P-3562

DESIGN OF INTELLIGENT DECISION MAKING SYSTEM AND ETL MANAGEMENT

PROJECT REPORT

Submitted by

R.JANAKI

Reg. No: 0920108006

In partial fulfillment for the award of the degree of

MASTER OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

KUMARAGURU COLLEGE OF TECHNOLOGY

(An Autonomous Institution Affiliated to Anna University, Coimbatore)

COIMBATORE – 641 049

APRIL 2011

KUMARAGURU COLLEGE OF TECHNOLOGY

(An Autonomous Institution Affiliated to Anna University, Coimbatore)

COIMBATORE – 641 049

Department of Computer Science and Engineering

PROJECT WORK

APRIL 2011

This is to certify that the project entitled

DESIGN OF INTELLIGENT DECISION MAKING SYSTEM AND ETL MANAGEMENT

is the bonafide record of project work done by

R.JANAKI

Register No: 0920108006

of M.E. (Computer Science and Engineering) during the year 2010-2011.

S. J.

Project Guide

Head of the Department

Submitted for the Project Viva-Voce examination held on 25/4/2011

DECLARATION

I affirm that the project work titled **DESIGN OF INTELLIGENT DECISION MAKING SYSTEM AND ETL MANAGEMENT** being submitted in partial fulfillment for the award of M.E Computer Science and Engineering is the original work carried out by me. It has not formed the part of any other project work submitted for the award of any degree or diploma, either in this or any other University.

R. Janaki

JANAKI R

Register No: 0920108006

I certify that the declaration made above by the candidate is true.

S. Thangasamy

Dr.S.Thangasamy Ph.D.,
Dean Research and Development(R&D)
Kumaraguru College of Technology,
(An Autonomous Institution)
Coimbatore-641 049.

MARAGURU COLLEGE OF TECHNOLOGY

Coimbatore - 641 049.

Department of Computer Science and Engineering

CSIR Sponsored

National Conference on Service Oriented Computing

Certificate

This is to certify that ~~Dr./Prof./Mr./Mrs.~~/Ms. R.JANAKI

Maraguru College of Technology has presented a technical paper

titled DESIGN OF INTELLIGENT DECISION MAKING SYSTEM AND ETL WORKFLOW MANAGEMENT in the National Conference on Service Oriented Computing(NCSOC 2011) held on 24th & 25th March 2011.

(Dr.D. Chitra Devakala)
Secretary

(Prof.P.Devaki)
HoD

(Dr.S. Ramachandran)
Principal

ACKNOWLEDGEMENT

If the words are considered as symbols of approval and tokens of acknowledge,then the words play the heralding role of expressing our gratitude.

First and foremost, I would like to thank the Lord Almighty for enabling me to complete this project.

I express my profound gratitude to our Chairman **Padmabhusan Arutselvar Dr.N.Mahalingam B.Sc., F.I.E.,** for giving this opportunity to pursue this course.

I would like to thank **Dr.S.Ramachandran Ph.D.,** *Principal* for providing the necessary facilities to complete my thesis.

I also thank **Mrs.P.Devaki M.E.,(Ph.D.,)** *HOD,* Department of Computer Science and Engineering, for her support and timely motivation.

I thank all project committee members for their comments and advice during the reviews.Special thanks to my project coordinators, **Mrs.V.Vanitha M.E., (Ph.D.,)**,*Associate Professor* and **Mr.V.Subramani M.Tech.,** *Associate Professor*, Department of Computer science and Engineering, for arranging brain storming project review sessions.

I register my sincere thanks to my Guide **Dr.S.Thangasamy Ph.D.,** *Dean, Research and Development,*my thesis advisor. I am grateful for his support, encouragement and ideas. I thank him for the countless hours he has spent with me for giving valuable suggestions and providing constant motivation which propelled me to complete this project.

I would like to convey my honest thanks to all **Teaching** staff members and **Non Teaching** staffs of the department for their support. I would like to thank all my classmates who gave me a proper light moments and study breaks apart from extending some technical support whenever I needed them most.

I dedicate this project work to my **parents** for no reasons but feeling from bottom of my heart, without their love this work wouldn't be possible.

ABSTRACT

In today's economy, the importance of effective and efficient decision making has become increasingly important in order to stay competitive in a global market set. Obtaining most relevant data and outputs is the key for best decisions on every management level. In this way, Business Intelligence (BI) is helpful in intelligent running of business.It needs a data warehouse for storing historic business related data and a computer based decision support system for end decision making. A Data warehouse assists in critical business decision making by collecting and organizing data from huge disparate operational sources of the organization. The three step process of extraction, transformation and loading (ETL) helps to build a data warehouse.

Though many companies have invested huge amounts in BI related projects, building a domain specific BI system is still a challenging one. In this project, retail based Business Intelligence system is proposed which merges retail data warehouse with an intelligent decision making system.

Sample operational data is collected from retail billing software where the data comes from heterogeneous sources and not in unique format.The proposed system, turns this inconsistent data into unique format through ETL process and loads into the target data warehouse with retail data.An intelligent decision making system is designed with various strategic questions put forth by retail decision makers as its basis. The system is developed using complex queries,statistical models such as AutoRegressive Integrated Moving Average Model (ARIMA) and regression model.Retail sellers can use this system for generating various reports,sales forecasting and other retail related decision making.

ஆய்வுச்சுருக்கம்

இன்றைய பொருளாதாரத்தில், பயன்பெறும் வகையில் திறமையாக முடிவெடுப்பது மேன் மேலும் அதிக முக்கியத்துவம் பெறுகிறது. உலக மயமாக்கப்பட்டுள்ள சந்தையில் போட்டியிடுவதற்கு இது மிகவும் தேவைப்படுகிறது. மிகவும் சம்பந்தமுடைய தகவல்களும், வெளியீடுகளும், நிர்வாகரீதியாக சிறந்த முடிவுகளை எடுக்க உதவும் திறவு கோலாகும். இவ்வகையில் தொழிலை செவ்வனேநடத்துவதற்கு வியாபார நுணுக்க அறிவுபயனளிக்கும், இதற்கு, ஒரு தகவல் களஞ்சியத்தில் வியாபாரம் பற்றிய வரலாற்றுக் குறிப்புகளை சேமிக்க வேண்டும். மேலும், கணிணி அமைப்பின் உதவியுடன் இறுதி முடிவைத் திறம்பட எடுக்க முடியும். ஒரு தகவல் களஞ்சியம், நிறுவனத்திற்கு பல்வேறு வகையில் இயங்கும் மூலஅமைப்புகளிலிருந்துவிவரங்களை சேகரித்து,சீர்படுத்திக்கொடுத்து, முக்கியமான முடிவுகளை எடுக்க உதவுகிறது. வெளிக்கொணர்தல், மாற்றியமைத்தல்,உட்செலுத்துதல் ஆகிய மூன்று படிகள் கொண்ட முறையின் மூலம் ஒரு தகவல் களஞ்சியத்தை நாம் உருவாக்கலாம்.

வியாபார நுணுக்க அறிவு சம்பந்தப்பட்ட திட்டங்கள் பலவற்றில் பல நிறுவனங்கள் பெரும் தொகைகளை முதலீடு செய்துள்ளன. இருப்பினும், ஒரு குறிப்பிட்ட துறை சார்ந்த வியாபார நுணுக்க அறிவு பெறும் அமைப்பை உருவாக்குவது இன்னமும் ஒரு சவாலாகவே இருக்கிறது. இந்த ஆய்வு, சில்லறை வியாபாரத்தகவல் களஞ்சியத்தை திறம்பட முடிவெடுக்கும் அமைப்புடன் இணைக்கிறது.

சில்லறை வியாபார விலை விவரங்கள் அடங்கிய மென்பொருளிலிருந்து மாதிரியாக இயக்க ரீதியான தகவல்கள் சேகரிக்கப்படுகின்றன. இதில் உள்ள விவரங்கள் மாறுபட்டமூலங்களிலிருந்து வருவன. அவற்றிற்கு சீரான வடிவமைப்பு இல்லை. இந்த ஆய்வில் உருவாக்கப்படவுள்ள அமைப்பு, இந்தப்பலவகையான சீரற்ற விவரங்களை மேலே கண்ட மூன்று படி முறை மூலம் சீரமைத்து, ஒரு ஒழுங்கான வடிவத்தை அவற்றிற்குக் கொடுக்கிறது. இவற்றை அது நம் இலக்காகிய தகவல் களஞ்சியத்தின் உள்ளே செலுத்துகிறது. ஒரு திறமையான முடிவெடுக்கும் அமைப்பு, சில்லறை வியாபார ரீதியான முடிவுகள் எடுப்போர் எழுப்பும் செயல் திட்டம் சார்ந்த வினாக்களை அடிப்படையாகக் கொண்டு உருவாக்கப்படுகிறது. இந்த அமைப்பு, சிக்கலான தகவல்-கோரிக்கைகள், விவரங்கள் சேகரிப்பில் எதிர்கொள்ளும் எண்ணற்ற விதிமுறைகள், புள்ளி விவரம் சம்பந்நதப்பட்ட மாதிரிக்கருவிகள் ஆகியனவற்றைப் பயன்படுத்தி உருவாக்கப்படுகிறது. புள்ளிவிவரம் என்ற வகையில் தானியங்கி பின்வாங்கு உருங்கிணைந்து அசையும் சராசரிக் கருவி, பின் நோக்கி இயக்கக்கூடிய கருவி ஆகியவை பயன்படுத்தப்படுகின்றன. சில்லறை வியாபாரிகள் பல்வேறு அறிக்கைகள் எதிர்கால விற்பனை வாய்ப்புகள், இதர சில்லறை வியாபார சம்பந்தப்பட்ட முடிவுகள் ஆகிய விஷயங்களை உற்பத்தி செய்ய, இந்த ஆய்வின் மூலம் உருவாக்கிய இந்த அமைப்பைப் பயன்படுத்திக் கொள்ளலாம்.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATION	EXPANSION
BI	Business Intelligence
IDSS	Intelligent Decision Support System
ETL	Extraction-Transformation-Loading
OLAP	OnLine Analytical Processing
OLTP	OnLine Transactional Processing
ARIMA	AutoRegressive Integrated Moving Average

LIST OF SYMBOLS

ξ	Autoregressive constant
ϕ_1, ϕ_2, ϕ_3	Autoregressive model parameters
μ	Moving average constant
θ_1	Moving Average parameter

CHAPTER 1

INTRODUCTION

1.1 BUSINESS INTELLIGENCE

Business Intelligence(BI) [4] means using data assets of the organization to make business related decisions. It involves the gathering, management, and analysis of data for the purpose of turning that data into useful information which is then used to improve decision making. BI is the technology and practice of applying information to make decisions. Data warehousing is the vehicle that drives business intelligence. It is also more comprehensive than data mining. Information shows its real value when many people can use and share it. This is the goal of business intelligence. For information to be usable, it must be trusted, timely, relevant, easy-to-use, and in context. These are all necessary aspects of business intelligence technology.

Different business intelligence tools address the each factor in different ways. Reporting delivers regular, timely information, with the ability to author reports or queries to get specific details. OLAP analysis, with its multiple dimensions, allows one to compare and contrast information against time and other factors to uncover trends. Business Intelligence is one of the emerging fields which are purely concerned with the intelligent running of the business.It involves extracting, gathering, management and analysis of vast amounts of data. It is used in order to gain insights to drive strategic business decisions, and to support operational processes with new functions.

Data mining is used for extracting the hidden information from the databases. It assists the managers within organization to perform important task of decision making. A data warehouse assists in business decision making by collecting data from different operational sources of the organization. Data warehousing consists of a three step process of Extraction,Transformation and Loading (ETL) to manage huge volumes of data.ETL is the most important and time consuming phase of any data warehousing project.

Business intelligence applications can be:

- Mission-critical and integral to an enterprise's operations or occasional to meet a special requirement.

- Enterprise-wide or local to one division, department, or project.
- Centrally initiated or driven by user demand.

Advantages of business intelligence are:

- The high responsiveness of the company to the needs of its customers.
- Recognition of customer needs.
- Ability to act on market changes.
- Optimization of operations.

BI can be seen as an umbrella [4] that covers a whole range of concepts. BI can be approached roughly as being a Data Warehouse (DW), with three layers on top of it: Queries & Reports (Q&R), Online Analytical Processing (OLAP) and Data Mining (DM). Authors and companies adopt this ordering widely. The following figure (figure 1.1) shows the BI pyramid consisting of the concepts described below:

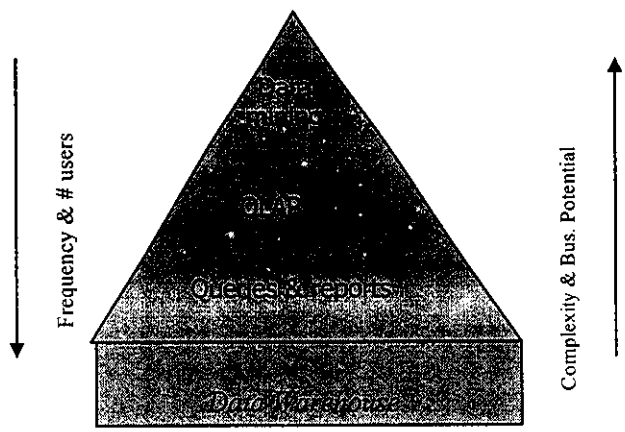


Figure. 1.1 BI pyramid

1.1.1 Need for Business Intelligence[3]

- To gain sustainable competitive advantage.

- Many technologies are available under BI through which data can be analyzed easily.

  --- Critical and strategic decision making.

- BI assists in

  ---Customer profiling, market basket analysis, customer contact analysis, market segmentation and inventory movement.

The following figure 1.2 depicts why many companies invest lot of money in Business intelligence.

