

# The Study on Distributed Data Warehouse Modelling



T07350

**MS Research Dissertation**

**By:**

**Atika Qazi**

**(517-FBAS/MSCS/F08)**

**Supervised By:**

**Prof. Dr Maqbool-Uddin Shaikh**

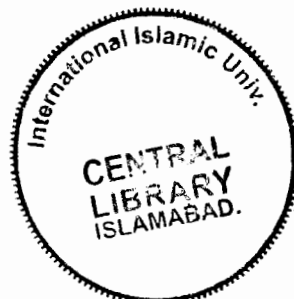
**Co-Supervised By:**

**Ms.Zareen Sharf**

**Department of Computer Science, Faculty of Basic and Applied Sciences,**

**International Islamic University, Islamabad.**

**2010**



Accession No. TH 7350

DATA ENTERED

MS  
005.758  
ATS

- 1- Distributed databases.
- 2- Database management.

**Department of Computer Science  
International Islamic University Islamabad**

**Dated:** -----

**Final Approval**

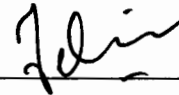
This is to certify that we have read the thesis submitted by **Atika Qazi**, registration No. **517AS/MSCS/F09**. It is our judgment that this project is of standard to warrant its acceptance by the International Islamic University, Islamabad, for the Degree of **MS in Computer Science**.

**Project Evaluation Committee**

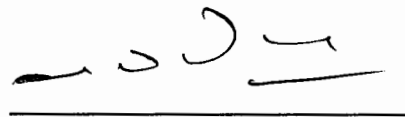
**External Examiner:**  
**Dr. Abdus Sattar**,  
Former Director General,  
Pakistan Computer Bureau, Islamabad.



**Internal Examiner:**  
**Ms. Tehmina Amjad**,  
Lecturer,  
Department of Computer Science,  
Faculty of Basic and Applied Sciences,  
International Islamic University, Islamabad.



**Supervisor:**  
**Prof. Dr. Maqbool Uddin Shaikh**,  
Computer Sciece Department,  
COMSATS Institute of Information Technology, Islamabad.



**Co-Supervisor:**  
**Ms. Zareen Sharaf**,  
Lecturer,  
Department of Computer Science,  
Faculty of Basic and Applied Sciences,  
International Islamic University, Islamabad.



## **Dedication**

**Dedicated To My Beloved Parents  
Especially To My Father**

**Atika Qazi  
517-FBAS/MSCS/F08**

**A dissertation Submitted To**  
**Department of Computer Science,**  
**Faculty of Basic and Applied Sciences,**  
**International Islamic University, Islamabad**  
**As a Partial Fulfillment of the Requirement for the Award of the**  
**Degree of *MS in Computer Science*.**

## **Declaration**

I hereby declare that this thesis "*The Study on Distributed Data Warehouse Modeling*" neither as a whole nor as a part has been copied out from any source. It is further declared that I have done this research with the accompanied research report entirely on the basis of my personal efforts, under the proficient guidance of my teachers especially my supervisor *Dr. Maqbool Uddin Shaikh and Ms. Zareen Sharf*. If any part of the system is proved to be copied out from any source or found to be reproduction of any project from any of the training institute or educational institutions, I shall stand by the consequences.

*Atika*

**Atika Qazi**

**517-FBAS/MSCS/F08**

## **Acknowledgement**

First of all I am obliged to Allah Almighty the Merciful, the Beneficent and the source of all Knowledge, for granting me the courage and knowledge to complete this thesis. There are a number of people to whom I am greatly indebted, without them this thesis might not have been completed. I want to thank my friends and colleagues, **Mr.Naveed Ahmed Khan, Mr.Saif-ur-Rahman, Ms. Zakia Jalil, Ms.Azra Shamim, Ms.Fatima Zaka, Ms.Sumera, Ms.Amna, Ms. Masuma and little Acer**, for their support through out this research work.

My deepest gratitude goes to my **family** for their unflagging love and support throughout my carrier, my sisters **Javaria** and **Sadaf** and brother **Usman** are always my carriers of hope to help me cope with any difficult situation, and my dearest **parents** who have supported me spiritually. Many thanks for all their honest efforts and love that makes me stand here successfully. I cannot repay them for their love and support throughout my life.

I would also like to thank some special names here. I owe my most sincere gratitude to **Mr. Nabeel-ur-Rahman** for his generous assistance, sharing experience and knowledge in writing and editing of my thesis. **Muhammad Huzaifa Khan**, whose extensive discussions related to my work has been very helpful for this study. I warmly thank **Mr. Kamran Khan** for reliable participation related to simulation. All of their help truly strengthen my thesis and lifted me up throughout this era.

I would like to express my warm thanks to my supervisor, **Dr. Maqbool Uddin Shaikh & Ms.Zareen Sharif**, without their guidance this research work would not have been presented.



---

**Atika Qazi**

**517-FBAS/MSCS/F08**

## **Project In Brief**

<b>Project Title</b>	<b>The Study on Distributed Data Warehouse Modeling</b>
<b>Undertaken By</b>	<b>Atika Qazi</b>
<b>Supervised By:</b>	<b>Professor Dr.Maqbool Uddin Shaikh</b>
<b>Start Date</b>	<b>10th April,2010</b>
<b>Completion Date</b>	<b>1st September,2010</b>
<b>Tools &amp; Technologies</b>	<b>Simulation tool is C#.Net</b>
<b>Documentation Tools</b>	<b>Microsoft Word 2003/2007</b>
<b>Operating System:</b>	<b>Windows XP</b>
<b>System Used:</b>	



## **Abstract**

This research study is for organized data warehouse architecture. The study includes finding efficient way of answering queries that are coming from multiple classes of users. The study also discusses the order of processing for the desired result. The category of users is differentiated with respect to priority level. The rank is pre assigned to a set of users according to the hierarchical order.

As priority level is already assigned to a set of users, the results are generated toward different classes of users accordingly. The threshold value is also maintained in order to accommodate the users that have very low query processing time. In this way the research study also focuses on time management and avoids long wait for modest query. The different architectures are studied and then a new phase name "Priority Allocation Layer" has been introduced in this research work. The layer's internal working and related algorithm is also presented here. The objective of research work is to introduce a new dimension in data warehouse architecture; the proposed research is towards query handling generated from multiple directions in an organized way.

# Table of Contents

<b>Dedication</b>	<b>III</b>
<b>Acknowledgment</b>	<b>VI</b>
<b>Abstract</b>	<b>VIII</b>
<hr/>	
<b>1 Introduction.....</b>	<b>1</b>
1.1 Motivations and Challenges .....	1
1.2 Background.....	2
1.3 Research Domain.....	2
1.4 Proposed Approach.....	3
1.5 Thesis Outline.....	4
<b>2 Literature Survey.....</b>	<b>6</b>
2.1 Introduction.....	6
2.1.1 Data Warehouse.....	6
2.1.2 Subject Oriented .....	6
2.1.3 Integrated .....	7
2.1.4 Time-variant .....	7
2.1.5 Non-volatile .....	7
2.2 Architecture of data Warehouse .....	7
2.2.1 Two Level Architecture of Data Warehouse .....	9
2.2.2 Independent Data Mart Data Warehouse Environment.....	10
2.2.3 Dependent Data Mart Data Warehouse Environment .....	11
2.2.4 Implementing a Data Warehouse.....	12
2.2.5 Tools for Accessing Data Warehouse.....	13
2.2.6 Data reporting tools .....	15
2.3 Limitations.....	15
2.4 Summary.....	16
<b>3 Requirement Analysis .....</b>	<b>18</b>
3.1 Introduction.....	18
3.2 Problem Scenarios .....	18
3.3 Focus of Research.....	19
3.3.1 Centralized Approach .....	20
3.3.2 Distributed Approach.....	20
3.3.3 OLAP (online analytical processing).....	20
3.3.4 Priority Allocation Algorithm.....	20
3.3.5 SQL Server 8 (Standard Query Language).....	21
3.3.6 Data Marts .....	21

