List of Tables
Table 1.1
Brief Summary of Industry wise utilization of Data Warehouse Applications

Table 1.2
Possible Data Quality Issues

Table 3.1
Kimball vs Inmon’s Ideology

Table 4.1
Comparative Analysis of ETL and ELT

Table 4.2
Build vs. Buy Analysis

Table 4.3
Workflow of ETL Prototype

Table 4.4
Software requirements for the ETL prototype

Table 4.5
Testing Goals for an ETL Application

Table 4.6
ETL Test Cases of Prime Importance

Table 5.1
Data Quality Concerns from User’s Perspective

Table 5.2
Anomalies Observed Before and After the Introduction of Automated Testing in Corresponding Samples
Comparison of Calculated Values of t and p with Tabulated Values at 5% and 1% Level of Significance
........................................................................................................................................... 126
Table 6.1
Source Student Table (Student Data in a Flat File) ........................................ 142
Table 6.2
Company Details (Normalized Table) ................................................................. 143
Table 6.3
Class Details (Normalized Table) ......................................................................... 14
Table 6.4
Stud (Normalized Table) ......................................................................................... 143
Table 6.5
Class Taken (Normalized Table) ........................................................................... 143

List of Figures
<table>
<thead>
<tr>
<th>Figure No.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1</td>
<td>Dissection of a Data Warehouse</td>
<td>5</td>
</tr>
<tr>
<td>Figure 1.2</td>
<td>The Star Schema</td>
<td>8</td>
</tr>
<tr>
<td>Figure 1.3</td>
<td>The Snowflake Schema</td>
<td>9</td>
</tr>
<tr>
<td>Figure 1.4</td>
<td>The Data Problem Initiation</td>
<td>18</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>The evolution of Data Warehouses</td>
<td>53</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Data Warehouse Architecture as Proposed by W.H. Inmon</td>
<td>57</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Data Warehouse Architecture as Defined by Ralph Kimball</td>
<td>57</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>Independent Data Marts</td>
<td>58</td>
</tr>
<tr>
<td>Figure 3.5</td>
<td>A Hub and Spoke Architecture</td>
<td>59</td>
</tr>
<tr>
<td>Figure 3.6</td>
<td>Centralized Data Warehouse Architecture</td>
<td>59</td>
</tr>
<tr>
<td>Figure 3.7</td>
<td>Federated Data Warehouse Architecture</td>
<td>60</td>
</tr>
<tr>
<td>Figure 3.8</td>
<td>The ETL Design Cycle</td>
<td>63</td>
</tr>
<tr>
<td>Figure 3.9</td>
<td>Estimated Percentage Usage of various ETL options</td>
<td>64</td>
</tr>
<tr>
<td>Figure 3.10</td>
<td>UML Representation of ETL Design Logic</td>
<td>65</td>
</tr>
<tr>
<td>Figure 3.11</td>
<td>Encryption and Masking Logic</td>
<td>67</td>
</tr>
</tbody>
</table>
Figure 3.12
Synthetic Records with Random Values ............................................. 68
Figure 3.13
A Scrambled Record ........................................................................... 68
Figure 3.14
Use of Look-up Fields ......................................................................... 69
Figure 3.15
Synthetic Data Set Generator Architecture ........................................ 70
Figure 3.16
The Multiplication logic of DMA Algorithm ....................................... 71
Figure 3.17
The ability of DSG to exploit resembling domain sets at different locations.... 72
Figure 3.18
The clubbing of resembling domain sets available at different locations ....... 73
Figure 3.19
Flow Chart depicting the working of DSG ........................................... 74
Figure 3.20
Liberty to specify desired column names and data types in the synthetic
database table ....................................................................................... 76
Figure 3.21
Selection of the Data Source Donor ..................................................... 77
Figure 3.22
DSG showing mapping options before performing table merge operation ..... 78
Figure 3.23
Mapping of resembling domain sets and declaration of number of synthetic
records needed ....................................................................................... 78
Figure 3.24
The DSG seeking directions before generating synthetic records ............. 79
Figure 3.25
Field values extracted from donor database are presented for user review .... 79
Figure 3.26
Sample of generated synthetic records ............................................... 80
Figure 4.1
Extraction Transformation and Loading Architecture .......................... 86
Figure 4.2
Extraction Loading and Transformation Architecture ................................. 87
Figure 4.3
A Simple Data Extraction Module ............................................................... 92
Figure 4.4
Data Transformation Logic ........................................................................... 93
Figure 4.5
Anatomy of ETL Routine ............................................................................. 96
Figure 4.6
Selection of Data Source and Source Data type ........................................... 101
Figure 4.7
Successful Connections between Source and Target Databases .................... 102
Figure 4.8
Replicated Schema of Source Database to be extracted ............................... 102
Figure 4.9
Replicated Schema of Target Database ....................................................... 103
Figure 4.10
Mapping of Source and Target Database Fields .......................................... 103
Figure 4.11
Validating Integrity Constraints ................................................................... 104
Figure 4.12
Source Data at Staging Area ........................................................................ 105
Figure 4.13
Data from different sources at Staging Area ................................................. 105
Figure 4.14
Log of Rejected Records .................................................................

106

Figure 4.15

Lexical Anomalies ...........................................................................

107

Figure 4.16

Format Errors .................................................................................. 

107

Figure 4.17

Irregularities in fact representation ..................................................

108

Figure 4.18

Integrity Constraint Violations ........................................................

108

Figure 4.19

Contradictions in records ............................................................... 

109

Figure 4.20

Duplicate records with different serial number ............................... 

109

Figure 4.21

Missing Values ................................................................................ 

110

Figure 4.22

The Loading Logic of ETL Prototype ............................................... 

118

Figure 5.1

Bar Chart Representation of Statistical Analysis ............................. 

127

Figure 6.1

The Data Sharehouse Architecture (Macro View) ............................ 

140

Figure 6.2