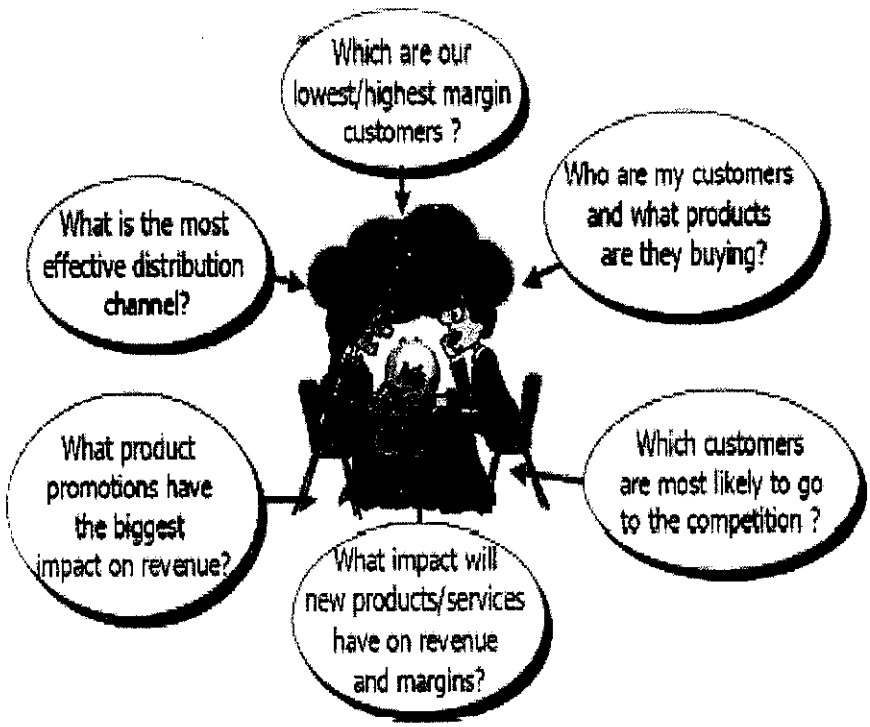


Figure.1.2 Need for Business Intelligence

1.1.2 Application areas of Business Intelligence

Industries that are known to use BI are data rich industries, such as[3]:

- Consumer goods

- Retailing industry
- Financial services
- Transport

Departments that are known to benefit most from BI are:

- (Database) Marketing
- Sales
- Finance
- IT(especially the Web) and the
- Higher Management

End-users:

- All types of end-users can use BI tools.
- End-users with different levels of expertise can apply BI applications to different levels of knowledge.
- With BI-tools it is possible to carry out analyses and reports on virtually all thinkable aspects of the underlying business, as long as the data about this business come in large amounts and are stored in a DW.

1.1.3 Main problems in data analysis

- Too much of data need to be analyzed
- Presence of bad data(redundant,missing,error prone data)
- Drowning in data
- Difficulty in turning data into actionable information.
  Thus the data available is
    - Weak ,difficult to detect patterns,confusing and does not indicate true change.
- Data analysis is complex in BI compared to querying a database because in BI data is extracted from multiple disparate sources.

The business environment changed in a way that led to the need of BI[3]:

- Increased speed of business
- Information overload
- Increased globalization
- Increased complexity and dynamics of internal processes and of the environment
- Speed of technological changes

1.2 COMPONENTS OF BUSINESS INTELLIGENCE

The following figure 1.3 shows the various components of Business Intelligence [4].

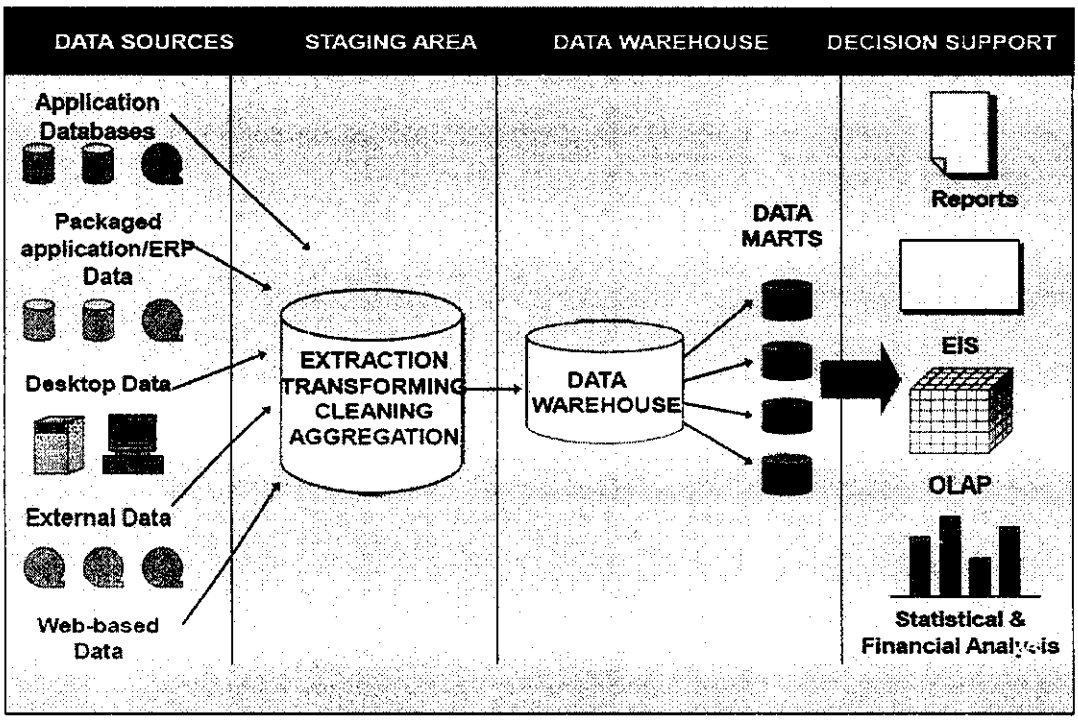


Figure. 1.3 Components of Business Intelligence

1.2.1 Sources of Data

Production data

This category of data comes from the various operational systems of enterprise. Based on the information requirements in the data warehouse, segments of data are chosen from the different operational sources.

Internal data

It includes documents, customer profiles and departmental databases. The IT department must work with the user department to gather internal data.

Archived data

Operational systems are primarily intended to run the current business. In every operational system, periodically take old data and store it in archived files.

External data

It refers to data of the competitors. They use standard values of financial indicators for their business to check their performance.

1.2.2 Data Warehousing

The data warehouse is at the center of the business intelligence Environment.The data warehouse represents the single version of truth for the Corporation and holds data at a granular level. A data warehouse is a subject-Oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process.

The reasons for building DWs were:

- Amount of data grew exponentially
- Amount of knowledge grew relatively slower
- Decision-time decreased
- Resulting in a need for bigger, more consistent and cheaper storing methods

The benefits of using a DW are:

- It supports management decisions

- It provides a cost-effective usage of IT resources
    - Cost reduction due to faster, automated search(one push on the button)
- It is linked to other sources
- It provides access to reliable, consistent and high-quality information
- It improves the information stream in an organization
- 'The unique truth':
    - everyone measures with the same units and
    - everyone has the exactly the same information
- Availability and access
    - Data are always available
    - Authority(different views)

1.2.3 Data mart

It is a subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart. It is data of a single business process. It is technology optimal for data access and analysis[3].

1.2.4 ETL(Extract-Transform-Load) process

The 3 major processes of the data warehouse are extract (data from the operational systems and bring it to the data warehouse), transform (the data into internal format and structure of the data warehouse) and load (cleanse data is put into the data warehouse)[1].

The three processes from extraction through loading often referred collectively as *Data Staging.*

Extract

Some of the data elements in the operational database can be reasonably be expected to be useful in the decision making, but others are of less value for that purpose. For this reason, it is necessary to extract the relevant data from the operational database before bringing into the data warehouse. Many commercial tools are available to help with the extraction process. **Data Junction** is one of the commercial products. The user

of one of these tools typically has an easy-to-use windowed interface by which to specify the following:

(i) Which files and tables are to be accessed in the source database?

(ii) Which fields are to be extracted from them? This is often done internally by SQL Select statement.

(iii) What are those to be called in the resulting database?

(iv) What is the target machine and database format of the output?

(v) On what schedule should the extraction process be repeated?

Transform

The operational databases developed can be based on any set of priorities, which keeps changing with the requirements. Therefore those who develop data warehouse based on these databases are typically faced with inconsistency among their data sources. Transformation process deals with rectifying any inconsistency (if any).

One of the most common transformation issues is 'Attribute Naming Inconsistency'. It is common for the given data element to be referred to by different data names in different databases. Employee Name may be EMP_NAME in one database, ENAME in the other. Thus one set of Data Names are picked and used consistently in the data warehouse. Once all the data elements have right names, they must be converted to common formats. The conversion may encompass the following:

(i) Characters must be converted ASCII to EBCDIC or vise versa.

(ii) Mixed Text may be converted to all uppercase for consistency.

(iii) Numerical data must be converted in to a common format.

(iv) Data Format has to be standardized.

(v) Measurement may have to convert. (Rs/ $)

(vi) Coded data (Male/ Female, M/F) must be converted into a common format.

All these transformation activities are automated and many commercial products are available to perform the tasks. .

Loading

Loading often implies physical movement of the data from the computer(s) storing the source database(s) to that which will store the data warehouse database, assuming it is different. This takes place immediately after the extraction phase. The most common channel for data movement is a high-speed communication link. Ex: Oracle Warehouse Builder is the API from Oracle, which provides the features to perform the ETL task on Oracle Data Warehouse.

1.2.5 Decision Support system

A decision support system (DSS) is a computer program application that analyzes business data and presents it so that users can make business decisions more easily[5]. . DSS is focused on the lower and middle management and makes it possible to look at and analyze data in different ways. Executive Information System(EIS) is the precursor focused on the higher management.

The general characteristics of a DSS which we also find within BI are:

- Support business decision making
- Based on data (from operational systems and/or a DW)

1.2.6 Online Analytical Processing(OLAP)

OLAP has proven to be the most extensive field in BI. OLAP is the concept that most authors have ventured to write about and most BI-companies claim to have in their portfolio of products and service. Without OLAP it would be quite a job to extract the right information using just regular Structured Query Language (SQL)-queries.

Eg. Three dimensional view of sales data i.e. sales by region, sales by product, sales by customer.

Typical OLAP operations are

- **Roll up (drill-up)**

  Summarize data by climbing up hierarchy or by dimension

  Reduction

- **Drill down (roll down)**

  Reverse of roll-up from higher level summary to lower level summary or detailed data, or introducing new dimensions

- **Slice and dice**

  Project and select

- **Pivot (rotate)**

  Reorient the cube, visualization, 3D to series of 2D planes.

1.2.7 Data Mining

Data mining is the process of extracting useful information from existing data[18].Knowledge discovery from hidden patterns supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools. This process is simply "learning from experience". It is a totally natural and routine part of every successful business. Data mining just helps you do it more quickly, accurately, and systematically.The methods used are more scalable, robust, reliable and intelligent.

Functionalities of data mining[18]

- **Characterization:**

  Data characterization is a summarization of general features of objects in a target class, and produces what is called characteristic rules.

- **Association analysis**:

  Association analysis is the discovery of what are commonly called *association rules*. It studies the frequency of items Occurring together in transactional databases, and based on a threshold called *support* and confidence.

- **Classification**:

  Classification analysis is the organization of data in given classes. Also known as *supervised classification*, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a *training set* where all objects are already associated with known class labels.

- **Prediction**:

  Prediction has attracted considerable attention given the Potential implications of successful forecasting in a business context.

- **Clustering**

  Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes.

- **Outlier analysis**

  Outliers are data elements that cannot be grouped in a given Class or cluster. Also known as *exceptions* or *surprises*, they are often very important to identify.

CHAPTER 2

LITERATURE SURVEY

2.1 EXISTING INTELIGENT DECISION MAKING SYSTEMS

2.1.1 Intelligent decision support system for a Car manufacturing Company

Zbigniew Michalewicz et al.[16] have developed an intelligent decision support system for a US based car manufacturing company.

Problem overview

US based car manufacturing car manufacturing company has more than 10,000 cars returned from leases or rental each year.

Challenges

- The main challenge is how to distribute the cars among the various auction sites around United States.
- The distribution must be such that the company must gain maximum profit.

The main decision making process in this involves,

- **Price prediction**

  Each car's initial price must be predicted based on parameters mileage,model,make,present condition,year it is manufactured,etc..,

- **Transportation issues**

  Calculating transportation costs involved and how many cars to ship at one time.

- Each day, the decision maker has to take atleast 100 cars and decide to which auction site it much be distributed.

Problem complexity

- **Volume effect:**
  - It's cheaper to ship a truckload of cars from one place to another than it is to ship one or a few cars at a time.
    - 1–6 cars cost $120 per car.

- 7–10 cars cost $95 per car.
- 11–14 cars cost $85 per car.

- **Scheduling:**
  - Every auction has a typical sales day.
  - The cars must be shipped to auction sites in advance atleast few days before sales day.

- The authors developed an intelligent decision support system that would solve all the above problems.
- The following are the building blocks of the intelligent system.
  - Prediction module
  - Optimization module
  - Adaptation module.

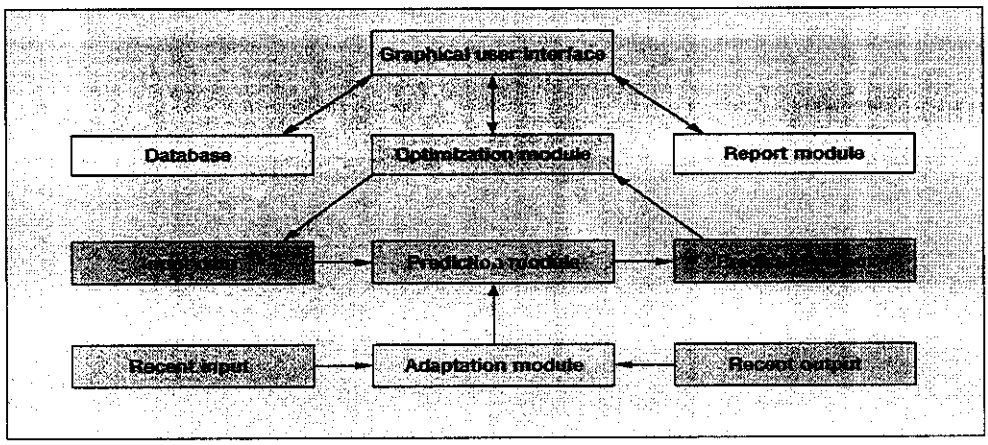


Figure. 2.1 IDSS Design for a Car manufacturing company

End Users of Car Manufacturing Company IDSS

- **Remarketing officer**
  - ➢ He must make use of the prize predicted by IDSS,can make changes if needed and update the prize of each car.
- **Transport officer in the Transport management department.**

- Responsible for shipping the cars to the destination auction site effectively.

- **Finance manager in finance management**
  - To set initial price for each car before shipping to the auction site

Prediction Module

Parameters:

- ***Default base prizing:***
  - Updated blackbook data
- ***Zip-code-based make/model adjustment:***
  - Uses historic sales records to adjust prize according to specific model or make.
- ***Mileage adjustment:***
  - Mileages are broken in to various categories like 0–10,000, 10,001–30,000,etc..,
- ***Seasonal prizing***

Optimization Module

- ***Need:***
  - To recommend the best distribution of cars to auction sites.
  - It uses evolutionary algorithms for performing optimization.
  - Each car is scheduled for transport to the closest auction site.
- ***What optimization module actually does?***
  - Calculates the expected total of all sales
  - Modifies the total to include price depreciation and volume effect
  - Subtracts costs related to the given distribution.