3.3.7	Metadata.....	21
3.3.8	Conceptual Model.....	21
3.3.9	Intrinsic & Contextual Quality.....	22
3.3.10	ETL Tools (Extract, Transform and Load).....	22
3.4	Summary.....	22
<b>4</b>	<b>System Design.....</b>	<b>24</b>
4.1	Introduction.....	24
4.2	Proposed Solution.....	24
4.3	Reference Architecture.....	25
4.4	Proposed Architecture.....	27
4.5	Proposed Algorithm.....	28
4.6	Summary.....	31
<b>5</b>	<b>Implementation.....</b>	<b>33</b>
5.1	Introduction.....	33
5.2	Simulation Environment.....	33
5.3	Sequence Diagram.....	34
5.4	Data Flow Control.....	35
5.5	The Pseudo Code.....	36
5.6	Summary.....	37
<b>6</b>	<b>Testing and Performance Evaluation.....</b>	<b>39</b>
6.1	Introduction.....	39
6.2	Test Scenarios.....	39
6.2.1	Description.....	39
6.5	Performance.....	46
	Screen shots.....	46
6.4	Summary.....	59
<b>7</b>	<b>Conclusion and Outlook.....</b>	<b>62</b>
7.1	Introduction.....	62
7.2	Achievements.....	63
7.3	Improvements.....	63
7.4	Future Recommendations.....	63
7.5	Summary.....	64
	<b>List of Acronyms.....</b>	<b>73</b>

## List of Figures

Figure 2-1 Basic Architecture of Data Warehouse by Bill Inmon .....	8
Figure 2-2 Detailed logical architecture of data warehouse 1.20 .....	8
Figure 2-3 Star schema [25] .....	9
Figure 2-4 Snow flake schema [25].....	9
Figure 2-5 Dependent Data Mart Data Warehouse Environment [21].....	12
Figure 2-6 independent Data Mart Data Warehouse Environment1 [21].....	11
Figure 4-1 Architecture of Web warehouse.....	25
Figure 4-2 Architecture of High way management Data Warehouse.....	26
Figure 4-3 Proposed Architecture of a data warehouse.....	27
Figure 5-1 shows the sequence diagram of algorithm.....	34
Figure 5-2 This figure shows the flow of a data in proposed architecture. ....	35
Figure 6-1 shows the present condition before processing.....	41
Figure 6-2 shows the processed queries .....	41
Figure 6-3 Processed.....	41
Figure 6-4 shows empty space in buffer after process.....	41
Figure 6-5 shows current state of buffer.....	42
Figure 6-6 shows pointing queries that are going to process.....	42
Figure 6-7 Shows Processed queries .....	42
Figure 6-8 shows empty space in buffer after process.....	42
Figure 6-9 shows Buffer after answering already processed queries .....	42
Figure 6-10 shows Buffer condition after applying sorting on queries .....	42
Figure 6-11 shows buffer according to priority levels.....	42
Figure 6-12 shows Buffer condition after apply sorting on processing/time bases.....	43
Figure 6-13 shows Time wise sorting.....	43
Figure 6-14 shows Buffer condition during execution of queries .....	43
Figure 6-15 shows processed query.....	43
Figure 6-16 Shows Buffer condition after processing Q5.....	43
Figure 6-17 shows the processed query.....	43
Figure 6-18 Shows Buffer condition after processing Q3.....	43
Figure 6-19 shows processed query.....	44
Figure 6-20 Shows Buffer condition after processing Q2.....	44
Figure 6-21 shows the processed query.....	44
Figure 6-22 Buffer condition after processing Q4.....	44
Figure 6-23 shows the processed query.....	44
Figure 6-24 Buffer condition after processing Q4.....	44
Figure 6-25 shows the processed query.....	44
Figure 6-26 Buffer condition after processing Q1.....	44
Figure 6-27 processed query.....	45

Figure 6-28 Buffer condition after processing Q6.....	45
Figure 6-29 processed query.....	45
Figure 6-30 Buffer condition after processing Q4.....	45
Figure 6-31 Buffer condition after processing all the queries .....	45
Figure 6-32 initial condition of buffer .....	46
Figure 6-33 after apply threshold. ....	48
Figure 6-34 after execution of values equivalent or less than threshold. ....	49
Figure 6-35 before applying priority .....	50
Figure 6-36 after applying priority sort on the basis of classes .....	51
Figure 6-37 time of the very first query before applying time sort .....	52
Figure 6-38 after applying time sort .....	53
Figure 6-39 query with high execution time is placed in last.....	54
Figure 6-40 during processing, the queries are processed accordingly .....	55
Figure 6-41 queries from class 'A' are answered.....	56
Figure 6-42 queries from class 'B' are answered .....	57
Figure 6-43 queries from class 'C' are processing .....	58
Figure 6-44 all the queries are processed, buffer is empty.....	59

## **1 Introduction**

# 1 Introduction

A data warehouse is a logical collection of information gathered from many different operational databases used to create business intelligence that supports business analysis activities and decision making tasks. The data warehousing market consists of tools, technologies, and methodologies that allow for the construction, usage, management, and maintenance of the hardware and software used for a data warehouse, as well as the actual data itself.

Data warehouses support business decisions by collecting, consolidating, and organizing data for reporting and analysis with tools such as online analytical processing (OLAP) and data mining. Although data warehouses are built on relational database technology, the design of a data warehouse database differs substantially from the design of an online transaction processing system (OLTP) database.

## 1.1 Motivations and Challenges

With rapid growth of traffic, efficient decision making is required. Query in bulky system is slow to process and this results in increased turnout time of applied query. In order to obtain rapid decision at right time, without delay is essential. To answer multiple queries at a time efficiently in order to get plentiful results is not possible in [1]. There is high demand of maintenances and construction work, tracks security, roots changing scenarios, security measures, etc. in day to day increasing with swift increase of traffic

In the present time complex system analysis, management, controls and decision making process of highway management, to answer multiple queries at same time in efficient way need better distributed system. There is also a need to connect all these Online Transaction Processing Systems (OLTPs), to meet the integrated data analysis need.

## 1.2 Background

This research study will make survey of existing technologies, propose a new architecture and then the validation of proposed architecture will be taken through a case study of educational sector. I will design architecture of data warehouse that may apply on any business application in a certain scenario, like Highway management system, Motorway, Educational Institutes, Industries, etc.

## 1.3 Research Domain

As Data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data therefore it supports management's decision-making process.

On-line analytical processing (OLAP) technology is a useful tool that can realize multidimensional analysis of data warehouse. OLAP can take advantage of data in data Warehouse to carry out different kinds of analysis operation and provide the result easily for the users to understand by direct viewing according to user's request.

OLAP is designed to support complex analysis operation, and it is mainly to help in decision-making. It can carry out complex inquiry processing in huge data rapidly and flexibly according to analyst's request, and provide the result to the decision-makers easily enabling them to understand in direct-viewing form.

Distributed computing refers to the use of distributed systems to solve computational problems. In distributed computing, a problem is divided into many tasks, each of which is solved by one computer.

Precisely, the research will work around distributed environment, Data warehouse, Query management, Data marts, Hierarchical distribution of data marts etc.



## 1.4 Proposed Approach

Proposed approach will improve the system in a way that brings improved and beneficial outcome as compared to previous study [1]. To acquire maximum benefits of OLAP and warehouse efficient decision making can take place in much improved way.

Research will purely focus on distributed approach. In order to improve the current work [1] ZHOU Qian, et.al.2009, the research study proposes an algorithm that works against user's priority. Here research takes place on hierarchical approach. For example data is handled first on top level by taking the data on country level, then split on provincial level, then on district level, and finally at the lower level of towns.

Here summaries of detailed data are copied at lower level of hierarchy and at higher level detailed information will be managed.

