



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai-400 050.

**MU**

**Sem  
4**



**Information Technology**

Name of Subject	Author
Applied Mathematics-IV	Baphana R M (Adjunct Faculty, COEP, Pune)
Computer Network and Network Design	Manoj S. Kavedia, Shilpa Ingoley
Operating System	Manoj S. Kavedia
Automata Theory	Ashish Budhrani
Computer Organization and Architecture	Velankar Shrikant, Shah Urvashi

For Orders Contact :

### Krishna Book Collections

Ground Floor, Krishna Niwas Building, Behind BEST Niwas  
Building, Near Napoo Hall, Chandavarkar Road, Matunga East, Mumbai 400019.

E-mail : dharmeshsota05@gmail.com

Mobile No.: Dharmesh Sota - 9820741455, Tulsidas Sota - 9833133921 / 9833082745 / 9833082761



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai-400 050.



Books are available on Flipkart amazon boohwalas.com

boohwalas.com  
the online book shop



University of Mumbai

# Computer Network and Network Design

(Course Code : ITC402)

Semester IV - Information Technology

Strictly as per the New Syllabus (REV-2019 'C' Scheme) of  
Mumbai University w.e.f. academic year 2020-2021



**Er. Manoj S. Kavedia**

Assistant Professor

Department of Electronics and Telecommunication

Thadomal Shahani Engineering College (TSEC), Bandra, Mumbai

**Shilpa Ingoley**

M.E. Computer Engineering (Mumbai University)

Assistant professor,


Computer Department

Thadomal Shahani Engineering College-Bandra

 **TECH-NEO**  
**PUBLICATIONS**

*Where Authors Inspire Innovation*

A Sachin Shah Venture

  
Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai-400 050.



M4-27

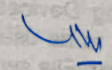




## Index

◆ Chapter 1 : Introduction to Computer Network .....	1-1 to 1-53
◆ Chapter 2 : Physical Layer and Data Link Layer .....	2-1 to 2-81
◆ Chapter 3 : Network Layer .....	3-1 to 3-62
◆ Chapter 4 : Transport Layer and Session Layer .....	4-1 to 4-36
◆ Chapter 5 : Presentation Layer and Application Layer .....	5-1 to 5-39
◆ Chapter 6 : Network Design Concepts .....	6-1 to 6-50

□□□

  
 Dr. G. T. Thampi  
 PRINCIPAL  
 Thadomal Shahani Engineering College  
 Bandra (W), Mumbai-400 059.



**MU****Sem  
5****Computer Engineering**New Syllabus  
2021-22

Compulsory Subject (CSC503)

# Computer Network

Manoj S. Kavedia

Shilpa Ingoley

(Thadomal Shahani Engineering College, Bandra, Mumbai)

With Typical  
**MCQ's****TECH-NEO  
PUBLICATIONS**

Where Authors Inspire Innovation

A Sachin Shah Venture

www.techneobooks.in  
info@techneobooks.in

- ☆ With Solved Latest  
**UNIVERSITY QUESTION PAPERS.**
- ☆ Self Explanatory **Diagrams.**
- ☆ Comparisons in **Tabular Form.**
- ☆ **Multiple Choice Questions.**

M5-55A



Price ₹ 325/-

Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai-400 050.

## About the Authors



**Er. Manoj S. Kavedia**

Assistant Professor,  
Electronics and Telecommunication Department,  
Thadomal Shahani Engineering College, Bandra, Mumbai  
(Total 24 years of Teaching Experience)



**Shilpa Ingoley**

Assistant Professor,  
M.E. Computer Engineering (Mumbai University) Computer Department  
Thadomal Shahani Engineering College, Bandra, Mumbai  
(Total 21 years of Teaching Experience)



Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai-400 050.



## Sem 5 Computer Engineering

Course Code	Compulsory Subjects
CSC501	Theoretical Computer Science
CSC502	Software Engineering
CSC503	Computer Network
CSC504	Data warehousing and Mining
Course Code	Departmental Elective
CSD05012	Internet Programming
CSD05013	Advance Database Management System

ISBN  
978-93-90904-11-2



e-books  
(PDF download)

Google Play  
Download App



SCAN TO VISIT

For Orders Contact :

### Krishna Book Collections

Ground Floor, Krishna Niwas Building, Behind BEST Niwas Building, Near Napoo Hall,  
Chandavarkar Road, Matunga East, Mumbai 400019. E-mail : dharmeshsota05@gmail.com

Mobile No.: Dharmesh Sota - 98207 41455, Tulsidas Sota - 98331 33921 / 98330 82745 / 98330 82761

Books are available on **Flipkart** **amazon** **boothwalas.com**  
the online book shop



University of Mumbai

# Computer Network

(Code : CSC503)

Semester V - Computer Engineering

Strictly as per the Choice Based Credit and Grading System  
(Revise 2019) of Mumbai University w.e.f. academic year 2021-2022

**Er. Manoj S. Kavedia**

Assistant Professor,

Department of Electronics and Telecommunication

Thadomal Shahani Engineering College (TSEC), Bandra, Mumbai

ME Electronics

Pursing PHD in Internet of Things

**Shilpa Ingole**

M.E. Computer Engineering (Mumbai University)

Assistant professor,

Computer Department

Thadomal Shahani Engineering College-Bandra

Dr. G. T. Thampi  
PRINCIPAL  
Thadomal Shahani Engineering College  
Bandra (W), Mumbai-400 050.



**TECH-NEO**  
PUBLICATIONS

*Where Authors Inspire Innovation*

A Sachin Shah Venture

M5-55



## Index

◆ Chapter 1 : Introduction to Networking.....	1-1 to 1-56
◆ Chapter 2 : Physical Layer.....	2-1 to 2-29
◆ Chapter 3 : Data Link Layer .....	3-1 to 3-46
◆ Chapter 4 : Network Layer .....	4-1 to 4-71
◆ Chapter 5 : Transport Layer.....	5-1 to 5-35
◆ Chapter 6 : Application Layer.....	6-1 to 6-26

### ► Multiple Choice Questions

□□□