Adaptation Module

- ***Need:***
  - To handle the rapidly changing environment.

- This is accomplished by by slightly altering the learned relationship between input and output as needed from recent history.

➢ ***Functions:***

- Using recent input and output, It trains the prediction module.
- Sales prize prediction:
  - **Recent input**: the auction sites, dates, and car information constitute the recent input.
  - **Recent output:** Actual prize.

Implementation

➢ Graphical user interface(GUI).

➢ A database to store information.

➢ After processing,The decisions are returned in user understandable format like reports,tables,etc.,

➢ **Starting Point:**

- Client's inventory system

➢ **Methodology:**

- optimization process

➢ **Target:**

- IDSS to be used by remarketing officer

Disadvantages of this work

➢ The authors did not clearly mention how the data needed for the IDSS is extracted and stored.

➢ They failed to mention how much the system is scalable.

2.1.2 Business Intelligence in Telecom Industry

J. O'Brien et al.[10](2008) have conducted a research as how business intelligence helps in Telecom industry. They have developed an intelligent decision support(IDSS) system which can be used in the Telecom industry.

Problem Overview

- In today's extremely challenging business environment, the telecommunications industry is measuring their success by the size and growth of their profit margins.
- Many critical telecommunications functions rely on fast, complex analysis of CDR (Call detail Records) data

Challenges

- Storing and accessing all of this data is highly valuable but technically challenging.
- Improving profitability in terms of achieving a competitive advantage through effective decision making.

Role of BI in telecom

- **CRM** - analyzing **behavioral data** to optimally target services.
- **Billing** - ensuring complete and accurate billing
- **Revenue Assurance** - modeling call behavior.

BI solution

- Summarize or filter the data for analysis
- Create a massive, complex and often custom **CDR warehouse** to analyze CDR information.

End Users

- Business Analysts:

  Individuals whose primary role is to track, understand, and manage information in order to pass it on to others in the organization .
- Knowledge workers:

  Role of these individuals is to make decisions and run the business .
- Chief Telecommunication Officer(CTO)
- Managers

Development of IDSS System

Starting Point

- **Computerization** has already taken place in the Telecom industry.
- There is a need for **analyzing** huge volume of CDR.

- To understand the special requirements of telecommunication decision support system.

Methodology

- Thorough and ongoing research into Telecom organization's informational needs (present and future).
- Telecom Decision making system is built using the techniques of data mining and artificial intelligence.

Target

- To help the end user in gathering and analyzing data effectively.
- To assist the end users in decision making.

Advantages of the developed system

- Improved revenue and profits by acquiring and retaining high-value customers.
- Reduce costs through more effective and responsive management;
- Improved customer loyalty through improved quality and superior service delivery.

2.2 EXISTING ETL MANAGEMENT SYSTEMS

2.2.1 Problems in each stage of ETL

Panos Vassiliadis et al.[28] have identified various problems in each stage of Extraction,Transformation and Loading which affects the ETL management which are described as follows:

The key factors underlying the main problems of ETL processes [28] are:

- **Vastness** of data volumes.
- **Quality problems**, since data is not always clean and has to be cleansed.
- **Performance**, since the whole process has to take place within a specific time window.
- **Evolution** of the sources and the data warehouse can eventually lead, even to daily maintenance operations.

Extraction

The extraction step is conceptually the simplest task of all, with the goal of identifying the correct subset of source data that has to be submitted to the ETL workflow for further processing. As with the rest of the ETL process, extraction also takes place at idle times of the source system - typically at night. Practically, the task is of considerable difficulty, due to two technical constraints:

- The source must suffer minimum overhead during the extraction, since other administrative activities also take place during that period and
- Both for technical and political reasons, administrators are quite reluctant to accept major interventions to their system's configuration, therefore, there must be minimum interference with the software configuration at the source side.

Transformation

Depending on the application and the tool used, ETL processes may contain a plethora of transformations. In general, the transformation and cleaning tasks deal with classes of conflicts and problems that can be distinguished in two levels [28]: the schema and the instance level.

Schema-level problems. The main problems with respect to the schema level are (a) naming conflicts, where the same name is used for different objects (homonyms) or different names are used for the same object(synonyms).Structural conflicts, where one must deal with different representations of the same object in different sources, or converting data types between sources and the warehouse.

Record-level problems. The most typical problems at the record level concern duplicated or contradicting records. Furthermore, consistency problems concerning the granularity or timeliness of data occur, since the designer is faced with the problem of integrating data sets with different aggregation levels (e.g., sales per day vs. sales per year) or reference to different points in time (e.g., current sales as of yesterday for a certain source vs. as of last month for another source).

Value-level problems. Finally, numerous low-level technical problems may be met in different ETL scenarios.To mention a few, there may exist problems in applying format masks, like for example, different value representations (e.g., for sex: `Male', `M', `1'), or different interpretation of the values (e.g., date formats: American `mm/dd/yy' vs. European `dd/mm/yy').

Loading

The end of the source records' journey through the ETL workfow comes with their loading to the appropriate table. A typical dilemma faced by inexperienced developers concerns the choice between bulk loading data through a DBMS-specific utility or inserting data as a sequence of rows.

2.2.2 Universal ETL Management Platform

He Li et al. [18] have proposed a universal ETL management platform for telecom industry. Considering the characteristic of massive data in telecom system, by means of thread pool technologies, this ETL management platform improves the multithreading technology to raise the ETL efficiency and resource utilization. At last, the system results prove this solution is efficient and valid.Aiming at the background of developing the decision support system of Hebei Province Netcom Corporation,the authors have proposed a universal ETL system architecture based on distributed, heterogeneous data sources, a flexible solution to deal with the problems of massive data, data sources complexity and data import efficiency. This system not only realized extracting, transforming and loading Meta data from heterogeneous data sources but also allowed users to monitor and manage the whole data process, and deal with abnormal exception to make the system more robust and reliable. This ETL management platform has been applied into the decision support system of Hebei Province Netcom Corporation.

System Design

The main goal of the system is to:

- Extract useful data from different business systems such as telephone billing systems, customer service subsystems to and deal with extract useful data and clean up to ensure data correctness.

- To achieve the real-time monitoring of the whole ETL process exceptions.

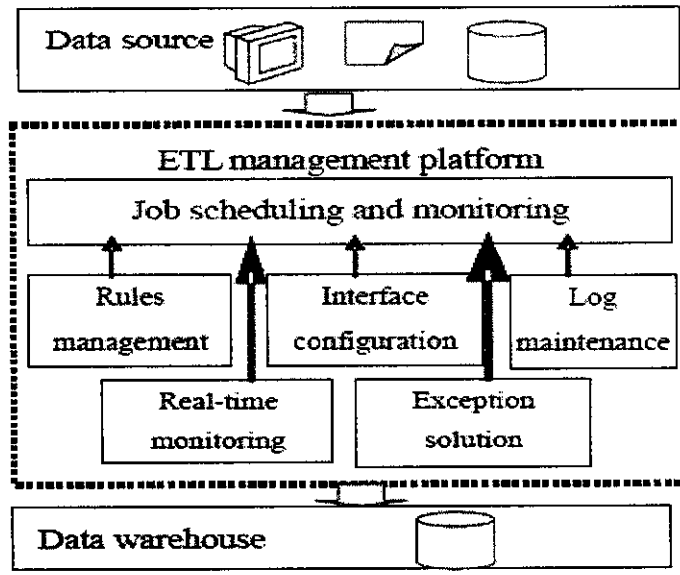


Figure.2.2 Universal ETL platform System architecture

- **Rules management module**

  This part function is to describe the data extraction, conversion, loading rules and provide a common rule description way so as to make this platform more expansible, this part is the core of the ETL platform.

- **Log maintenance module**

  The main task is to record and manage important information during the process of system running, providing the basis for real-time monitoring module.

- **Real-time monitoring and exception handling module**

  The main function is to deal with exception coming from the running system,and implement system recovery so as to improve the system reliability.

Drawbacks of the system

- More complexities are involved in each stage
- Many works have to be done to bring them in to a unified platform.

CHAPTER 3

SYSTEM ANALYSIS

3.1 DRAWBACKS OF THE EXISTING SYSTEMS

Many retail trade enterprises[25] have accumulated huge amount of data from OLTP (Online Transactional Processing) systems, which might not have been used for many years.How to use use the data to provide decision support for managers is still a challenging one. The end users are unaware of the internal structure of the databases, difficult to generate their own queries and also manual analysis is also complex as data which needs to be analysed is huge. In domestic retail industry, there has been tremendous growth mainly due to the computerization of transactional activities.As a result of this automation, lot of historical data which has been accumulated.Existing systems which are developed only provide the users to analyze the data in traditional static formats.These reports can neither be viewed from different perspectives at different times nor can they provide critical insight for retailers. The following are some of the available tools which can be used to build business intelligence systems:-

- Oracle's warehouse builder
- IBM's warehouse manager
- SAP Business Objects
- Informatica's datastage
- Pentaho Business Intelligence suite
- PaloETL
- SAS Business Intelligence
- Microstrategy

These tools automate entirely the process of building a warehouse and decision making.But these tools are very costly and suits only large organization.Small organizations find it very costly and complex to understand.Moreover these tools are not domain specific.Business intelligence system built for a banking sector cannot be used by a retail organization.

There is a wide variety of reporting requirements, and whether to buy or build a reporting tool for business intelligence needs is also heavily dependent on the type of requirements.

With BI tools,the system can be built quickly,shielding the complex task of coding,improving the speed and reducing the degree of difficulty,but the main shortcoming is the lack of flexibility. The following are shortcomings of existing BI platform

- The limited available user interfaces, not satisfying the users' secondary development request;
- High prices and development costs;
- Different national conditions. Although these tools have a certain amount of commonality, they still can not meet the needs of domestic Users.

After a detailed study on the existing business intelligence system,This project deals with designing a retail business intelligence system with low cost and in a short period of time which can be used by a small retail organization for their decision making.

3.2 PROPOSED SYSTEM

In recent years,there has been tremendous growth in the retail industry due to automation and universal standardizaton of business rules.This has led to the introduction of many Point of Sale (POS) solutions,Customer Relationship managent (CRM),Enterprise Resource Planning(ERP) packages and son on in the market.In this proposed project,the features of retail Point of Sale Solution is extended to manageable Business Intelligence System to assist the retail sellers in future planning and decision making.

In general,Retail is the most diverse industry handling widest range of products from equally large numbers of suppliers and, naturally, the highest number of customers. Also, the retail market trends change most frequently than any other industry. Keeping an eye on all the operations in retail business is crucial for seamless profitable corporate performance. Understanding customer requirements and offering them what they want and still maintaining profitability requires highly analyzed information on management part to deal with ever changing market conditions quickly to beat competition.

Some of the areas where BI Solutions can be applied to retail:

- **Sales and profitability analysis** - Sales and profitability analysis enables to track product demands, trends, & opportunities in sales. Helps you to optimize strategy for marketing promotions & planning according to periodic phase.

- **Store Operations Analysis** - Store Operations Analysis facilitates to monitor multiple store performance, efficiency of sales management, and operations across different Stores. Helps in empower management with information-based execution and budgeting.

- **Customer Analysis** - Customer Analysis evaluates profitability by customer segmentation and develop continuing relationships with customers. Helps to plan more strategic campaigns to maximize customer acquisition and retention.

Other Analysis Areas:

- Merchandise Management
- Inventory Management
- Supplier Performance Management
- Marketing and E-commerce Analysis
- Market Basket Analysis
- Category Management
- Brand & Marketing Research
- Market Share Analysis

After a detailed understanding of need of business intelligence in retail,a simple system is proposed in which a small portion of the retail analysis area is concentrated.The system can further extended in future to handle more complex requirements.The

following figure 3.1 is the diagrammatic representation of the proposed system.The detailed description of the proposed architecture is given in the upcoming chapter 6 under the title project description.

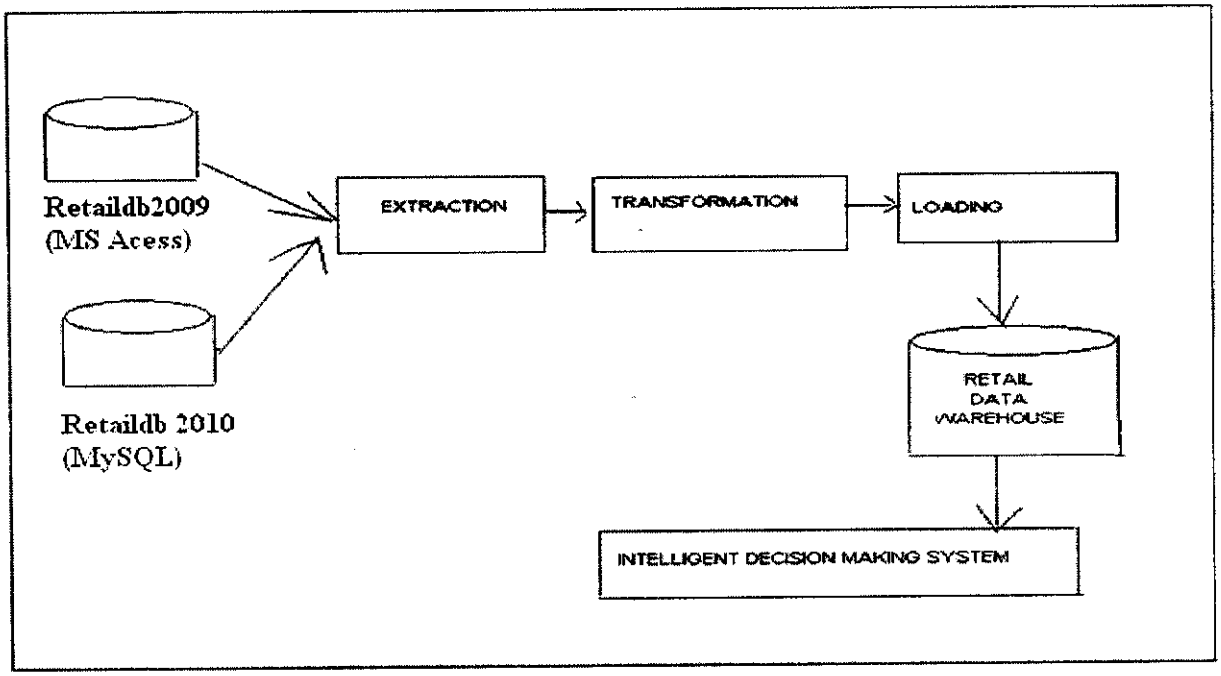


Figure. 3.1 Proposed System Architecture

Methodology for the proposed system development

Starting Point

- Computerization has been already taken place in the retail industry.
- Raw data from a retail based billing software is taken as input for the proposed system
- There is a need for analyzing huge volume of retail data like product details, customer details,sales details and so on.
- To understand the special requirements of retail decision support system.