The main focus is on better management of desired data and use of OLAP in efficient way, so that queries take less time to be executed, and load will be divided on separate individual systems in order to be on safe side, as data is being distributed on separate systems. Reliability is increased in this way as load is also divided on separate individual marts. Query processing time will be reduced and querying will be better optimized. Query will be executed only on desired domain. Due to this approach turnaround time will be decreased and users don't have to wait long to get a reply of small query.

A subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process is a key idea of this thesis.

On-line analytical processing (OLAP) technology is a useful tool that can realize multidimensional analysis of data warehouse. OLAP can take advantage of data in data warehouse is used to carry on different kinds of analysis operation and provides

the result easily for the users to understand by direct viewing according to user's request.

The architecture discussed in [1] is not using the advantages of distributed system at its fullest. This study proposes a distributed architecture in order to gain maximum benefits of OLAP and warehouse technologies.

## 1.5 Thesis Outline

This research study focuses on distribution of queries on separate data marts and query optimization. The hierarchical architecture and Priority Allocation procedure will be used to cope with related problems of a data warehouse.

First chapter is about the introduction and background of the problem. The proposed solution is also discussed here.

The second chapter describes literature study. The related work and its relevance and further limitations are discussed.

This third chapter explains problem analysis, functional as well as non functional requirements of data warehouse. This chapter discusses the main points related to proposed work.

The fourth chapter briefly explains the proposed solution to the problem. The related architectures are also discussed here.

The fifth chapter explains the classes and diagrams to work in UML. Data flow charts are also discussed here.

the result easily for the users to understand by direct viewing according to user's request.

The architecture discussed in [1] is not using the advantages of distributed system at its fullest. This study proposes a distributed architecture in order to gain maximum benefits of OLAP and warehouse technologies.

## 1.5 Thesis Outline

This research study focuses on distribution of queries on separate data marts and query optimization. The hierarchical architecture and Priority Allocation procedure will be used to cope with related problems of a data warehouse.

First chapter is about the introduction and background of the problem. The proposed solution is also discussed here.

The second chapter describes literature study. The related work and its relevance and further limitations are discussed.

This third chapter explains problem analysis, functional as well as non functional requirements of data warehouse. This chapter discusses the main points related to proposed work.

The fourth chapter briefly explains the proposed solution to the problem. The related architectures are also discussed here.

The fifth chapter explains the classes and diagrams to work in UML. Data flow charts are also discussed here.

In the sixth chapter the simulation of algorithm is described. The code and snapshots are given to show the work in detail.

The last chapter describes conclusion of related and proposed work. In this chapter future work is also described.

## **2 LITERATURE SURVEY**

## 2 Literature Survey

This chapter is about the literature study and all the material related to this survey.

### 2.1 Introduction

The Data warehouse is used to create business intelligence that supports business analysis activities and an optimum way for decision-making process. Market of data warehouse consists of tools, technologies, and methodologies that allow for the construction, usage, management, and maintenance of the hardware and software. The normal size of data warehouse varies from hundreds of gigabytes to terabytes. Different scans, joins, and aggregates are performed while querying the data warehouse. The queries on data warehouse are ad hoc and multi-faceted. Throughput of query determines the success of data warehousing project. The query response time is also important factor in data warehouse success. The allocation of facts & dimension in a certain schema also affect on query success.

#### 2.1.1 Data Warehouse

Connolly et al. defined data warehouse as “*Data Warehouse is a subject-oriented, integrated, time-variant, non volatile collection of data in support of management decisions*”. References [14][17] and [18] define data warehouse as “*A set of materialized views over data sources*”.

#### 2.1.2 Subject Oriented

A data warehouse works on subject oriented approach, where subject could be customer, dealer, product or sales. It doesn't deal with day to day transactions. Hence, data warehouses typically provide a simple and concise view about particular subject.

### 2.1.3 Integrated

The data from multiple heterogeneous sources are gathered to data warehouse, such as relational databases, flat files, and on-line transaction records. To ensure consistency, data cleaning & integration techniques are applied.

### 2.1.4 Time-variant

A warehouse focuses on change over time to analyze trend in different time. Archive data is also stored to provide information from historical viewpoint.

### 2.1.5 Non-volatile

Data warehouse always store the transformed data from the operational environment into a physically separate storage. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms.

## 2.2 Architecture of data Warehouse

Basic architecture of a data warehouse consists of data source layer, integration process layer, storage area, Metadata and end user query and analysis tools. Figure 2.1 shows the basic architecture of a data warehouse. Detailed logical architecture of a data warehouse is shown in Figure 2.2 Integration and transformation programs convert data from multiple sources (system of record) into a unified format. Unified data then stored into data warehouse storage areas. As shown in Figure 2.2 data warehouse contains old data, current data, highly summarized data and data marts. The design of data warehouse includes Schemas like snow flake and star schema to hold facts and dimensions. A fact is used to record measures or states concerning an event or situation. The states are analyzed through different criteria organized in dimension.[5].The Data warehouse graph is also very helpful for design of warehouse.[5].

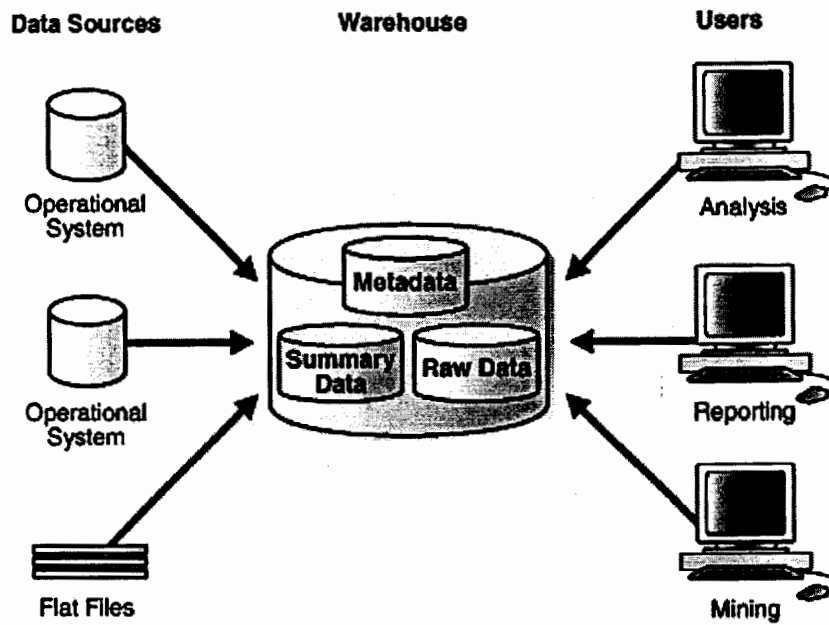


Figure 2-1 Basic Architecture of Data Warehouse by Bill Inmon

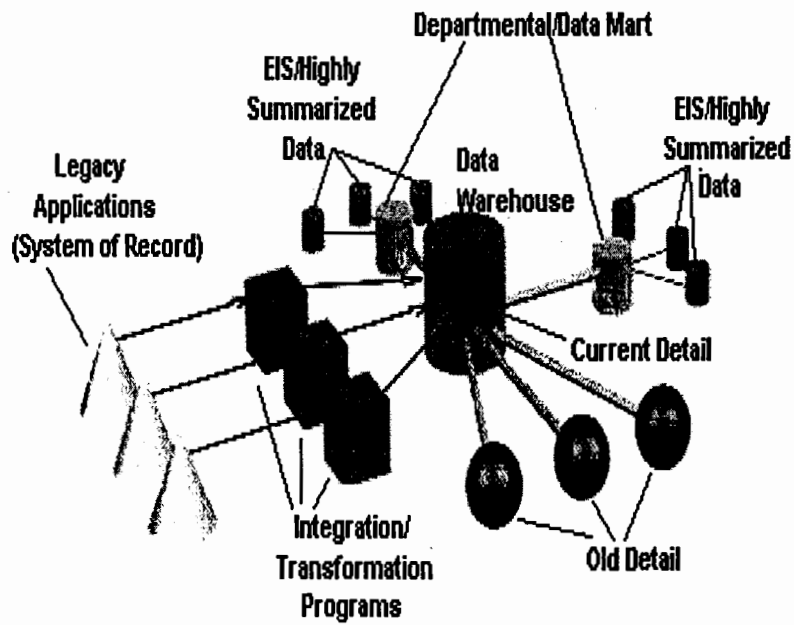


Figure 2-2 Detailed Logical Architecture of Data Warehouse [19]

