

Table of Contents

Introduction.....	xvii
Chapter 1: Introduction to Data Warehousing	1
1.1 The Need for Data Warehousing	2
1.1.1 Increasing Demand for Strategic Information	3
1.1.2 Inability of Past Decision Support System.....	5
1.2 Operational vs. Decisional Support System.....	6
1.3 Data Warehouse Defined	6
1.3.1 Benefits of Data Warehousing.....	7
1.3.2 Features of Data Warehouse.....	7
1.3.3 Information Flow Mechanism.....	8
1.4 Role of Metadata.....	9
1.4.1 Classification of Metadata	9
1.5 Data Warehouse Architecture.....	10
1.5.1 Data Sources.....	11
1.5.2 Data Staging.....	12
1.5.3 ETL Process.....	12
1.5.4 Data Aggregation	14
1.6 Different Types of Architecture of Data Warehouse	14
1.6.1 Data Warehouse Architecture (Basic).....	15
1.6.2 Data Warehouse Architecture with a Staging Area.....	15
1.6.3 Data Warehouse Architecture with a Staging Area & Data Marts ...	16
1.7 Data Warehouse and Data Mart	16
1.7.1 Differences between Data Warehouse and Data Mart.....	17
1.7.2 Types of Data Marts	18
1.8 Data Warehousing Design Strategies	18

1.8.1 Top-down Approach.....	18
1.8.2 Bottom-up Approach.....	19
Summary.....	19
Review Exercise	20
Multiple Choice Questions.....	20
Descriptive Questions.....	22

Chapter 2: Dimensional Modeling.....31

2.1 Data Warehouse Modeling versus Operational Modeling.....	32
2.2 Dimensional Model versus Entity-Relationship Model	33
2.3 Features of a Good Dimensional Model	34
2.4 The Star Schema	35
2.4.1 Query Execution in Star Schema.....	37
2.5 The Snowflake Schema	38
2.6 Introducing the Fact Table.....	40
2.7 Introducing the Dimension Table	40
2.8 The Factless Fact Table.....	41
2.8.1 Factless Fact Tables for Events.....	42
2.8.2 Factless Fact Tables for Conditions	42
2.9 Updates to Dimension Tables.....	42
2.10 Slowly Changing Dimensions.....	43
2.10.1 Slowly Changing Dimension Type 1	44
2.10.2 Slowly Changing Dimension Type 2	44
2.10.3 Slowly Changing Dimension Type 3	44
2.11 Large Dimension Tables.....	44
2.12 Rapidly Changing or Large Slowly Changing Dimensions.....	46
2.13 Junk Dimensions.....	47
2.14 Keys in the Data Warehouse Schema	48
2.14.1 Primary Key.....	48
2.14.2 Surrogate Key.....	49

2.14.3 Foreign Keys.....	50
2.15 Aggregate Table	50
2.16 Fact Constellation Schema or Family of Stars.....	51
Summary.....	52
Review Exercise	52
Multiple Choice Questions.....	52
Descriptive Questions.....	54
Chapter 3: ETL Process.....	69
3.1 Overview of ETL Process.....	70
3.1.1 Identification of Data Sources	71
3.1.2 Challenges in ETL Functions.....	71
3.2 Data Extraction.....	72
3.2.1 Logical Extraction Methods.....	72
3.2.2 Physical Extraction Methods	75
3.2.3 Change Data Capture	76
3.2.4 Ways of Extracting Data.....	77
3.3 Data Transformation: Tasks Involved in Data Transformation.....	78
3.3.1 Transforming Data Using SQL.....	81
3.3.2 Transforming Data Using PL/SQL.....	82
3.3.3 Transforming Data Using Table Functions.....	83
3.4 Data Loading: Techniques of Data Loading.....	83
3.4.1 Data Quality.....	86
3.4.2 Cleaning.....	86
3.4.3 Need for Data Cleaning	87
3.4.4 Issues in Data Cleansing.....	88
3.5 Loading the Fact Tables.....	88
3.6 Loading the Dimension Tables	89
Summary.....	90
Review Exercise	90