Methodology

- Thorough understanding of various strategic questions posed by the retail sellers for their business decision making.
- Building a consistent data warehouse through the three step process of Etraction,Tansformation and Loading(ETL).

- Development of intelligent decision making system using the techniques of statistical models to discover the hidden retail related decisions from the data warehouse

Target

- To help the end user in gathering and analyzing data effectively through friendly user interface.
- To assist the end users in decision making.

Expected Advantages of the system

- To Improve organization sales by acquiring and retaining high-value customers;
- Reduce costs through more effective and responsive management;
- The end users must be provided with future planning feature like inventory optimization,sales forecasting and so on.

CHAPTER 4

SYSTEM SPECIFICATION

4.1 HARDWARE REQUIREMENTS

Processor	: Pentium III and above
Clock speed	: 550MHz
Hard Disk	: 20GB
RAM	: 128MB or above
Cache Memory	: 512KB
Operating System	: any Windows version,Linux,Unix
Monitor	: Color Monitor
Keyboard	: 104Keys
Mouse	: 3Buttons

4.2 SOFTWARE REQUIREMENTS

IDE	: Eclipse with JDK 1.6
Database	: MS Access,MySQL 5.0
Database drivers	: ODBC driver,MySQL-JDBC driver
Java plug-in	: JFreechart
Other softwares	: R-Software(Statistical package)

4.3 SOFTWARE DESCRIPTION

- **Eclipse IDE**

  It is a project aiming to provide a universal toolset for development. Open Source *IDE*, mostly provided in Java, but the development language is platform independent. The Eclipse IDE for Java Developers provides superior Java editing with validation, incremental compilation, cross-referencing, code assist;an XML EditorIt provides improved debugging support, including hyperlink Stepping and also assists in task-focused development.

- **Java Development Kit(JDK 1.6)**

  The JDK is a superset of the JRE, and contains everything that is in the JRE, plus tools such as the compilers and debuggers necessary for developing applets and applications. The Java Runtime Environment (JRE) provides the libraries, the Java Virtual Machine, and other components to run applets and applications written in the Java programming language.

- **MySQL 5.0**

  MySQL is a relational database management system (RDBMS) that runs as a server providing multi-user access to a number of databases. It is a popular choice of database for use in web applications, and can be used as central component for open source project development.

- **ODBC (Open database Connectivity)driver**

  Open Database Connectivity (ODBC) is Microsoft's strategic interface for accessing data in a heterogeneous environment of relational and non- relational database management systems. Based on the Call Level Interface specification of the SQL Access Group, ODBC provides an open, vendor- neutral way of accessing data stored in a variety of proprietary personal computer, minicomputer, and mainframe databases.

- **MySQL –JDBC Driver**

The Java Database Connectivity interfaces allows developers to write applications that can be used with different databases with a minimum of porting effort. Once a driver for a given server engine is installed, JDBC applications can communicate with any server of that type. By using MySQL-JDBC Connector Java programs can access MySQL databases.

- **JFree chart**

JFreeChart is a Java chart library that makes it easy for developers to display professional quality charts in their applications. It is particularly effective for when a user needs to regenerate graphs that change on a frequent basis.As this Jfree chart is built based on java,this can be integrated with java programs.

- **R-Software**

R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

CHAPTER 5

PROJECT DESCRIPTION

5.1 PROBLEM DEFINITION

A retail organization contains lot of operational data collected from various sources.The data available is huge and not in unique format.It is difficult for a retail seller to manually analyze these data and make decisions. The problem here is to extract the operational data, tranform them into a consistent format and load them into the target retail data warehouse.Retail sellers are totally unaware of the internal structure of the data and also how to write queries to view the internal data.A system must be designed which provides a user interactive interface to generate decision making reports. An Intelligent decision making system must be designed for retail related decision making and to assist the retail sellers in future planning.

5.2 PROJECT OVERVIEW

In this project,ETL system[26] and intelligent decision making system are merged together.The source for the project is from a retail based day to day billing software. The entire project is divided in to two phases as follows:-

- **Phase I:** Building a sample retail based data warehouse.
- **Phase II:** Design of intelligent decision making system.

5.2.1 BUILDING A RETAIL BASED DATA WAREHOUSE

As per definition,"A Data warehouse is a subject-oriented,integrated,non-volatile and time variant collection of data in support of management's decision making process".The data in data warehouse are from OLTP(Online Transactional Processing) database.In this project the OLTP database of a ASSET 2005,a retail billing software.In this phase,A retail data warehouse is built after Extraction,Transformation and Loading.

- **Billing software**

ASSET 2005 - Retail Software Solution suitable for all kind of retailers and distributors. Covers Value Added Tax(VAT), billing , Sales , Purchase inventory , payments , receipts ,etc., This software provides DOS mode printing and Windows front end gives excellent performance. Integrated with other popular accounting software like Tally . Customised to suit Hardware shops, paint shops,electrical shops, automobile shops, sanitary and fittings dealers,Jewellery, Textile shops,Hotel & Sweet Shops etc., It provides features for exporting data into Excel sheet format. E -returns for VAT are taken easily.

The 30 day trial version of this software is downloaded from retailplus website The databases for this project is downloaded from this software.

- **About the data sources**

The database contains the transactional data of a retail based hardware and paint shop for two different years namely 2009,2010.The databases are in various formats .This is taken as input to the project.The database which contain 2009 transaction details is in MS Access format and other one is MySQL format.Both the databases contains entities like customer(customerid,name,address,phone),product(id,name,rateid,supplierid,etc),supplier billing,sales,etc.The two source databases contain 120 transaction tables each and occupies nearly 15 MB space.After loading into the target retail data warehouse,the fact and dimension tables have to be designed.Dimension tables which are collection of attributes and fact tables are measure of facts or dimensions.

5.2.2 DESIGNING INTELLIGENT DECISION MAKING SYSTEM

The intelligent decision making system built in this project covers the following business intelligence analysis areas other than generating traditional reports based on complex queries.

- **Customer segmentation using clustering algorithm**

Customer segmentation is dividing customers into groups that are relevant for performing sales and marketing activities.There are many clustering data mining techniques available for segmentation.Clustering data mining divides data objects into several groups or clusters automatically so that the objects in the same group have high

similarties, and the objects in different groups have big differences.In this project,customer segmentation is performed based on retail sales data using k-means algorithm.In k-means,every group is represented by the average of each group,assign each remaining object to one of the k groups to which the object has the shortest distance.Calculate the average of each of each group,and regroup them until the groups are no longer changed.

- **Sales Forecasting based on ARIMA model**

Sales forecasting is the prediction of the future sales of products over a specific period of time based on past sales performance.It is based on time-series analysis. According to Hill *et al.*[11],The ARIMA procedure analyzes and forecasts equally spaced univariate time series data, transfer function data, and intervention data using the **A**uto**R**egressive **I**ntegrated **M**oving-**A**verage (ARIMA) or autoregressive moving-average(ARMA) model. An ARIMA model predicts a value in a response time series as a linear combination of its own past values, past errors (also called shocks or innovations), and current and past values of other time series.

In **autoregressive process,**Most time series consist of elements that are serially dependent in the sense that can estimate a coefficient or a set of coefficients that describe consecutive elements of the series from specific, time-lagged (previous) elements. This can be summarized in the equation:

$$x_t = \xi + \phi_1 * x_{(t-1)} + \phi_2 * x_{(t-2)} + \phi_3 * x_{(t-3)} + \ldots + \varepsilon \quad \textbf{(5.1)}$$

Where,

ξ is a constant (intercept), and
ϕ_1, ϕ_2, ϕ_3 are the autoregressive model parameters.

The **Moving average process** which is independent from the autoregressive process, each element in the series can also be affected by the past error (or random shock) that cannot be accounted for by the autoregressive component, that is:

$$\varepsilon = \mu + \varepsilon_t - \theta_1 * \varepsilon_{(t-1)} - \theta_2 * \varepsilon_{(t-2)} - \theta_3 * \varepsilon_{(t-3)} - \ldots \quad \textbf{(5.2)}$$

Where,

μ is a constant, and

$\theta_1, \theta_2, \theta_3$ are the moving average model parameters

The analysis performed by ARIMA is divided into three stages, corresponding to the stages described by Box and Jenkins[15].The IDENTIFY, ESTIMATE, and FORECAST statements perform these three stages, which are summarized below.

i. In the **identification stage**, the IDENTIFY statement is used to specify the response series and identify candidate ARIMA models for it. The IDENTIFY statement reads time series that are to be used in later statements, possibly differencing them, and computes autocorrelations, inverse autocorrelations, partial autocorrelations, and cross correlations.

ii. In the **estimation stage**, the ESTIMATE statement is used to specify the ARIMA model to fit to the variable specified in the previous IDENTIFY statement, and to estimate the parameters of that model.

iii. In the **forecasting stage** , the FORECAST statement is used to forecast future values of the time series and to generate confidence intervals for these forecasts from the ARIMA model produced by the preceding ESTIMATE statement.

- **Inventory Optimization**

In general, Inventory optimization (IO) includes a range of planning functionality that uses analytics and modeling to achieve the right balance of stock and meet service levels. Inventory optimization requires detailed historic data input .Inventory optimization is a key player as organizations search for a flexible supply chain and improved forecast accuracy, especially as increased uncertainty makes it harder to control inventory correctly. In this project,reorder quantity for each product is calculated based on the history of stock.The reorder point ("ROP") is the level of inventory when an order should be made with suppliers to bring the inventory up. method of calculating reorder level involves the calculation of usage rate per day, lead time which is the amount of time between placing an order and receiving the goods and the safety stock level expressed in terms of several days' sales.

CHAPTER 6
IMPLEMENTATION

6.1 MODULES DESCRIPTION

The main aim in this phase is to build a retail based data warehouse which is the input for the decision making system. The proposed BI system is developed using Java platform.Operational data is converted in to a consistent format using the three step process of Extraction (E), Transformation (T) and Loading (L).The following are the modules in the phase I which is building retail data warehouse.

i) Accessing heterogeneous data sources

The input data source is taken from retail based billing software. Sources are two heterogeneous data sources. The datasource1 is in MS Access format which contains transactions for the year 2009.The data source2 is in MySQL format which contains transactions for the year 2010.The end target data warehouse format is MySQL.In this module, choice is made between the source data sources and the target data sources.

ii) Extraction module

Extraction involves selecting only the relevant databases, tables and set of attributes from the source system according to the end decision making perspective. In this proposed architecture, the end target data warehouse is built for product, customer and sales analysis. The source databases contains nearly 120 tables each. Only 80 tables are extracted for end decision making. User can view the attributes in each table,drop unwanted tables,etc.This module is developed based using JSwing and JTable.

iii) Transformation module

The data in the source databases and the target database are compared. In this module, the data types in MS Access are transformed into MySQL format.This module takes input from config.txt which contains the source and target URL. The rows which contain null values are also filtered. The databases in different time periods are merged into a single database.

iv) Loading module

This module is designed for copying the tables from the source databases to the target data warehouse.Now, the target data warehouse built will be in consistent format and can be used for decision making. The target data warehouse now contains around 80 tables and nearly 1000 transaction details.In this module,user can view the contents of the table loaded into the target MySQL database.

After a data warehouse which contains retails sales data is built,next is to generate various sales related reports.The intelligent decision making system designed in this project provides a user interactive interface based on some of the strategic questions imposed by the retail sellers for their decision making.

- Who/Which are my top customers/ selling products?
- Which are my various customer(retail sellers) segments ?
- What will be my net sales next year with respect to current two year sales?

The following some of the modules in this phase.This involves designing complex queries for various sales reports generation,customer data clustering and sales forecasting using Autoregressive Integrated Moving Average (ARIMA) model.

i) Reports generation module

As there are nearly 80 tables in the data warehouse,retail sellers find it difficult to Generate complex queries to view the sales data in their required format.In this module two year data is taken and sales reports are generated combining various related tables.Eg.If the user needs to view the top products by sales.The system gets the top number,generates the query combining mst_product,det_sales and det_billing by aggregating the total sales from tables in two timelines.The following are some of the scenarios based on which queries are generated:-

- Top products by sales and profits
- Products purchased by customers
- Productwise sales
- Customerwise sales
- Supplierwise sales

ii)Segmentation Analysis module

In this module,customer(retail sellers) data is divided into various segments. Total sales of each retailer is calculated. Based on this total sales value the data is segmented into top,medium and low.Performance of each customer is evaluated based on their contribution to their net total sales in the previous two years.The data is visualized in barchart format using java based JFreechart package.

iii)Sales Forecasting module

This module computes the extrapolation forecasts of a univariate ARIMA model for a time series Y[t] (for t = 1, 2, ..., T). The user may specify a cut-off period K which implies that the ARIMA model is estimated based on Y[t] for t = 1, 2, ..., T-K and such that the extrapolation forecast F[t] for t = T-K+1, ..., T is computed and compared with the actual values that were dropped: various extrapolation forecast statistics are computed (MPE, RMSE, MAPE, ...). In addition, the following probabilities are computed: P(F[t]>Y[t-1]), P(F[t]>Y[t-s]), and P(F[t]>Y[T-K]).Input parameters and results obtained is explained in the forthcoming chapter.This module is developed using R-language which contains various packages for fitting statistical models.Forecast package is used to fit the ARIMA model.The net sales data for past 5 years is given as input and next three years net sales is forecasted.The following are the steps in this module.

1. Get the input data x and initial parameters of ARIMA model.
2. Calculate the length of the input data and store it in a variable.
3. Fit the input parameters to the model
4. Pass the fitted model as parameter to the predict function which is found in the forecast package.
5. Calculate the lower bound and upper bound for the predicted vaues.
6. Plot the graph for the forecasted data.
7. Display the results in table.