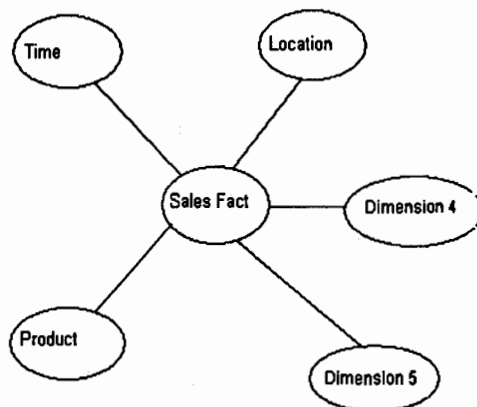


Figure 2-3 Star Schema [7]

The design of data warehouse includes Schemas, such as snow flake and star, to old facts and dimensions.

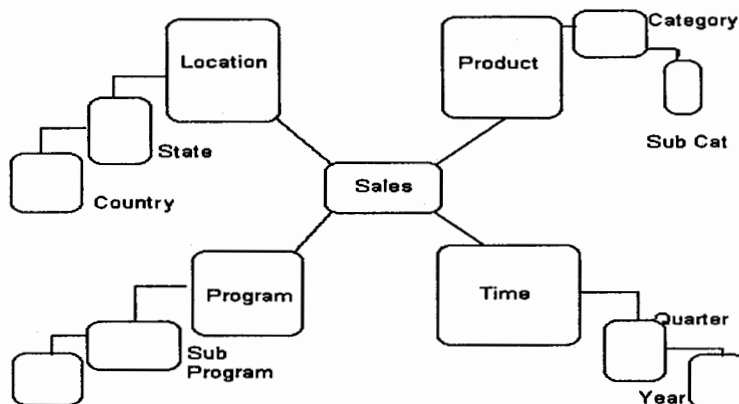


Figure 2-4 Snow Flake Schema [7]

### 2.2.1 Two Level Architecture of Data Warehouse

Hoffer et. al. presented generic two level architecture of data warehouse [22].

#### I. Data Source Systems

Data from multiple internal and external sources are extracted to load into a data warehouse. Data cleaning and data transformation operations are performed on data before loading it into the data warehouse. Data cleaning enhances the quality of data whereas transformation process converts data into a single format.



**Data sources can be quite diverse, such as**

- a. Databases
- b. Excel Sheets.
- c. Flat files.

## **II. Data Staging Area**

Data staging area layer perform three functions on data; data cleaning, data transformation and loading data into data warehouse. It cleans data that comes from multiple internal and external sources. After cleaning, data is transformed into one unified format and then loaded in to data warehouse.

## **III. Data and Metadata Storage Area (Data Warehouse)**

Data and Metadata storage area is a place where data is organized, stored, and made available for the direct querying by the end users. This layer consists of Metadata, current detailed data, old detail data, highly and lightly summarized data.

## **IV. End User Presentation Tools**

User interacts with a data warehouse by means of end user presentation tools. A variety of end user presentation tools are available that provides user multi dimension view of the data. End user access tools can be query and reporting tools, executive information system and data visualization tools.

### **2.2.2 Independent Data Mart Data Warehouse Environment**

Implementing a data warehouse is a complicated task. Some organizations break down implementation of a data warehouse into data marts. Organization first creates different independent data marts and then integrates them to provide company wide view of data.

A data mart is a data warehouse that is limited in scope. Contents of data mart is obtained from independent ETL (extract, transform and load) process or derived from the data warehouse [22]. Figure 2.3 shows independent data mart data warehouse architecture.

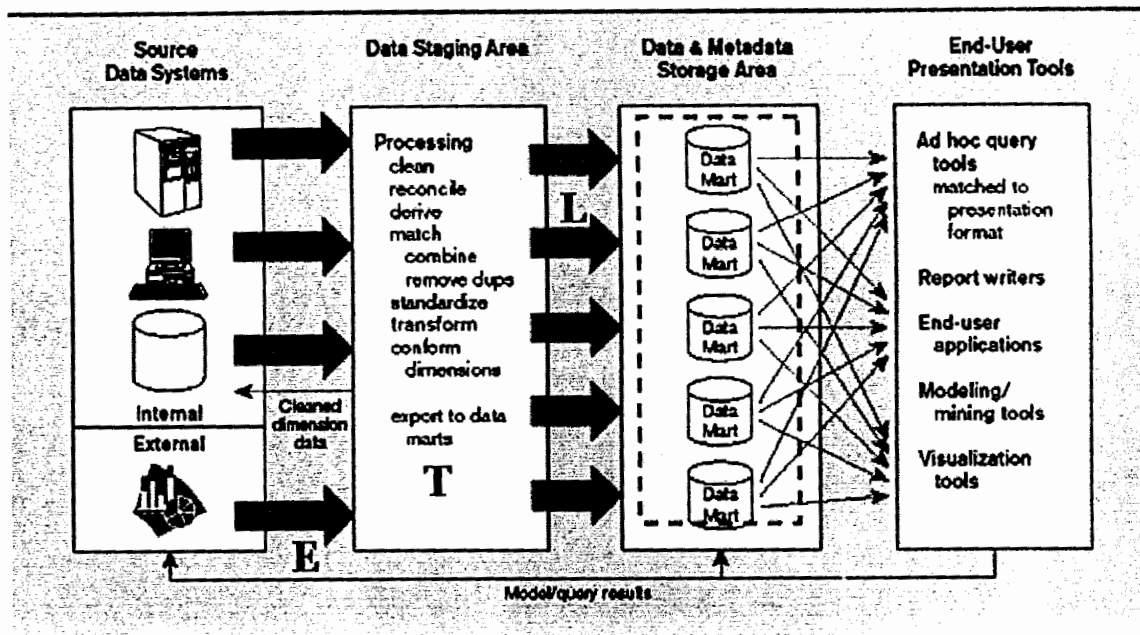


Figure 2-5 Independent Data Mart Data Warehouse Environment [22]

### 2.2.3 Dependent Data Mart Data Warehouse Environment

The study depicted few limitations of independent data mart Kimball strongly supported development of independent data mart data [21]. Dependent data mart overcomes some limitations of independent data mart. A data mart filled fully from enterprise data warehouse and its submissive data is called dependant data mart [22] Figure 2.4 shows dependant data mart data warehouse architecture.