Multiple Choice Questions..... 90
Descriptive Questions..... 92

Chapter 4: Online Analytical Processing (OLAP)..... 107

4.1 Need for Online Analytical Processing 108
4.1.1 Requirement of Fast Access and Fast Calculations of
Complex Data..... 108
4.1.2 Need for OLAP..... 109
4.1.3 Drawbacks of Other Methods of Analysis..... 109
4.2 OLTP vs. OLAP 110
4.3 OLAP and Multidimensional Analysis 111
4.3.1 Multidimensional Analysis..... 112
4.3.2 Need for multidimensional Analysis..... 113
4.3.3 Advantages of Multidimensional Analysis 114
4.4 Hypercubes..... 115
4.4.1 Multicubes..... 116
4.5 OLAP Operations in Multidimensional Data Model 117
4.5.1 Roll Up..... 117
4.5.2 Drill Down 118
4.5.3 Slice 118
4.5.4 Dice 119
4.5.5 Pivot..... 119
4.6 OLAP Models 120
4.6.1 ROLAP (Relational OLAP)..... 120
4.6.2 MOLAP (Multidimensional OLAP) 121
4.6.3 HOLAP (Hybrid OLAP) 122
4.6.4 DOLAP 123
4.6.5 Other Types of OLAP Servers..... 124
4.7 Popular OLAP tools..... 124

4.7.1	InetSoft's Web-based OLAP Server Solution.....	124
4.7.2	Pentaho.....	125
4.7.3	Arbor Essbase OLAP Server 5.....	126
4.7.4	Microstrategy Analytics.....	126
	Summary.....	127
	Review Exercise.....	127
	Multiple Choice Questions.....	127
	Descriptive Questions.....	128
Chapter 5: Introduction to Data Mining.....		135
5.1	What is Data Mining?.....	136
5.2	Knowledge Discovery in Database (KDD).....	136
5.3	What Type of Data can be Mined?.....	138
5.4	Concepts Related to Data Mining.....	143
5.4.1	Structure of data mining.....	143
5.4.2	Architecture of data mining system.....	144
5.4.3	Classification of Data Mining Systems.....	145
5.4.4	Integration of a Data Mining System with a Data Warehouse.....	146
5.5	Data Mining Techniques.....	147
5.6	Applications of Data Mining.....	149
5.7	Issues in Data Mining.....	150
5.7.1	Ethical issues in data mining.....	151
	Summary.....	152
	Review Exercise.....	153
	Multiple Choice Questions.....	153
	Descriptive Questions.....	154
Chapter 6: Data Exploration.....		161
6.1	Exploring Data.....	162
6.2	Types of Data Attributes.....	163

6.2.1	Continuous and Discrete Attributes.....	164
6.3	Statistical Description of Data.....	165
6.3.1	Basic Statistical Methods.....	166
6.3.2	Dispersion of data.....	167
6.3.3	Visualization of Statistical Description of Data.....	168
6.4	Data Visualization.....	172
6.4.1	Data Visualization process.....	172
6.4.2	Visualization Techniques.....	173
6.4.3	Advantages of Visualization.....	183
6.5	Measuring Similarity and Dissimilarity in Data.....	184
6.5.1	Data and dissimilarity matrix.....	185
6.5.2	Distance Measures and Dissimilarity Measures.....	186
6.5.3	Cosine similarity.....	188
	Summary.....	190
	Review Exercise.....	190
	Multiple Choice Questions.....	190
	Descriptive Questions.....	191

Chapter 7: Data Preprocessing 197

7.1	Why Preprocessing?.....	198
7.1.1	Characteristics of Good Quality Data.....	198
7.1.2	Steps in Data Preprocessing.....	199
7.2	Data Cleaning.....	200
7.2.1	Missing Values.....	200
7.2.2	Noisy Data.....	202
7.2.3	Data Cleaning as a Process.....	204
7.3	Data Integration.....	205
7.4	Data Reduction.....	207
7.4.1	Data Cube Aggregation.....	208

7.4.2	Attribute Subset Selection.....	209
7.4.3	Dimensionality Reduction.....	211
7.4.4	Numerosity Reduction.....	211
7.4.5	Regression and Log-Linear Models.....	211
7.4.6	Histograms.....	212
7.4.7	Clustering and Sampling.....	213
7.5	Data Transformation.....	217
7.5.1	Normalization	218
7.6	Data Discretization and Concept Hierarchy Generation	219
7.6.1	Binning.....	220
7.6.2	Histogram Analysis.....	220
	Summary.....	220
	Review Exercise	221
	Multiple Choice Questions.....	221
	Descriptive Questions.....	223
Chapter 8: Classification and Prediction.....		233
8.1	Basic Concepts.....	234
8.1.1	Data Classification.....	234
8.1.2	Data Preparation.....	235
8.1.3	Data Types	235
8.2	Classification Methods.....	236
8.2.1	Decision Tree Induction	238
8.2.2	Steps for Decision Tree Induction	239
8.2.3	Classification Using Decision Tree	240
8.2.4	Decision Tree Algorithm	242
8.2.5	Decision Tree Inducers.....	243
8.2.6	Attribute Selection Measures.....	245
8.2.7	Tree Pruning.....	250