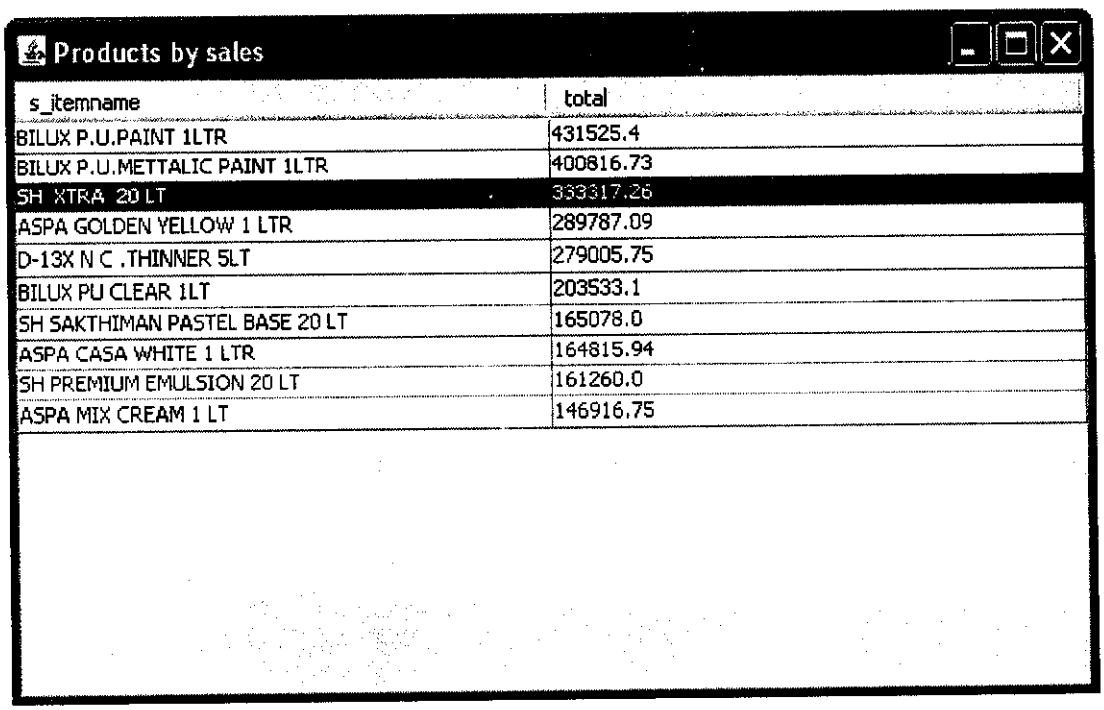
Figure.6.1 ARIMA model algorithm

iv)Inventory Optimization module

In this module, the reorder quantity is calculated for every product.The reorder quantity ensures that the stock of each product does not go below ReOrder Level (ROL).The user selects the product name from the list.The system calculated the average sales quantity and maximum sales quantity.The ROL quantity is calculated based on the
Supplier lead time.

6.2 RESULTS

The following are some of the sales related reports generated in the report generation module.Figure 6.2 shows the top products based on sales.



s_itemname	total
BILUX P.U.PAINT 1LTR	431525.4
BILUX P.U.METTALIC PAINT 1LTR	400816.73
SH XTRA 20 LT	333317.26
ASPA GOLDEN YELLOW 1 LTR	289787.09
D-13X N C .THINNER 5LT	279005.75
BILUX PU CLEAR 1LT	203533.1
SH SAKTHIMAN PASTEL BASE 20 LT	165078.0
ASPA CASA WHITE 1 LTR	164815.94
SH PREMIUM EMULSION 20 LT	161260.0
ASPA MIX CREAM 1 LT	146916.75

Figure.6.2 Top 10 products by sales report

The following report which is shown in figure 6.4 shows the products which are purchased by individual customers(retail sellers).

Customerwise products

s_name	s_itemname	totQty	Total	i_retprice	Avg
JOHNSON LIFTS ...	B WASTE 1KG	0.5	62.4	130.0	305.0
JOHNSON LIFTS ...	CLOTHEMERY M...	500.0	5280.0	11.0	7606.14
JOHNSON LIFTS ...	D-13X N.C.THIN...	2.0	142.4	80.0	305.0
JOHNSON LIFTS ...	ROBERT POLISH...	2.0	71.2	40.0	305.0
JOHNSON LIFTS ...	SANDPAPER 60	10.0	57.6	6.0	490.0
JOHNSON LIFTS ...	SANDPAPER 80	3.0	17.28	6.0	263.0
JOHNSON LIFTS ...	SH SUPERLAC A...	6.0	961.2	180.0	15579.0
JOHNSON LIFTS ...	SH SUPERLAC A...	53.0	4245.3	90.0	9852.5
JOHNSON LIFTS ...	SH SUPERLAC B...	480.0	66216.0	155.0	9787.09
JOHNSON LIFTS ...	SH SUPERLAC D...	1.0	80.1	90.0	263.0
JOHNSON LIFTS ...	SH SUPERLAC D...	3.0	413.85	155.0	376.5
JOHNSON LIFTS ...	SH SUPERLAC F...	81.0	6488.1	90.0	5396.5
JOHNSON LIFTS ...	SH SUPERLAC G...	142.0	6192.62	49.0	9237.67
JOHNSON LIFTS ...	SH UNI PRIMER ...	1.0	106.8	120.0	490.0
JOHNSON LIFTS ...	TURPENTINE 1 LT	525.0	21026.25	45.0	8172.92
JOHNSON LIFTS ...	WILLSON 545 B...	324.0	11534.4	40.0	11102.5

Figure.6.3 Retail seller wise products report

The following figure 6.5 shows the supplierwise report the total sales is summarized based on the two year data and the report is generated.

out - Notepad

File Edit Format View Help

SUPPLIERWISE SALES ANALYSIS

CUSTOMERNAME	QTY	Total value
ASIAN PAINTS LTD	1237.15	1315041.2
ASIAN PPG INDUSTRIES LTD	5850.9	1270350.8
SHALIMAR PAINTS LTD	1976.1	1028041.4

Figure.6.4 Supplier wise total sales report

The following figure Figure.6.6 shows performance analysis of the top customer..

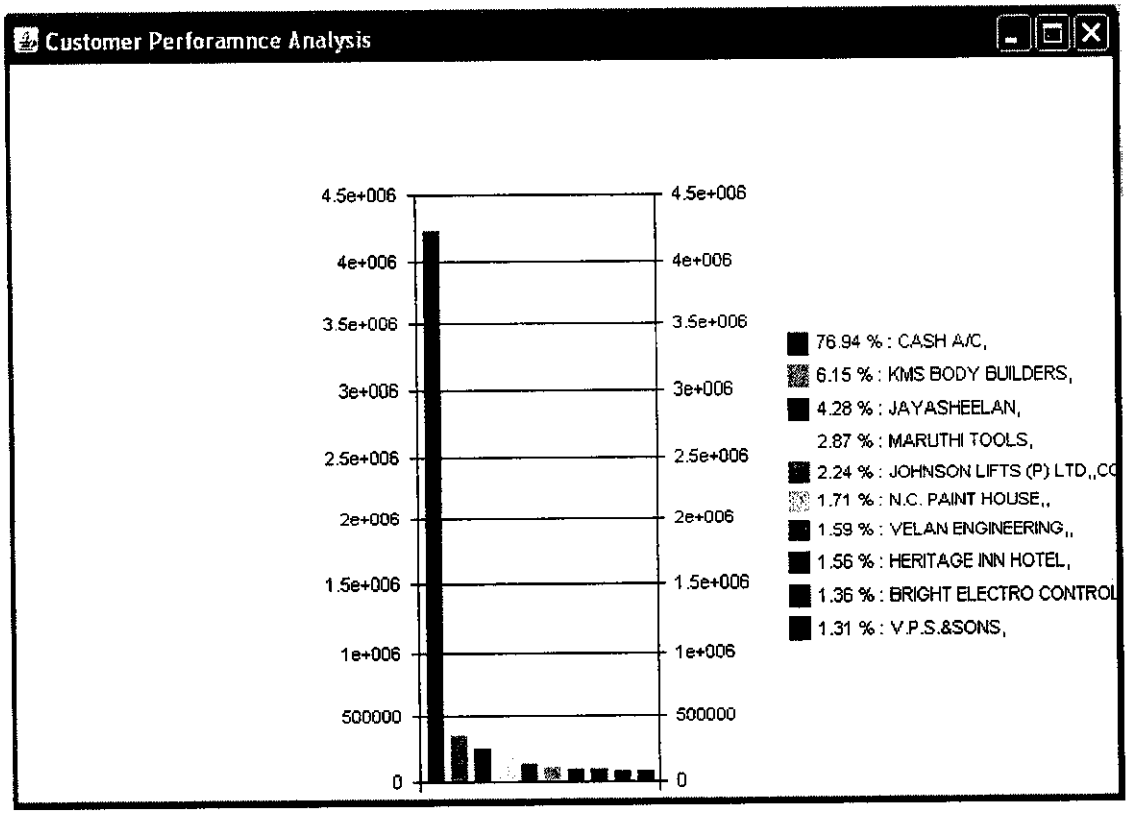


Figure.6.5 Top retail sellers Performance Analysis

Sales forecasting is done using ARIMA model which predicts the future values based on the input data,seasonal and non-seasonal parameters. **ARIMA models** - autoregressive, integrated, moving average models which express the data value at time t as a linear combination of previous data values and/or current and previous random shocks to the system.

Arima(x,order=c(p,d,q),seasonal=list(order(P,D,Q),period=S),include mean);

The above function shows the general form of fitting the parameters to the ARIMA model.In this project,sales data computed for the past 5 years is given as Input to the model.The following table Table 6.1 shows the Input parameters with range.

No.	Parameter name	Range
1	Testing period	(0-24)
2	Box -cox Lambda transformation parameter	(-2 to 1)
3	Degree of non-seasonal differencing(d)	(0,1,2)
4	Degree of seasonal differencing(D)	(0,1)
5	Seasonal Period(S)	(0,1,2,3,4,6,12)
6	Auto Regression(AR) order(p)	(0,1,2,3)
7	Moving Average(MA) order(q)	(0,1,2)
8	Seasonal Auto Regression(SAR) order(P)	(0,1,2)
9	Moving Average(MA) order(Q)	(0,1)
10	Include mean	(true/false)

Table: 6.1 ARIMA model parameters

For forecasting next three years sales,the following are given as input and the algorithm explained in the previous chapter.The following are the inputs given and the table 6.2 shows F(t) which is computed by the ARIMA model with confidence index 95%.

Input:

X← (4086789,4796493,5953621,6445670,7107664) [sales data],testing period=3,Lambda=1,d=0,D=0,S=1,p=2,q=2,P=2,Q=1,include mean=False

Time	Y[t]	F[t]	95% LB	95% UB	p-value (H0: Y[t] = F[t])	P(F[t]>Y[t-1])
1	4086789	-	-	-	-	-
2	4796493	-	-	-	-	-
3	5953621	5505581.863	5464686.457	5546477268	0	1
4	6445670	6214015.532	6121407.506	6306623.558	0	1
5	7107664	6921553.955	6759617.097	7083490.813	0.0121	1

Table: 6.2 Univariate ARIMA extrapolation forecast

The following Table 6.3 shows the Univariate ARIMA Extrapolation Forecast Performance.This is the difference between the actual data value at time t and the forecast that would have been made for that period given the data through time t-1. Included are the:

- **MSE** - the average or mean of the squared errors.
- **RMSE** - the square root of the MSE.
- **MAE** - the mean of the absolute values of the one-ahead errors.
- **MAPE** - the mean of the absolute values of the errors, as a percentage of the actual values. This is only calculated if all data values are greater than 0.
- **ME** - the average or mean of the errors.
- **MPE** - the mean error as a percentage of the actual values. This is only calculated if all data values are greater than 0.

The MSE, RMSE, MAE, and MAPE measure the magnitude of the forecast errors. Better models will show smaller values for these statistics.

Time	% S.E.	PE	MAPE	Sq.E	MSE	RMSE
3	0.0038	0.0814	0	2.00739E+13	0	0
4	0.0076	0.0373	0.0593	5.36639E+12	1.27202E+13	3566534.749
5	0.0119	0.0269	0.0485	3.46373E+12	9.63469E+12	3103979.768

Table 6.3 Univariate ARIMA Extrapolation Forecast Performance

The following graph Figure 6.7 shows the ARIMA forecast performance having nce Index 95% with time.

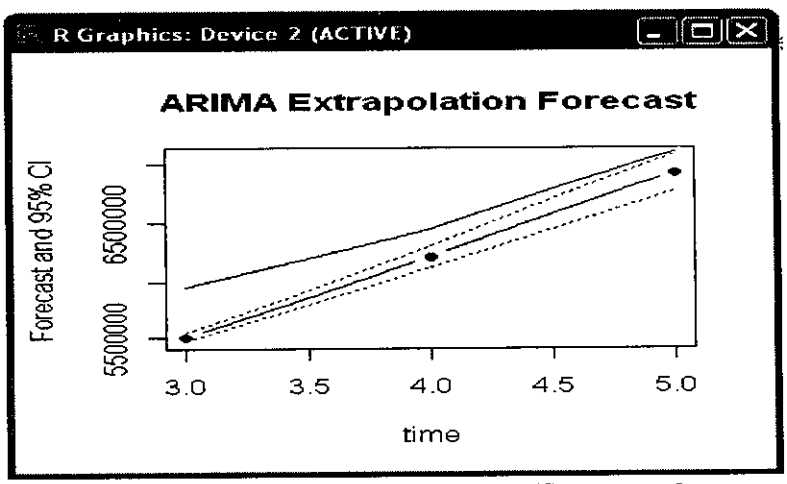


Figure.6.6 ARIMA Extrapolation Forecast plot

CHAPTER 7

CONCLUSIONS AND FUTURE ENHANCEMENTS

The retail based Business Intelligence system developed in this project is platform independent as it is developed using java platform and MySQL, both are open source software. Using the BI system implemented in this project,users can generate Reports which delivers regular, timely information and also they can author reports or queries to get specific details.Hence the BI system developed integrates ETL management system and the intelligent decision making system into single system .The system can be used by the retail seller for generating various sales related reports and analysis of sales data based on products,customers,suppliers,etc. This system is cost effective and can be used by any small retail stores.Sales forecasting is done based on ARIMA model and its model its performance is evaluated.Further enhancements can be made to the Business intelligence system by understanding various current issues.Complex ETL rules can be considered to build a datawarehouse.Real-time monitoring and updation of warehouse is still a challenging one.Various data mining techniques and artificial neural networks algorithms can be used to analyze the data.Further data visualization can be improved by presenting the data in the form of dashboards,scorecards,etc.