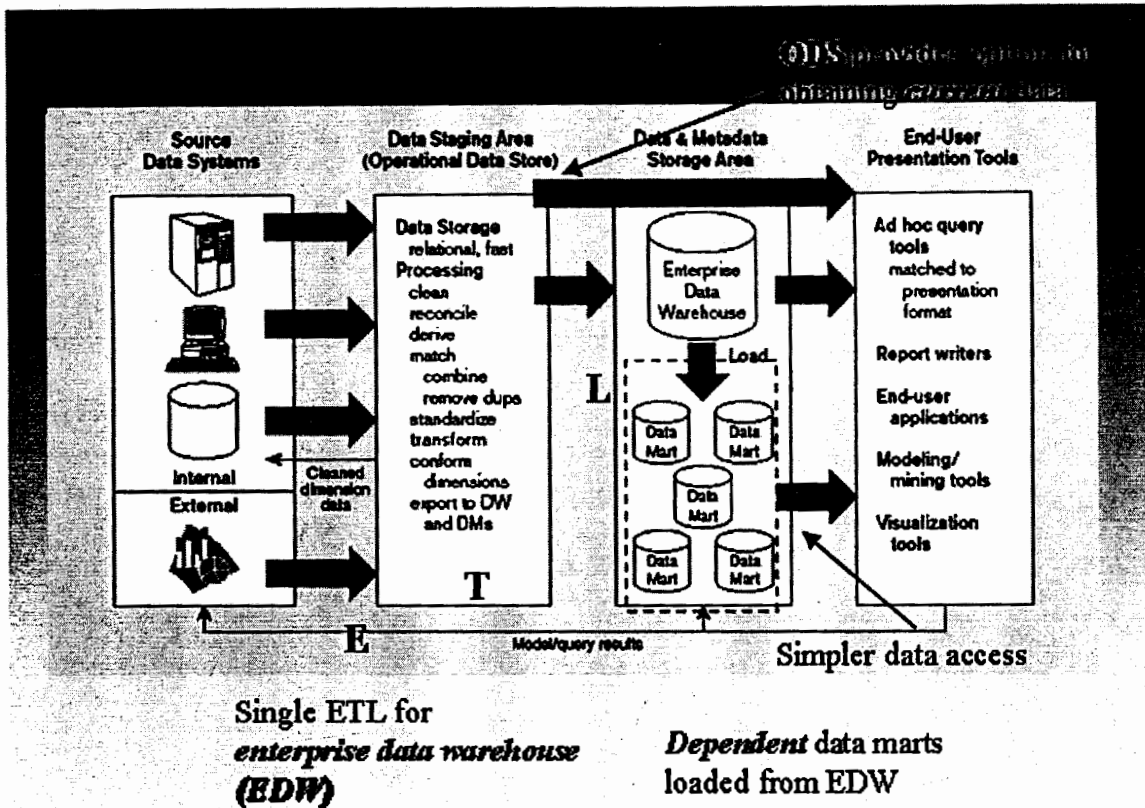


Figure 2-6 Dependent Data Mart Data Warehouse Environment [22]

### 2.2.4 Implementing a Data Warehouse

Implementation of a data warehouse involves four phases [20]. Each step is explained below.

#### I. Monitoring Phase

Transfer data from multiple sources to a data warehouse is called monitoring. There is variety of data source for data warehouse such as relational, flat file, excel sheets, news-wire, and databases. Different monitoring techniques are used such as periodic snapshots, database triggers, log shipping, data shipping (replication service), transaction shipping, polling (queries to source), screen scraping and application level monitoring.

Monitoring relates with some issues such as data compatibility and data transfer rate. Transformation is applied to compose the data into single format that is coming from multiple sources.

## **II. Integrating phase**

Processing step involves query processing, creation of indexes and materialized views. Once data is loaded into data warehouse, it should be maintained in data warehouse and later on accessed to fulfill the requirement of a user. Many access methods are available and used to access data such as B-trees, hash tables, R-trees and grids. But these methods are not used in data warehouse. Data access methods used in data warehouse are inverted lists, bit map indexes, join indexes and text indexes.

## **III. Managing Phase**

Management of data warehouse is mandatory after building it up. There are several management techniques such as metadata creation, summarized data creation, archive and back up activities.

### **2.2.5 Tools for Accessing Data Warehouse**

Special query tools have been developed and used with data warehouse. These tools include OLAP, ETL, data mining and data visualization. These query tools provide user multi dimension view of data.

#### **I. Online Analytical Processing (OLAP)**

The study of various research papers including [23] "*Online Analytical Processing (OLAP) is an efficient way to access data warehouse for multidimensional analysis and decision support*". Three types of OLAP operations are available to provide user multi dimensional view of data; roll up, drill down and pivoting. Three types of OLAP servers are available; relational, multidimensional and hybrid OLAP servers. The purpose of OLAP is also to discover trends.[12].

## II. ETL

ETL, Extract, transform, and load (ETL) is a process in database usage and especially in data warehousing that involves:

- Extracting data from outside sources
- Transforming it to fit operational needs (which can include quality levels)
- Loading it into the end target (database or data warehouse)

## III. Data Mining

Data mining is the process to discover hidden facts. Data mining explores massive datasets in order to search valuable information and knowledge. Data mining uses a variety of different techniques such as clustering, association analysis, classification, and regression and deviation detection to analyze massive data sets. Data mining have been applied to diverse domains like, biosciences, marketing, financial industry, banking and pharmaceutical companies.

## IV. Data Visualization

*“Visualization is the graphical presentation of information, with the goal of providing the viewer with a qualitative understanding of the information contents”[24].*

To recognize the data precisely data visualization takes place. Data visualization represents data in the form of graph to make it easily understandable for viewer. Many analytic techniques use data visualization to represent their results in suitable format.

### 2.2.6 Data reporting tools

Reporting tools further divide in to two parts.

- 1 **Production reporting tools** will let companies generate regular operational reports or support high level batch job, such as calculating and printing paychecks. [11].
- 2 **Report writer**, are expensive desktop tools designed for end users.

### 2.3 Limitations and Issues

The existing architecture of the data Warehouse (DW) has the following limitations and issues which need to be addressed.

- The existing architecture of DW is basic and at abstract level.
- In presents architecture data is coming from different sources & centralized over country level for decision making purpose.
- How data will be gathered and extracted from different sources.
- How the stored data will be accessed and processed using warehouse tools.
- How to distribute related data over different levels of hierarchy.
- How captured data will be distributed using different technologies.
- How to handle multiple users on the basis of priority level.
- How to resourcefully manage multiple quires at a time.
- How to reduce query response time.
- How to provide summaries of data & reduce redundancy of required data.
- How to efficiently structure schemas for warehouse.
- How to provide updates to operational level (lower level) of hierarchy.
- How to maintain the historical data of data warehouse in an improved way.

## 2.4 Summary

In this chapter literature review different terminologies related to architecture of data warehouse are discussed. It views how to develop new architecture of the data warehouse from the existing information.

The study identifies the limitation of the existing architecture. The coming chapter will propose the improved architecture of data warehouse which would more efficient methodology than is currently available. In the subsequent chapter, the research study will focus on the educational sector with emphasis on data warehouse architecture.

### **3 REQUIREMENT ANALYSIS**



the design of current architecture of a data warehouse, the required data is completely copied to desired destination which involves a lot of memory and time.

The main focus of research is on the design of data warehouse architecture in distributed manner and to handle multiple queries in an organized way, to provide better management of desired data with the use of OLAP in efficient way. In this way queries take less time to be executed and load will be divided on separate individual system (data marts). It is also safer, as data is being distributed on separate individual systems. Queries are executed only on the desired domain. Due to this approach turnaround time is decreased.

In short the research study works on distributed architecture of data warehouse, OALP query optimization, Priority Allocation Process (PAP), data marts and conceptual model.

The distributed approach of a data warehouse can be applied in diverse areas and scenarios. For example it can be applied on education, health, media, highway or other types of business or organization.

### 3.3 Focus of Research

In this research work, the priority allocation procedure in distributed environment is proposed. Research focuses on architecture of a data warehouse for education environment, as universities are the main source of getting knowledge. To build such an architecture that is suitable for related organization, but also bring ease to related group of users is also studied.

The research work takes the university as a case study for the development of data warehouse architecture. However, proposed data warehouse architecture is also efficient to work in different business environment as well.

### **3.3.1 Centralized Approach**

In the present architecture the centralized approach is applied. In existing system the scattered data is organized. The data is coming from different operational sources and centralized on provisional level. The centralized approach is not well structured and not fulfilling the today's promise. The proposed research will work on distributed approach to improve current architecture requirements.

### **3.3.2 Distributed Approach**

The present approach will enhance the architecture in distributed manner. The data coming from useful operational sources will be distributed over different levels of architecture, In order to gain full advantage of a data warehouse.

### **3.3.3 OLAP (Online Analytical Processing)**

It's a useful tool that can realize multidimensional analysis of a data warehouse. The use of OLAP is appropriate to connect all the online transaction processing systems to meet the corporate analysis need. OLAP is designed to support complex analysis operation, and it is mainly to help decision-making.