8.2.8	Rule-Based Classification	252
8.2.9	Rule Induction Using a Sequential Covering Algorithm	253
8.3	Bayesian Classification.....	254
8.3.1	Bayes' Theorem	255
8.3.2	Bayes' Rule	255
8.3.3	Naïve Bayes Classifier.....	256
8.4	Classification by Artificial Neural Networks	257
8.4.1	Working of a Multi-Layer Neural Network.....	258
8.4.2	Backpropagation.....	259
8.5	Lazy Learners	259
8.6	Associative Classification	260
8.7	Other Classification Methods.....	260
8.7.1	Genetic Algorithms.....	260
8.7.2	Fuzzy Classifiers	261
8.7.3	Rough Sets	262
8.7.4	k-Nearest Neighbor Classifier (k-NN)	263
8.8	Prediction.....	264
8.8.1	Structure of Regression Model.....	265
8.8.2	Correlation	265
8.8.3	Simple Linear Regression	266
8.8.4	Multiple Linear Regression (Multivariable Linear Regression).....	267
8.8.5	Nonlinear Regression.....	268
8.8.6	Least Squares Method	269
8.9	Model Evaluation and Selection.....	270
8.9.1	Data Partitioning.....	272
8.9.2	Holdout.....	274
8.9.3	Random Sampling	275
8.9.4	Cross-Validation	275

8.9.5	Bootstrap	276
8.9.6	Comparing Classifier Performance Using ROC Curves.....	277
8.10	Combining Classifiers (Ensemble Methods).....	279
8.10.1	Bagging.....	279
8.10.2	Boosting.....	280
8.10.3	Random Forests.....	283
	Summary.....	284
	Review Exercise	284
	Multiple Choice Questions.....	284
	Descriptive Questions.....	286

Chapter 9: Clustering and Trends in Data Mining 293

9.1	Cluster Analysis	293
9.1.1	Requirements of a Good Clustering Algorithm.....	295
9.2	Types of Data in Clustering.....	296
9.3	Categorization of Major Clustering Methods.....	297
9.4	Partitioning Methods.....	298
9.4.1	Simple K-Means Clustering	299
9.4.2	K-Medoids Clustering Method.....	300
9.5	Hierarchical Methods	301
9.5.1	Distance Metrics for Hierarchical Clustering	303
9.5.2	Advantages and Disadvantages of Hierarchical Methods.....	305
9.5.3	Hierarchical Algorithms - AGNES & DIANA Agglomerative Nesting (AGNES)	306
9.5.4	Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH).....	307
9.6	Density-Based Clustering	309
9.6.1	Density-Based Spatial Clustering of Applications with Noise (DBSCAN)	310

9.6.2	Ordering Points to Identify the Clustering Structure (OPTICS).....	312
	Summary.....	314
	Review Exercise	315
	Multiple Choice Questions.....	315
	Descriptive Questions.....	316
Chapter 10:	Frequent Pattern Mining	329
10.1	Market Basket Analysis.....	330
10.1.1	Frequent Itemsets, Closed Itemsets, and Association Rules.....	332
10.1.2	Frequent Pattern Mining Technique.....	334
10.2	Efficient and Scalable Frequent Pattern Mining Methods	335
10.2.1	Apriori Algorithm for Finding Frequent Itemsets Using Candidate Generation.....	336
10.2.2	Generating Association Rules from Frequent Itemsets.....	340
10.2.3	Improving Efficiency of Apriori Algorithm.....	341
10.2.4	Pattern Growth Approach for Mining Frequent Itemsets.....	342
10.2.5	Mining Frequent Itemsets Using VDFs.....	344
10.2.6	Mining Closed and Maximal Patterns	346
10.3	Multilevel and Multidimensional Association Rules.....	348
10.4	Association Rule Mining to Correlation Analysis.....	351
10.4.1	Pattern Evaluation	352
10.5	Constraint-Based Association Mining.....	354
	Summary.....	355
	Review Exercise	355
	Multiple Choice Questions.....	355
	Descriptive Questions.....	357
Index	365