CHAPTER 8

APPENDIX

8.1 Source codes

 # Extraction module

```
import java.awt.*;
import java.awt.event.*;
import javax.swing.*;
import javax.swing.event.*;
import java.sql.*;
import java.util.*;
import java.io.*;


class Extraction extends JFrame implements ListSelectionListener,ActionListener,ItemListener {

    private JLabel tableLbl;
    private JComboBox tableBox;
    private JLabel attributesLbl;
    private JList  attributesList;
    private JTextField conditionField;
    private JLabel qTypeLbl;
    private JComboBox qTypeBox;
    private JLabel whereLbl;
    private JTextField whereField;
    private JLabel oprationLbl;
    private JComboBox oprationBox;
    private JLabel valueLbl;
    private JTextField valueField;
```

```
private JRadioButton andButton;
private JRadioButton orButton;
private JButton executeButton;
private JButton genrateButton;
private JButton refreshButton;
private JLabel queryLbl;
private JTextField queryField;
private JButton resetButton;
private JTable table;
private JScrollPane personScrollPane;
private JScrollPane attributeScrollPane;
private JLabel imageLbl;
private  JLabel bgLbl;
private JFileChooser chooser;
/private FileNameExtensionFilter filter;
private JButton dbButton;
private JLabel dbLbl;
private JTextField dbField;
private static String mdb;
private final int x = 110,  y = 50;
private int l = 150,  w = 20;
private int xInc = 100,  yInc = 40;

// data base variable
private    Connection con=null;
private    Statement st=null;
private    ResultSet rs=null;
private    DatabaseMetaData dbmd;
private    ResultSetMetaData rsmd;
final static String jdbcDriver = "sun.jdbc.odbc.JdbcOdbcDriver";
private    boolean start=false;
```

```
public static void main(String[] arg)
{
        Extraction obj=new Extraction();

}

public Extraction()
{

        super("Extraction");
        try
        {

            UIManager.setLookAndFeel(UIManager.
                  getSystemLookAndFeelClassName());
         }
         catch (Exception unused) {}
        setLayout(null);
         try{
           Class.forName(jdbcDriver);}
           catch(Exception errr){

           }
        setBounds(150, 150, 835, 551);
        addFields();
         resetFields();
        addImageAndBackGround();
        setDefaultCloseOperation(JFrame.EXIT_ON_CLOSE);
        setVisible(true);
   }
```

```
private void loadDataBase(String database)
{

                    try{
                    //connect to database
                    con=DriverManager.getConnection("jdbc:odbc:Driver={Microsoft
                     Driver (*.mdb)};DBQ="+database+";DriverID=22}");
                    //con=DriverManager.getConnection(jdbcURL);
                    st=con.createStatement();

                    dbmd=con.getMetaData();
                    }
                    catch(Exception eee){

                    JOptionPane.showMessageDialog(null, eee.toString(), "Error
                     Message2", JOptionPane.ERROR_MESSAGE);
                    }
    }       //end of database connection

  private void addFields()
  {

        dbLbl = new JLabel("Database :");
        dbField = new JTextField(10);
        dbLbl.setBounds(x, y-yInc, l, w);
        dbField.setBounds(x + xInc, y-yInc, l, w);
        dbField.setEditable(false);
        dbButton=new JButton("Browse");
        dbButton.setBounds(x+3*xInc,y-yInc,l-60,w);
        dbButton.addActionListener(this);
        add(dbLbl);
```

```
add(dbField);
add(dbButton);
tableLbl = new JLabel("TableName :");
tableBox = new JComboBox();
tableBox.addItemListener(this);
tableLbl.setBounds(x, y, l, w);
tableBox.setBounds(x + xInc, y, l, w);
add(tableLbl);
add(tableBox);
attributesList = new JList();
attributesList.setSelectionMode(ListSelectionModel.
   MULTIPLE_INTERVAL_SELECTION);
attributesList.addListSelectionListener(this);
attributesList.setSelectionBackground(Color.blue);
attributeScrollPane=new JScrollPane( attributesList);
attributeScrollPane.setBounds(x, y + 2*yInc, l, w*4);
attributeScrollPane.setAutoscrolls(true);
add(attributeScrollPane);

String[] qType={"Select Query","Update Query","Delete Query"};
      qTypeLbl = new JLabel("Query Type");
qTypeBox = new JComboBox(qType);
      qTypeBox.addItemListener(this);
qTypeLbl.setBounds(x+2*xInc, y +  yInc, l-50, w);
      qTypeBox.setBounds(x + 3*xInc, y + yInc, l, w);
add(qTypeLbl);
add(qTypeBox);
whereLbl = new JLabel("Where");
whereField = new JTextField(10);
whereLbl.setBounds(x+2*xInc, y + 2* yInc, l-50, w);
whereField.setBounds(x + 3*xInc, y + 2*yInc, l, w);
```

```
 add(whereLbl);
add(whereField);
String[] oprant={"=","!=",">",">=","<","<="};
oprationLbl = new JLabel("Operations");
oprationBox = new JComboBox(oprant);
oprationBox.addItemListener(this);
oprationLbl.setBounds(x+2*xInc, y + 3 * yInc, l, w);
oprationBox.setBounds(x + 3*xInc, y + 3 * yInc, l, w);
add(oprationLbl);
add(oprationBox);
valueLbl = new JLabel("Value");
valueField = new JTextField(10);
valueLbl.setBounds(x+2*xInc, y + 4 * yInc, l, w);
valueField.setBounds(x + 3*xInc, y + 4 * yInc, l, w);
add(valueLbl);
add(valueField);
 queryLbl = new JLabel("Query");
 queryField = new JTextField(10);
 queryLbl.setBounds(x, y + 6 * yInc, l, w);
 queryField.setBounds(x +xInc, y + 6 * yInc, l*2, w);
 add(queryLbl);
 add(queryField);
 refreshButton = new JButton("Refresh");
 refreshButton.addActionListener(this);
 refreshButton.setBounds(x+3*xInc , y , l - 60, w);
 add(refreshButton);
 executeButton = new JButton("Execute");
 executeButton.addActionListener(this);
 executeButton.setBounds(x+10+4*xInc , y + 6 * yInc, l - 60, w);
  add(executeButton);
  resetButton = new JButton("Reset");
```

```
        resetButton.addActionListener(this);
        resetButton.setBounds(x + 4* xInc, y + 5 * yInc, l - 60, w);
        add(resetButton);
        genrateButton = new JButton("Genrate Query");
        genrateButton.addActionListener(this);
        genrateButton.setBounds(x+3*xInc-60 , y + 5 * yInc, l - 20, w);
        add(genrateButton);
        attributesLbl = new JLabel("Attributes");
        attributesLbl.setBounds(x, y + yInc, l, w);
        add(attributesLbl);


    }

    private void addImageAndBackGround()
{


        imageLbl = new JLabel("", new ImageIcon(getClass().getResource(
                "imgs/query.PNG")), SwingConstants.LEFT);
        imageLbl.setBounds(x + 450, y-30, 256, 256);
        add(imageLbl);

        bgLbl = new JLabel("", new ImageIcon(getClass().getResource(
                "imgs/bground1.jpg")), SwingConstants.LEFT);
        bgLbl.setBounds(0, 0, 835, 551);
        add(bgLbl);

    }

private void displayResult()
```

```
{
        Vector rows = new Vector();
        Vector cols = new Vector();

 Vector<String> row;

try{
        rsmd=rs.getMetaData();
while (rs.next())


        {
                row = new Vector<String>();
               for(int i=1;i<=rsmd.getColumnCount();i++)
               switch(rsmd.getColumnType(i)){
                  case Types.VARCHAR:
                       row.add(rs.getString(i)); break;
                  case Types.INTEGER:
                       row.add(String.valueOf(rs.getInt(i))); break;
                 case Types.LONGVARCHAR:
                      row.add(rs.getString(i)); break;
                 case Types.DATE:
                      row.add(String.valueOf(rs.getDate(i))); break;
                default :
                System.out.println("Type is :"+rsmd.getColumnTypeName(i));
         }


   rows.add(row);
 }
```

```
            for(int col = 1; col <=rsmd.getColumnCount(); col++)
                    cols.add(rsmd.getColumnLabel(col));
    }
        catch(Exception eee){
                JOptionPane.showMessageDialog(null, eee.toString(), "Error
                 Message2", JOptionPane.ERROR_MESSAGE);
     }

    table=null;
    table = new JTable(rows, cols);
    table.setBackground(new Color(238, 238, 238));
        //table.setEnabled(false);
    personScrollPane = new JScrollPane(table);
    personScrollPane.setBounds(x, y+7*yInc, 500, 180);
    personScrollPane.setAutoscrolls(true);
    add(personScrollPane);

            remove(imageLbl);
      remove(bgLbl);
      addImageAndBackGround();
}
    public void valueChanged(ListSelectionEvent lse){
    whereField.setText((String)attributesList.getSelectedValue());
    }

    public void itemStateChanged(ItemEvent ie){

    if(ie.getSource()==tableBox){
    String table=(String)tableBox.getSelectedItem();
    Vector colName=new Vector();
```

```
try{
    ResultSet rs = st.executeQuery("SELECT * FROM "+ table);
    rsmd = rs.getMetaData();
    for(int col = 1; col <=rsmd.getColumnCount(); col++)
      colName.add(rsmd.getColumnLabel(col));
    attributesList.setListData(colName);


}
catch(Exception ee){
System.out.println(ee.toString());
}


}}

 public void actionPerformed(ActionEvent aEvent) {
 if(aEvent.getActionCommand().equals("Browse")){
  chooser=new JFileChooser();

       int returnVal=chooser.showOpenDialog(null);
 if(returnVal==JFileChooser.APPROVE_OPTION)
       {
       mdb=(String.format("%s",new File(chooser.getSelectedFile().getPath())));
       dbField.setText(String.format("%s",new
                    File(chooser.getSelectedFile().getName())));
       }
 }

 else if(aEvent.getActionCommand().equals("Refresh")){
 if(dbField.getText().length()==0){
```

```
JOptionPane.showMessageDialog(null, "Select DataBase First", "Error Message",
                JOptionPane.ERROR_MESSAGE);
}
else{
loadDataBase(mdb);
try{
tableBox.removeAllItems();
 String[] tableTypes = { "TABLE" };
 ResultSet allTables = dbmd.getTables(null,null,null,tableTypes);
while(allTables.next()) {
 String table_name = allTables.getString("TABLE_NAME");
  tableBox.addItem(table_name);



}}
catch(Exception ee){}
}}
else if(aEvent.getActionCommand().equals("Reset")){

resetFields();
}
else if(aEvent.getActionCommand().equals("Execute")){
int typeIndex=qTypeBox.getSelectedIndex();
if(typeIndex==0){
try{
rs=st.executeQuery(queryField.getText());
displayResult();
}
catch(Exception ee){}
}
else if(typeIndex==2){
```

```
try{
st.executeUpdate(queryField.getText());
}catch(Exception eee){}
}}
else if(aEvent.getActionCommand().equals("Genrate Query")){
 int typeIndex=qTypeBox.getSelectedIndex();
 String table=(String)tableBox.getSelectedItem();
  String opration=(String)oprationBox.getSelectedItem();
 if(typeIndex==0){
 if(whereField.getText().length()==0&&valueField.getText().length()==0)
  queryField.setText("Select * from "+table);
  else  if(whereField.getText().length()==0||valueField.getText().length()==0)
  JOptionPane.showMessageDialog(null, "Conditon is  Incorrect", "Error
         Message", JOptionPane.ERROR_MESSAGE);
 else
 queryField.setText("Select * from "+table+" where "+whereField.getText()+"
         "+opration+" "+valueField.getText());
 }
 else if(typeIndex==1){
  if(whereField.getText().length()==0&&valueField.getText().length()==0)
  JOptionPane.showMessageDialog(null, "Condition is required", "Error Message",
      JOptionPane.ERROR_MESSAGE);
 else  if(whereField.getText().length()==0||valueField.getText().length()==0)
  JOptionPane.showMessageDialog(null, "Conditon is  Incorrect", "Error
         Message", JOptionPane.ERROR_MESSAGE);
 else
  queryField.setText("Update "+table+" set columnName=columnValue,.. "+"
            where "+whereField.getText()+" "+opration+" "+valueField.getText());
 }
 else if(typeIndex==2){
 if(whereField.getText().length()==0&&valueField.getText().length()==0)
```

```
        queryField.setText("Delete  from "+table);
       else  if(whereField.getText().length()==0||valueField.getText().length()==0)
        JOptionPane.showMessageDialog(null, "Conditon is Incorrect", "Error Message",
              JOptionPane.ERROR_MESSAGE);
       else
       queryField.setText("Delete  from "+table+" where "+whereField.getText()+"
"+opration+" "+valueField.getText());
        }
       //resetFields();
       }
        }// end of ActionPerformed
}//end of  Class
```

#Code for transformation module

```
import com.datamover.db.*;
import java.util.*;
import java.sql.*;

public class DataMover
{
 /**
  * The source database.
  */
 private Database source;

 /**
  * The target database.
  */
 private Database target;
```

```
/**
 * The list of tables, from the source database.
 */
private List<String> tables = new ArrayList<String>();

public Database getSource()
{
 return source;
}

public void setSource(Database source)
{
 this.source = source;
}

public Database getTarget()
{
 return target;
}

public void setTarget(Database target)
{
 this.target = target;
}
public void createTable(String table) throws DatabaseException
{
 String sql;

 // if the table already exists, then drop it
 if (target.tableExists(table))
 {
```

```
    sql = source.generateDrop(table);
    target.execute(sql);
  }

  // now create the table
  sql = source.generateCreate(table);
  target.execute(sql);
}


private void createTables() throws DatabaseException
{
  System.out.println("Create tables.");
  Collection<String> list = source.listTables();
  for (String table : list)
  {
    try
    {
      System.out.println("Create table: "+table);
      createTable(table);
      tables.add(table);
    } catch (DatabaseException e)
    {
      e.printStackTrace();
    }
  }
}

private void copyTable(String table) throws DatabaseException
{
  StringBuffer selectSQL = new StringBuffer();
```

```
StringBuffer insertSQL = new StringBuffer();
StringBuffer values = new StringBuffer();

Collection<String> columns = source.listColumns(table);

System.out.println("Begin copy: " + table);

selectSQL.append("SELECT ");
insertSQL.append("INSERT INTO ");
insertSQL.append(table);
insertSQL.append("(");

boolean first = true;
for (String column : columns)
{
 if (!first)
 {
  selectSQL.append(",");
  insertSQL.append(",");
  values.append(",");
 } else
  first = false;

 selectSQL.append(column);
 insertSQL.append(column);
 values.append("?");
}
selectSQL.append(" FROM ");
selectSQL.append(table);

insertSQL.append(") VALUES (");
```

```
insertSQL.append(values);
insertSQL.append(")");

// now copy
PreparedStatement statement = null;
ResultSet rs = null;

try
{
 statement = target.prepareStatement(insertSQL.toString());
 rs = source.executeQuery(selectSQL.toString());

 int rows = 0;

 while (rs.next())
 {
  rows++;
  for (int i = 1; i <= columns.size(); i++)
  {
   statement.setString(i, rs.getString(i));
  }
  statement.execute();
 }

 System.out.println("Copied " + rows + " rows.");
 System.out.println("");
}
catch (SQLException e)
{
 throw (new DatabaseException(e));
}
```

```
finally
{
 try
 {
  if( statement!=null )
    statement.close();
 }
 catch (SQLException e)
 {
  throw (new DatabaseException(e));
 }
 try
 {
  if( rs!=null )
    statement.close();
 }
 catch (SQLException e)
 {
  throw (new DatabaseException(e));
 }
}
}

private void copyTableData() throws DatabaseException
{
 for (String table : tables)
 {
  copyTable(table);
 }

public void exportDatabse() throws DatabaseException
```

```
 {
  createTables();
  copyTableData();
 }
}
```