OLAP can take advantage of data in a data warehouse to carry on different kinds of analysis operations and provide the results easily for the users to understand by direct viewing according to user's request.

### **3.3.4 Priority Allocation Algorithm**

This algorithm is in proposed solution that works to bring convenience for users, as multiple users form diverse sites are interlinked with a data warehouse. The proposed algorithm is built to handle multiple queries in an efficient way and avoids maximum chances to user suffering from long wait for desired response.

### **3.3.9 Intrinsic and Contextual Quality**

The quality of a data warehouse must affect decision making. The intrinsic qualities related to research are accuracy, concurrency and completeness. The contextual quality related to the decision makers.

### **3.3.10 ETL Tools (Extract, Transform and Load)**

ETL is the process in a data warehouse to extract, transform and load. To extract the data from different outside useful sources then transform it according to operational needs and finally loading it into data warehouse.

## **3.4 Summary**

In this chapter, the research indicates different tools, terms and application. The useful martial that is necessary in the development of a distributed data warehouse is discussed. In the coming chapter research study will develop the architecture of a data warehouse with proposed solution.

## **4 SYSTEM DESIGN**

## 4 System Design

In this chapter the detailed solution of the problem is focused.

### 4.1 Introduction

The basic focus of this research study is to build distributed architecture of a data warehouse and to provide the way to handle multiple queries in well organized way to improve efficiency. The reporting tools are used to access the stored data in efficient way. On-line analytical processing (OLAP) technology is a useful tool that can realize multidimensional analysis of data warehouse. The literature survey helped in understanding different architectures that are related to warehouse architecture.

### 4.2 Proposed Solution

The data marts will be created in order to distribute the data over network. The summaries of required data will be generated and transmitted to requested nodes. The data warehouse architecture follows the hierarchical architecture. The operational level work will be processed at lower level of proposed architecture. The updates are copied to related nodes.

With the help of appropriate tools queries are generated and answered. Query will be executed only on desired domain. Due to distributed approach and query, the turnaround time will be decreased. The Priority Allocation Process will take place in order to allow user to generate queries in limited amount of time. In this way, the coming user will not suffer long wait. The algorithm will allocate priority according to already allocated time on the basis of queries complexity. The already processed queries will be answered first then queries with less executing time will be processed.

### 4.3 Reference Architecture

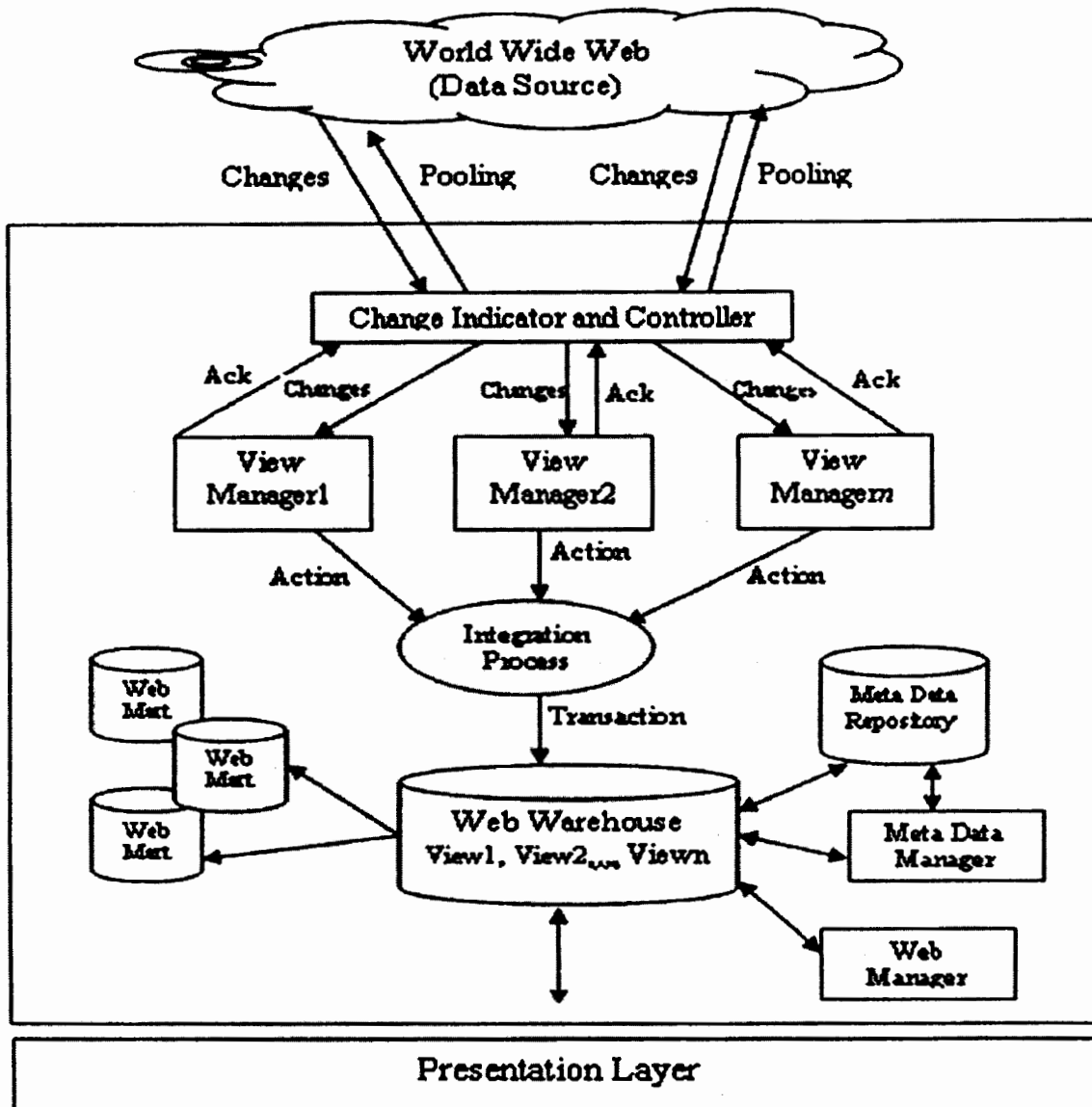


Figure 4-1 Architecture of Web Warehouse

Finally, the web warehouse and data warehouse architectures are studied. The features of both architectures are discussed. As an example, architecture of highway management is shown in figure 4-2. The proposed architecture of COMSAT Data Warehouse is shown in figure 4-3.