#Code for Report generation module

```
import java.io.*;
import java.awt.event.*;
import javax.swing.*;
import javax.swing.event.*;
import java.lang.*;
import java.sql.*;
import java.util.*;
import java.io.*;
import java.awt.BorderLayout;

public class SalesAnalysis extends JFrame implements ActionListener
{
      private JLabel lbl;
      private JRadioButton prodrbtn;
      private JRadioButton custrbtn;
      private JRadioButton supplierbtn;
      private JButton previewbtn,refreshbtn;
      private final int x=110,y=50;
      private int l=150,w=20;
      private int xInc=100,yInc=40;
      // data base variable
         private Connection con=null;
         private Statement st=null;
         private ResultSet rs=null;
         private DatabaseMetaData dbmd;
```

```
    private ResultSetMetaData rsmd;
    final static String jdbcDriver = "sun.jdbc.odbc.JdbcOdbcDriver";
    private boolean start=false;

public static void main(String[] args)
    {
        SalesAnalysis obj = new SalesAnalysis();

    }
public SalesAnalysis()
{
     super("SalesAnalysis");
      try
      {

    UIManager.setLookAndFeel(UIManager.getSystemLookAndFeelClassName());
          }
              catch(Exception unused)
       {  }
      setLayout(null);
       try{Class.forName(jdbcDriver);}
       catch(Exception ee) {}
       setBounds(50,50,500,400);
       addFields();
       loadDataBase();
       //productwise();
            //Customerwise();
            //supplierwise();

            setDefaultCloseOperation(JFrame.EXIT_ON_CLOSE);
      setVisible(true);
```

```
    }
private void addFields()
    {
            lbl=new JLabel("Sales reports:Select the category");
            lbl.setBounds(x-xInc,y-yInc,l+20,w);
            prodrbtn=new JRadioButton("Productwise sales");
            prodrbtn.setBounds(x,y,l,w);
            custrbtn=new JRadioButton("Customerwise sales");
            custrbtn.setBounds(x,y+50,l,w);
            supplierbtn=new JRadioButton("Supplierwise sales");
            supplierbtn.setBounds(x,y+100,l,w);
            add(lbl);
            add(prodrbtn);
            add(custrbtn);
            add(supplierbtn);
            previewbtn=new JButton("Preview");
           //previewbtn.addActionListener(this);
            previewbtn.setBounds(x-xInc,y+200,l,w);
            add(previewbtn);
            previewbtn.addActionListener(this);

           refreshbtn=new JButton("Refresh");
           refreshbtn.setBounds(x+50,y+200,l,w);
           add(refreshbtn);
    }
public void productwise()
{
   String str=null;
   String itemname=null;
   float totqty=0;
   Float qty=null;
```

```
String sqty=null;
  try
  {
     File file = new File("out.txt");
     BufferedWriter out = null;
     out = new BufferedWriter(new FileWriter(file));
      rs=st.executeQuery("SELECT b.s_itemname, sum(a.i_qty) AS totqty,
        round(sum(a.i_tot),0) AS Totalvalue FROM det_sales AS a,
        mst_productnews AS b, vwmstSales AS c WHERE a.i_productid=b.i_id and
        a.i_productid=c.i_productid and (c.s_billno=a.s_billno and
        c.i_compid=a.i_companyid and c.s_docid=a.s_docid) GROUP BY
        s_itemname ORDER BY s_itemname");
      out.write("PRODUCTWISE SALES ANALYSIS");
           out.newLine();
      out.write("-----------------------------------------------------------------------------
                  ----------------------------");
      out.newLine();
      out.write("ITEMNAME                                                QTY          Total
        value");
      out.newLine();
      out.write("-----------------------------------------------------------------------------
            ----------------------------");
      out.newLine();
      while(rs.next())
      {

        itemname=(rs.getString("s_itemname")+"                                                ");
        itemname=itemname.substring(0,50);
        out.write(itemname+"\t \t "+rs.getFloat("totqty")+"\t \t"
                              +rs.getFloat("Totalvalue"));
        out.newLine();
```

```
            //System.out.println(((rs.getString("s_itemname")+"
                        ")+"\t \t "+rs.getInt("totqty")));
          }
          out.newLine();
          out.write("-----------------------------------------------------------------------");
          out.write("                                    Total :  "+totqty);
          out.newLine();
          out.close();
        }
       catch(Exception ee) {}
    }
}
```

#Sales forecasting source code in R language

```
x <- c(4086789,4796493,5953621,6445670,7107664)
 par10 = 'FALSE'
 par9 = '1'
 par8 = '2'
 par7 = '2'
 par6 = '2'
 par5 = '1'
 par4 = '0'
 par3 = '0'
 par2 = '1'
 par1 = '3'
par1 <- as.numeric(par1) #cut off periods
par2 <- as.numeric(par2) #lambda
par3 <- as.numeric(par3) #degree of non-seasonal differencing
par4 <- as.numeric(par4) #degree of seasonal differencing
par5 <- as.numeric(par5) #seasonal period
par6 <- as.numeric(par6) #p
par7 <- as.numeric(par7) #q
```

```
par8 <- as.numeric(par8) #P
par9 <- as.numeric(par9) #Q
if (par10 == 'TRUE') par10 <- TRUE
if (par10 == 'FALSE') par10 <- FALSE
if (par2 == 0) x <- log(x)
if (par2 != 0) x <- x^par2
lx <- length(x)
first <- lx - 2*par1
nx <- lx - par1
nx1 <- nx + 1
fx <- lx - nx
if (fx < 1) {
fx <- par5
nx1 <- lx + fx - 1
first <- lx - 2*fx
}
first <- 1
if (fx < 3) fx <- round(lx/10,0)
(arima.out <- arima(x[1:nx], order=c(par6,par3,par7), seasonal=list(order=c(par8,par4,par9),
period=par5), include.mean=par10, method='ML'))
(forecast <- predict(arima.out,par1))
(lb <- forecast$pred - 1.96 * forecast$se)
(ub <- forecast$pred + 1.96 * forecast$se)
if (par2 == 0) {
x <- exp(x)
forecast$pred <- exp(forecast$pred)
lb <- exp(lb)
ub <- exp(ub)
}
if (par2 != 0) {
x <- x^(1/par2)
```

```
forecast$pred <- forecast$pred^(1/par2)
lb <- lb^(1/par2)
ub <- ub^(1/par2)
}
if (par2 < 0) {
olb <- lb
lb <- ub
ub <- olb
}
(actandfor <- c(x[1:nx], forecast$pred))
(perc.se <- (ub-forecast$pred)/1.96/forecast$pred)
opar <- par(mar=c(4,4,2,2),las=1)
ylim <- c( min(x[first:nx],lb), max(x[first:nx],ub))
plot(x,ylim=ylim,type='n',xlim=c(first,lx))
usr <- par('usr')
rect(usr[1],usr[3],nx+1,usr[4],border=NA,col='lemonchiffon')
rect(nx1,usr[3],usr[2],usr[4],border=NA,col='lavender')
abline(h= (-3:3)*2 , col ='gray', lty =3)
polygon( c(nx1:lx,lx:nx1), c(lb,rev(ub)), col = 'orange', lty=2,border=NA)
lines(nx1:lx, lb , lty=2)
lines(nx1:lx, ub , lty=2)
lines(x, lwd=2)
lines(nx1:lx, forecast$pred , lwd=2 , col ='white')
box()
par(opar)
#dev.off()
prob.dec <- array(NA, dim=fx)
prob.sdec <- array(NA, dim=fx)
prob.ldec <- array(NA, dim=fx)
prob.pval <- array(NA, dim=fx)
perf.pe <- array(0, dim=fx)
```

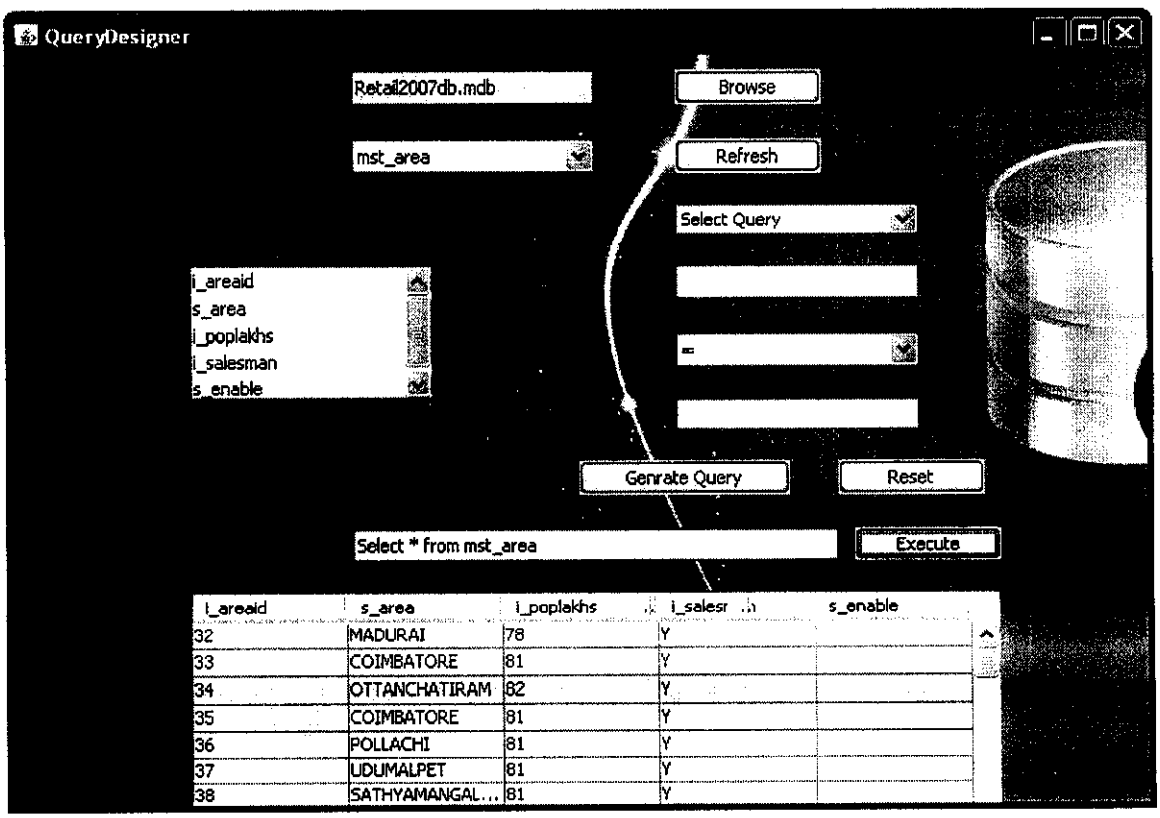
```
perf.mape <- array(0, dim=fx)
perf.mape1 <- array(0, dim=fx)
perf.se <- array(0, dim=fx)
perf.mse <- array(0, dim=fx)
perf.mse1 <- array(0, dim=fx)
perf.rmse <- array(0, dim=fx)
for (i in 1:fx) {
locSD <- (ub[i] - forecast$pred[i]) / 1.96
perf.pe[i] = (x[nx+i] - forecast$pred[i]) / forecast$pred[i]
perf.se[i] = (x[nx+i] - forecast$pred[i])^2
prob.dec[i] = pnorm((x[nx+i-1] - forecast$pred[i]) / locSD)
prob.sdec[i] = pnorm((x[nx+i-par5] - forecast$pred[i]) / locSD)
prob.ldec[i] = pnorm((x[nx] - forecast$pred[i]) / locSD)
prob.pval[i] = pnorm(abs(x[nx+i] - forecast$pred[i]) / locSD)
}
perf.mape[1] = abs(perf.pe[1])
perf.mse[1] = abs(perf.se[1])
for (i in 2:fx) {
perf.mape[i] = perf.mape[i-1] + abs(perf.pe[i])
perf.mape1[i] = perf.mape[i] / i
perf.mse[i] = perf.mse[i-1] + perf.se[i]
perf.mse1[i] = perf.mse[i] / i
}
perf.rmse = sqrt(perf.mse1)
plot(forecast$pred, pch=19, type='b',main='ARIMA Extrapolation Forecast', ylab='Forecast and
95% CI', xlab='time',ylim=c(min(lb),max(ub)))
dum <- forecast$pred
dum[1:par1] <- x[(nx+1):lx]
lines(dum, lty=1)
lines(ub,lty=3)
lines(lb,lty=3)
```

```
#dev.off()
load(file="D:/charting/createtable.xls")
a<-table.start()
a<-table.row.start(a)
a<-table.element(a,'Univariate ARIMA Extrapolation Forecast',9,TRUE)
a<-table.row.end(a)
a<-table.row.start(a)
a<-table.element(a,'time',1,header=TRUE)
a<-table.element(a,'Y[t]',1,header=TRUE)
a<-table.element(a,'F[t]',1,header=TRUE)
a<-table.element(a,'95% LB',1,header=TRUE)
a<-table.element(a,'95% UB',1,header=TRUE)
a<-table.element(a,'p-value<br />(H0: Y[t] = F[t])',1,header=TRUE)
a<-table.element(a,'P(F[t]>Y[t-1])',1,header=TRUE)
a<-table.element(a,'P(F[t]>Y[t-s])',1,header=TRUE)
mylab <- paste('P(F[t]>Y[',nx,sep='')
mylab <- paste(mylab,'])',sep='')
a<-table.element(a,mylab,1,header=TRUE)
a<-table.row.end(a)
for (i in (nx-par5):nx) {
a<-table.row.start(a)
a<-table.element(a,i,header=TRUE)
a<-table.element(a,x[i])
a<-table.element(a,'-')
a<-table.element(a,'-')
a<-table.element(a,'-')
a<-table.element(a,'-')
a<-table.element(a,'-')
a<-table.element(a,'-')
a<-table.element(a,'-')
a<-table.row.end(a)
```

```
}
for (i in 1:fx) {
a<-table.row.start(a)
a<-table.element(a,nx+i,header=TRUE)
a<-table.element(a,round(x[nx+i],4))
a<-table.element(a,round(forecast$pred[i],4))
a<-table.element(a,round(lb[i],4))
a<-table.element(a,round(ub[i],4))
a<-table.element(a,round((1-prob.pval[i]),4))
a<-table.element(a,round((1-prob.dec[i]),4))
a<-table.element(a,round((1-prob.sdec[i]),4))
a<-table.element(a,round((1-prob.ldec[i]),4))
a<-table.row.end(a)
}
a<-table.end(a)
table.save(a,file="D:/charting/3rwb71301491635.xls")
a<-table.start()
a<-table.row.start(a)
a<-table.element(a,'Univariate ARIMA Extrapolation Forecast Performance',7,TRUE)
a<-table.row.end(a)
a<-table.row.start(a)
a<-table.element(a,'time',1,header=TRUE)
a<-table.element(a,'% S.E.',1,header=TRUE)
a<-table.element(a,'PE',1,header=TRUE)
a<-table.element(a,'MAPE',1,header=TRUE)
a<-table.element(a,'Sq.E',1,header=TRUE)
a<-table.element(a,'MSE',1,header=TRUE)
a<-table.element(a,'RMSE',1,header=TRUE)
a<-table.row.end(a)
for (i in 1:fx) {
a<-table.row.start(a)
```

```
a<-table.element(a,nx+i,header=TRUE)
a<-table.element(a,round(perc.se[i],4))
a<-table.element(a,round(perf.pe[i],4))
a<-table.element(a,round(perf.mape1[i],4))
a<-table.element(a,round(perf.se[i],4))
a<-table.element(a,round(perf.mse1[i],4))
a<-table.element(a,round(perf.rmse[i],4))
a<-table.row.end(a)
}
a<-table.end(a)
table.save(a,file="D:/charting/4znqm1301491636.xls")
```