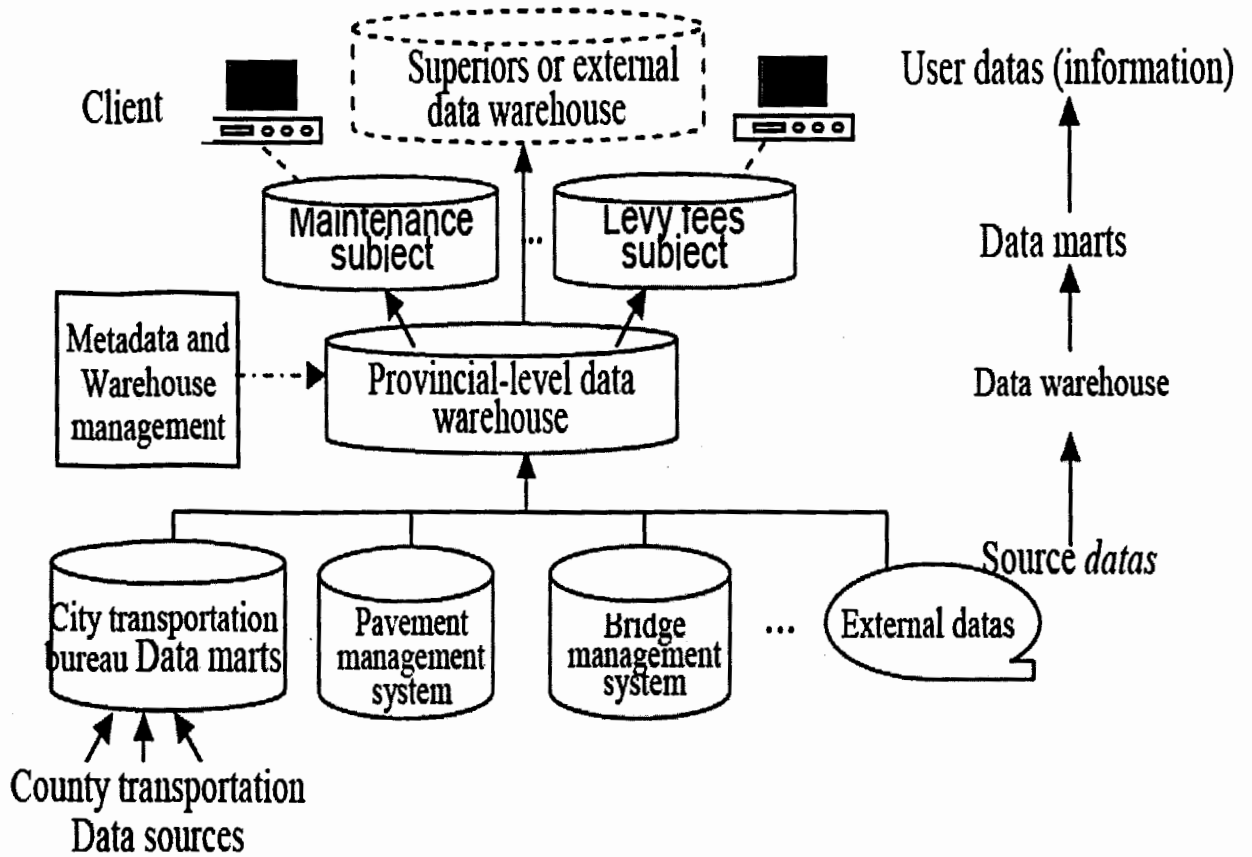


Figure 4-2 Architecture of Highway Management Data Warehouse

### 4.4 Proposed Architecture

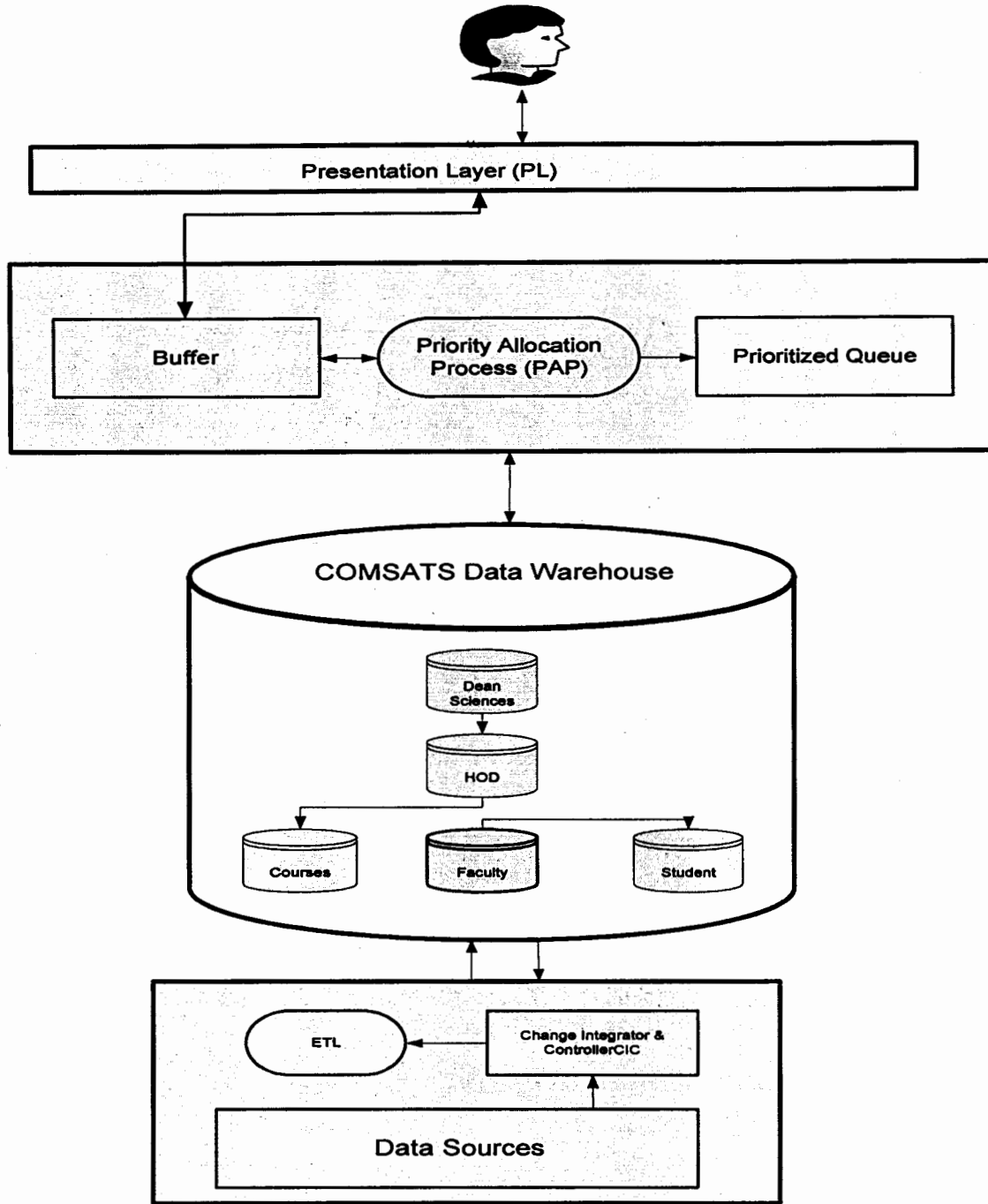


Figure 4-3 Proposed Architecture of Data Warehouse



## 4.5 Proposed Algorithm

### 4.5.1 Description

The proposed architecture works in distributed environment. Users from many directions are sending set of queries to Priority Allocation Process. The user interacts with the system through presentation layer. The priority level of user is already defined. The set of queries generated by users are denoted with  $Q$ . The  $Q$  is sent to Priority Allocation process. Here  $Q$  is filtered on the basis of a threshold value. The threshold is a predefined value and can be optimized according to system. It is assumed that query processing time is already calculated on the basis of their complexity. In practical query complexity will be calculated through an efficient algorithm. The complexity is determined by the number of joins a query may hold. The queries in  $Q$  that are less than or equal to Threshold value are filtered out accordingly.  $Q_{Thresh}$  is the set of filtered queries. The  $Q_{Thresh}$  is processed first and replied back to requested user. The queries that have greater time than threshold are kept in PAQ (Priority Allocation Queue) to apply sorting techniques. After applying sorting the PAQ forwards the set of sorted queries to PQ (Prioritized Queue). The PQ keeps the queries in the form of queue and forwards them to a data warehouse in order to get result. If the desired result against any query is not found in data warehouse then queries are further directed to data sources. The required result is then generated back to user by passing through ETL. The ETL (Extract Transform and Load) is a procedure that is called to transform the queries in required standard structure.

Initially the users from multiple directions interact with the data warehouse through presentation layer. The set of unordered queries are random queries that are coming from multiple directions and from multiple classes of users. The unordered sets of queries are denoted with  $Q$ . The queries that have less than or equivalent to threshold value are separated from  $Q$  and forwarded to data warehouse to generate results.

The queries are filtered out in PAP, filtered queries are denoted with  $Q_{Thresh}$ . The remaining set of queries are denoted with  $Q_i$ , the  $Q_i$  is sent from PAP (Priority Allocation Process) to PQ, where  $Q_i$  is sorted first on priority bases than time wise sort is applied. After PAP the queries are forwarded to PQ and then directing to data warehouse in order to find results. The results are generated back to users once found.