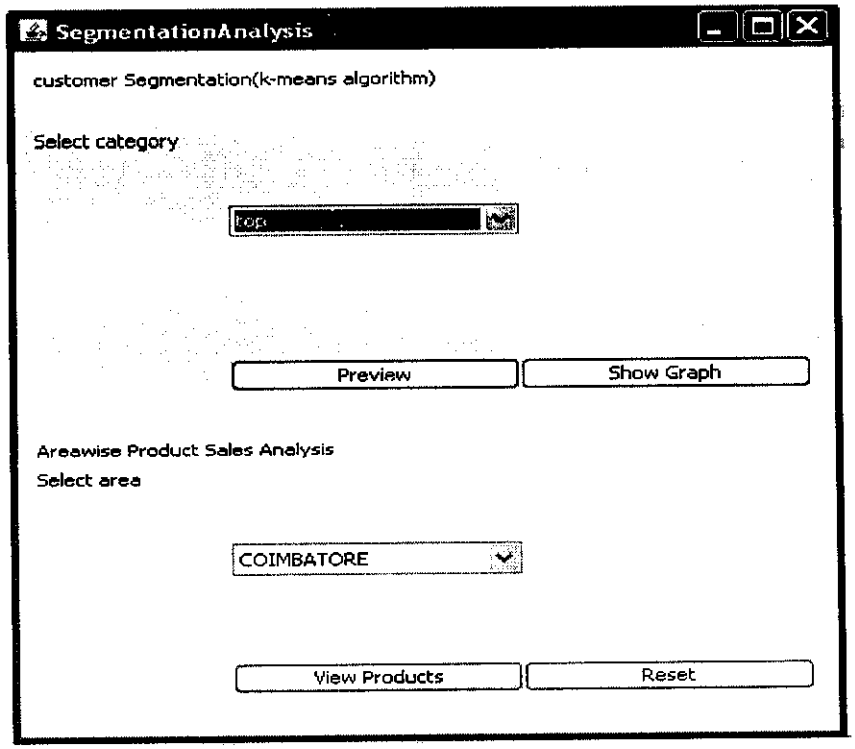
8.2 SNAPSHOTS

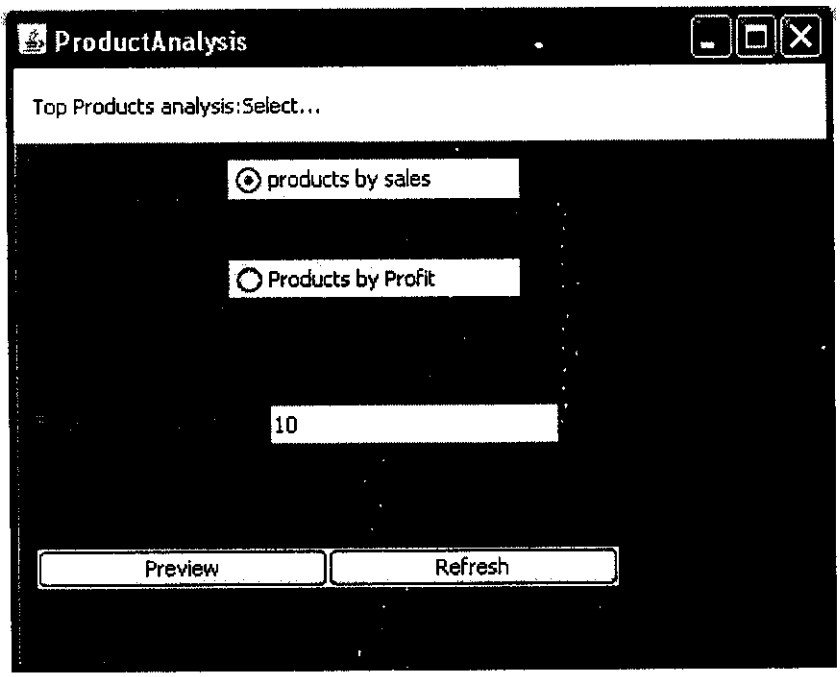


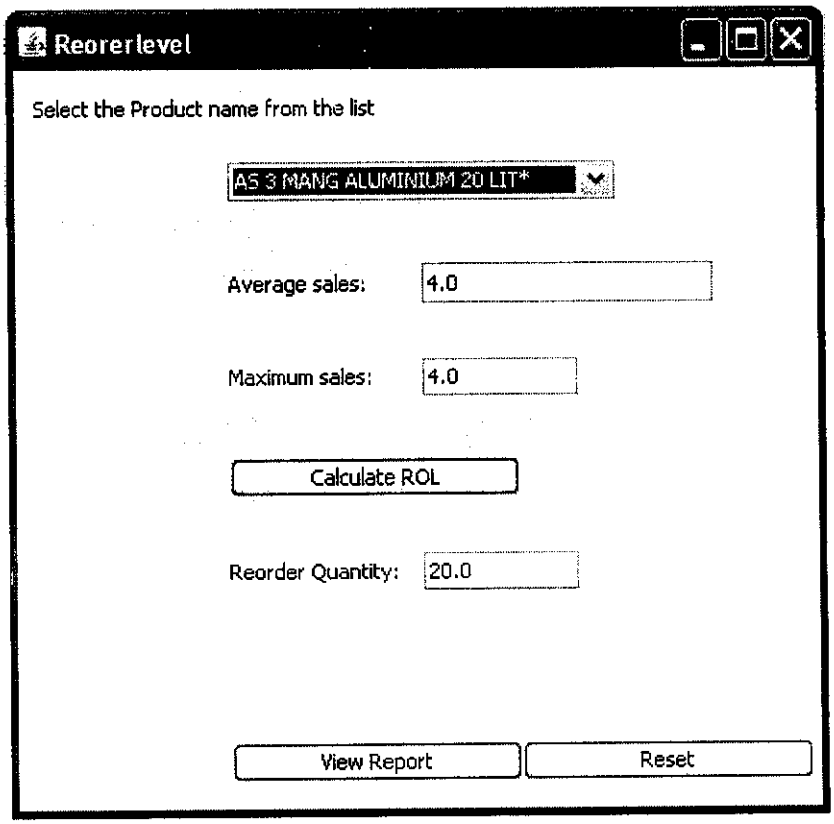
ResultSet

mst_productnew

i_Id	i...	s_itemname	i_basicunit	i_taxable	i_companyid	i_rateid
1	5	brilliant white	5	1.0	11	2
2	9	Yellow	6	1.0	7	1
3	5	p o.red	5	1.0	7	1
4	5	DULUX ENL. D...	5	1.0	7	1
5	5	ACE PAINTS	7	0.0	12	1
6		TEST ITEM ONE	5	0.0	12	1
7		TEST ITEM TWO	8	1.0	13	4
8	5	ANITHA THINN...	5	1.0	14	5
9		CHETTINAD CE...	12	1.0	15	7
10		SNOWCEM PL...	62	1.0	2	15
11		AP EMU SAFFR...	56	1.0	1	
12		GLASS DOOR ...	55	1.0	1	
13		BR TEMPLE KN...	5	1.0	1	
14		AP EMU MISTYV...	56	1.0	1	
15		GILLARICO PAI...	56	1.0	2	15
16		LUX EN GOLDE...	56	1.0	7	
17		SST FAST YELL...	56	1.0	7	
18		AP EMU MARO...	56	1.0	7	
19		ALU RADIO BR...	5	1.0	1	
20		BR VIPUL HING...	95	1.0	1	
21		DX AE RENOIR ...	56	1.0	1	
22		AP EN MERCE...	56	1.0	1	
23		AP EN DEEP O...	34	1.0	1	
24		SONY CUTTIN...	5	1.0	1	
25		WOOD SCREW...	83	1.0	1	




ProductAnalysis
Top Products analysis:Select...
products by sales
Products by Profit
10
Preview
Refresh


Reorerlevel
Select the Product name from the list
AS 3 MANG ALUMINIUM 20 LIT*
Average sales:
4.0
Maximum sales:
4.0
Calculate ROL
Reorder Quantity:
20.0
View Report
Reset

REFERENCES

1. Alkis Simitsis, Panos Vassiliadis, Timos Sellis. (2005) 'Optimizing ETL Processes in Data Warehouses', 21st International Conference on Data Engineering (ICDE'05), Vol.72, pp.564-575.

2. Bannon, L.J. (2000) 'Group Decision Support Systems: An analysis and critique', In Proceedings of 5th European Conference on Information Systems, Cork, Vol.20, pp. 526-539.

3. Cody, W.F., Kreulen, J.T, Krishna, V. and Spangler, W.S. (2002) 'The Integration Of Business Intelligence and Knowledge Management', IBM Systems Journal,Vol.41, No. 4, 180-184.

4. Darbe and Michael, D. (2003) 'A Business Intelligence is more than a buzzword', Credit Union Magazine', Vol.12, pp.33-34.

5. Davenport, Tom et al. (2002) 'Data to Knowledge to Results: Building an Analytical Capability',California Management Review, Vol.14, pp.117-138.

6 .Faguo Zhou, Bingru Yang, Linna Li and Zhuo Chen. (2008) 'Overview of the New Types of Intelligent Decision Support System', IEEE An Intetnational Conference on Innovative Computing Information and Control, pp.260-267.

7. Fausett, L. (1998) 'Fundamentals of Neural Networks: Architectures, Algorithms and Applications', Prentice Hall .

8. Gottinger, H.W. and Weimann, P. (1992) 'Intelligent Decision Support Systems', Journal of Decision Support Systems, Vol. 8, No. 4, pp. 317-332.

9.George Papastefanatos and Panos Vassiliadis. (2008) 'Policy regulated management of ETL evolution', 22nd conference on data management, pp.5-10.

10.He Li Zang Miao and Cai Liang. (2009) 'Research and Implementation of a Universal ETL Management Platform based on Telecom Industry', IEEE International conference on computational intelligence, Vol.3, pp.1-4.

11.Hill.T and Lewicki.P (2007) 'Statistical methods and applications',Statsoft Inc.

12.Hong, X.K., Liu, J.L and Xie, J.C. (2004) 'Interactive Decision Support Algorithm and Its Application', Journal of Systems Science and Systems Engineering, Vol.32,pp.475-480.

13.Jie Song Yubin Bao Jingang Shi. (2010) 'A Triggering and Scheduling Approach for ETL in a Data Warehouse', Vol.21, pp.91-98.

14.Kudyba, S. and Hoptroff, R. (2001) 'Data Mining and Business Intelligence: A Guide to Productivity', Idea Group mc, Vol.2, pp.23-26.

15.H.Lohinger (1999) 'Teach Me Data Analysis',Springer-Verlag publications,pp.120-160.

16.Michalewicz, Z., Schmidt, Michalewicz, M. (2008) 'Case Study:Intelligent Decision Support system', IEEE intelligent systems, Vol.20, No.4 ,pp.44-47.

17.Nemati, H.R., Steiger, D.M., Lyer, L. and Herschel, R.T. (1996) 'Knowledge warehouse:an architectural integration of knowledge management, decision support, Artificial Intelligence and data warehousing ', AlMagazine, Vol.17, No.3, pp.37- 54.

18.O'Brien, J. and Kok, (2006) 'Business Intelligence and Telecommunication Industry',Centre for Information and Knowledge Management, University of Johannesburg,Vol.15, pp.1-5.

19.Panos Vassiliadis and Alkis Simitsis. (2006) 'Extract transform Load', IBM journal,Vol.3, pp.40-45.

20.Paulraj Ponniah, (2007) 'Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals', Wiley India Pvt. Ltd.,

21.Shaikh, M.U. and Sarah Yaqoob. (2010) 'Intelligent decision making based on data mining using differential evolution algorithms and framework for ETL workflow management',IEEE international conference on Computer Engineering and Applications (ICCEA), Vol.1,pp.22-26.

22.Storn, R. and Price, K. (1997) 'Differential Evolution-A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces', Journal of Global Optimization, Vol. 11, pp. 341-359.

23.Slowik and Bialko, M. (2007) 'Training of Artificial Neural Networks Using Differential Evolution Algorithm', IEEE International Conference on Human System Interactions, Vol.19, pp.10-15.

24.Thiele, M., Fischer, U. and Lehner, W. (2007) 'Partition-based workload Scheduling in living data warehouse environments, IEEE DOLAP,Vol.13, pp.57-64.

25.Tong Gang Cui Kai Song Bei. (2008) 'Research and application of Business Intelligence in Retail Industry',vol.45,pp.87-91.

26. Tong Jun and Cuikai(2009) 'The Research and application of ETL in Business Intelligence project',2009 International Forum on Information Technology,Vol.4,pp.27-31.

27.Thomas Connolly and Carolyn Begg. (2008) ' Database Systems:A Practical Approach to Design, Implementation, and Management', Addison-Wesley.

28. Vasiliki Tziovara, Panos Vassiliadis, Alkis Simitsis. (2007) 'Deciding the Physical Implementation of ETL', Proceedings of the ACM tenth international workshop on Data warehousing and OLAP, Vol.42, pp.49-56.

29. Weizhuo Lu ,Yaowu Wang and Bing Yao. (2006) 'Decision Support System for Work Flow Management', IEEE international conference on convergence intelligence,Vol.15, pp.990-996.

30. Wang, H. and Wang, S.H. (2008) 'A knowledge management approach to data mining process for business intelligence', Industrial Management & Data Systems, Vol.10, pp.622–634.

DESIGN OF INTELLIGENT DECISION MAKING SYSTEM AND ETL MANAGEMENT

PROJECT REPORT

Submitted by

R.JANAKI

Reg. No: 0920108006

In partial fulfillment for the award of the degree of

MASTER OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

KUMARAGURU COLLEGE OF TECHNOLOGY

(An Autonomous Institution Affiliated to Anna University, Coimbatore)

COIMBATORE – 641 049

APRIL 2011