Following is the equation to calculate set of remaining queries  $Q_i$ .

$$Q_i = Q - Q_{Thresh}$$

where  $Q$  is the set of all unordered queries. The unordered queries are all the queries that are coming from multiple direction and classes.  $Q_{Thresh}$  is the set of queries less than Threshold value and  $Q$  is the set of all unordered queries.

The  $Q_i$  is sorted in two steps. In first step,  $Q_i$  is sorted on the basis of user's priority in order to get  $Q_{ij}$  where  $j$  is priority number. In second step, sorting is applied on the set  $Q_{ij}$  individually (i.e. each set of priority is sorted separately) on the basis of processing/time.

Time  $T_{ij}$  is defined as the time slot assigned to set  $Q_{ij}$  in order to interact with DWH in predefined time.

The results are generated to requested users accordingly.

TA 7350

### 4.5.2 The Algorithm

The Algorithm of Priority Allocation Process is given below

**Input:**  $Q_{Thresh}$  set of unsorted queries less than threshold value,  $Q_i$ , set of queries after subtracting  $Q_{Thresh}$  from  $Q$  and  $T_{ij}$ , set of time slots corresponding to each priority level. In  $Q_{ij}$ , subscript  $i$  denote the number of queries and  $j$  denote the priority level and in  $T_{ij}$  the subscript  $i$  describes the time required for processing and  $j$  denotes the priority level.

Output:  $Q_o$ , queue of queries to be executed.

**Begin**

$Q_o = 0$

For each query in  $Q_{Thresh}$

$Q_o = \text{Sort } Q_{Thresh}$  on the basis of required processing/ time

End

For each query in  $Q_i$

$Q_{ij} = \text{Sort } Q_i$  on the basis of priority levels

End

For each priority level  $j$

$Q_{Tmp} = \text{Sort } Q_{ij}$  on the basis of processing/time

$Q_{Tmp} \leftarrow T_{ij}$

$Q_o = Q_o \cup Q_{Tmp}$

End

End

where  $Q$  is the random set of queries,  $Q_{Thresh}$  is the set of unsorted queries less than threshold value and  $Q_i$  is the set of remaining queries and  $Q_{Temp}$  is a temporary variable that stores the resultant queries.

## 4.6 Summary

This chapter includes related architecture, proposed architecture, proposed methodology and detailed description of proposed algorithm. The proposed architecture will be able to organize the queries effectively. It will be designed to provide the desired results to related users according to priority basis. The educational institute will be considered as a case study to accomplish this task.

where  $Q$  is the random set of queries,  $Q_{Thresh}$  is the set of unsorted queries less than threshold value and  $Q_i$  is the set of remaining queries and  $Q_{Temp}$  is a temporary variable that stores the resultant queries.

#### 4.6 Summary

This chapter includes related architecture, proposed architecture, proposed methodology and detailed description of proposed algorithm. The proposed architecture will be able to organize the queries effectively. It will be designed to provide the desired results to related users according to priority basis. The educational institute will be considered as a case study to accomplish this task.

## **5 IMPLEMENTATION**

## 5 Implementation

In this chapter the related data flow charts, pseudo code of proposed work is given.

### 5.1 Introduction

This chapter introduces the flow chart and sequence diagram and example to present the proposed research work. Here the case study with example is discussed in detail. The software tool C# is used for simulation of proposed phase. The proposed knowledge is applied on COMSATS University Computer Science Dept. The departmental level decision making will become much easier and organized. The user to interact with data warehouse is from three different classes' higher level, medium level and lower level. The scenario in which these classes are involved is of courses offered for semester /summer breaks. The HOD (higher level) has highest rights to recommend a course. The lecturer (medium level) has to contribute their course of interest and staffs (lower level) for availability of resources are concerned.

The research work introduces the priority allocation processes in order to organize and prioritize the random queries generated by different classes of users. The detail description of PAP (Priority Allocation Processes) algorithm is discussed in fourth chapter. Here in this chapter the pseudo code, flow chart and sequence diagram is discussed.

### 5.2 Simulation Environment

The simulation environment takes place in C# programming language. It's the latest programming language that helps to simulate and validate the proposed algorithm.

### 5.3 Sequence Diagram

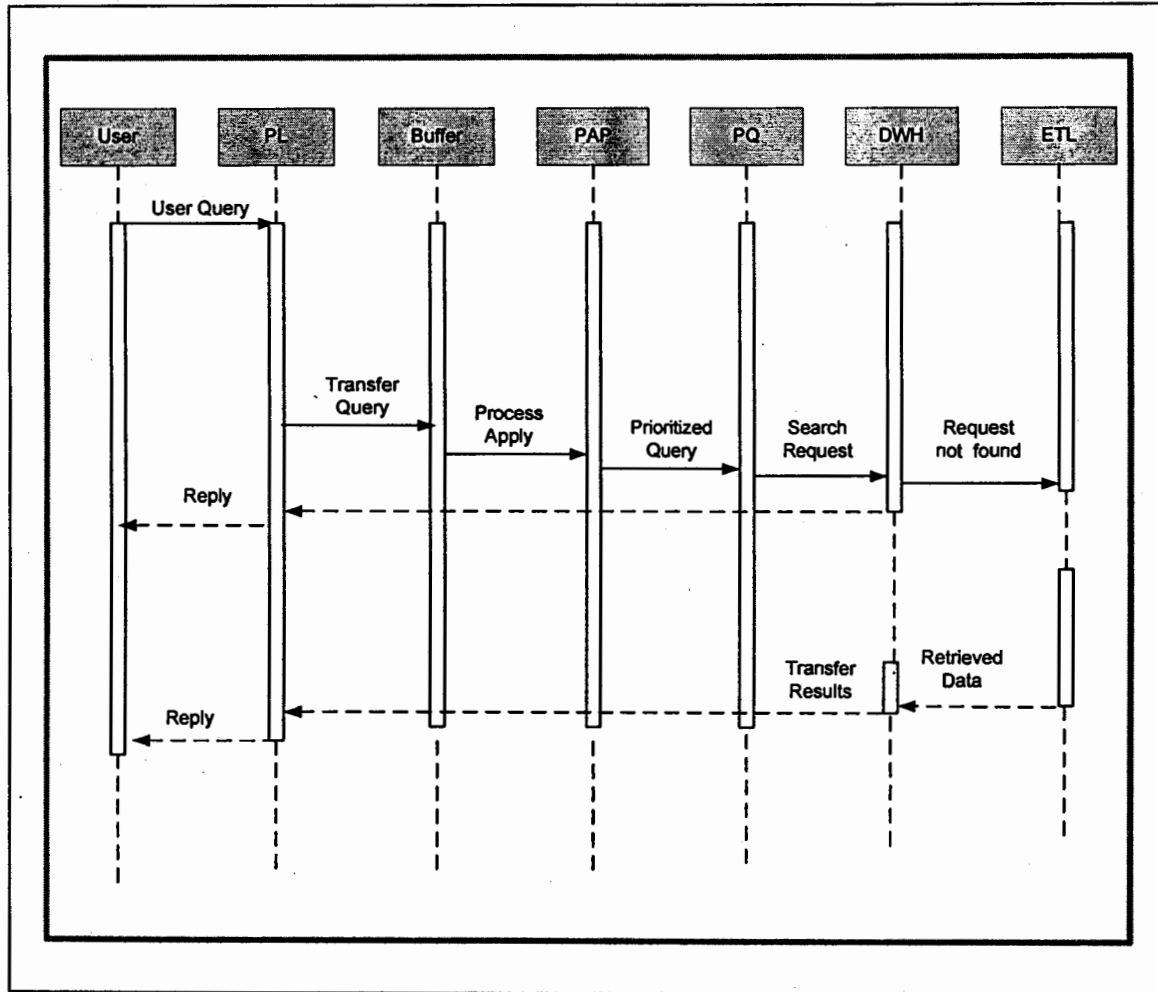


Figure 5-1 Sequence Diagram of Algorithm



## 5.4 Data Flow Control

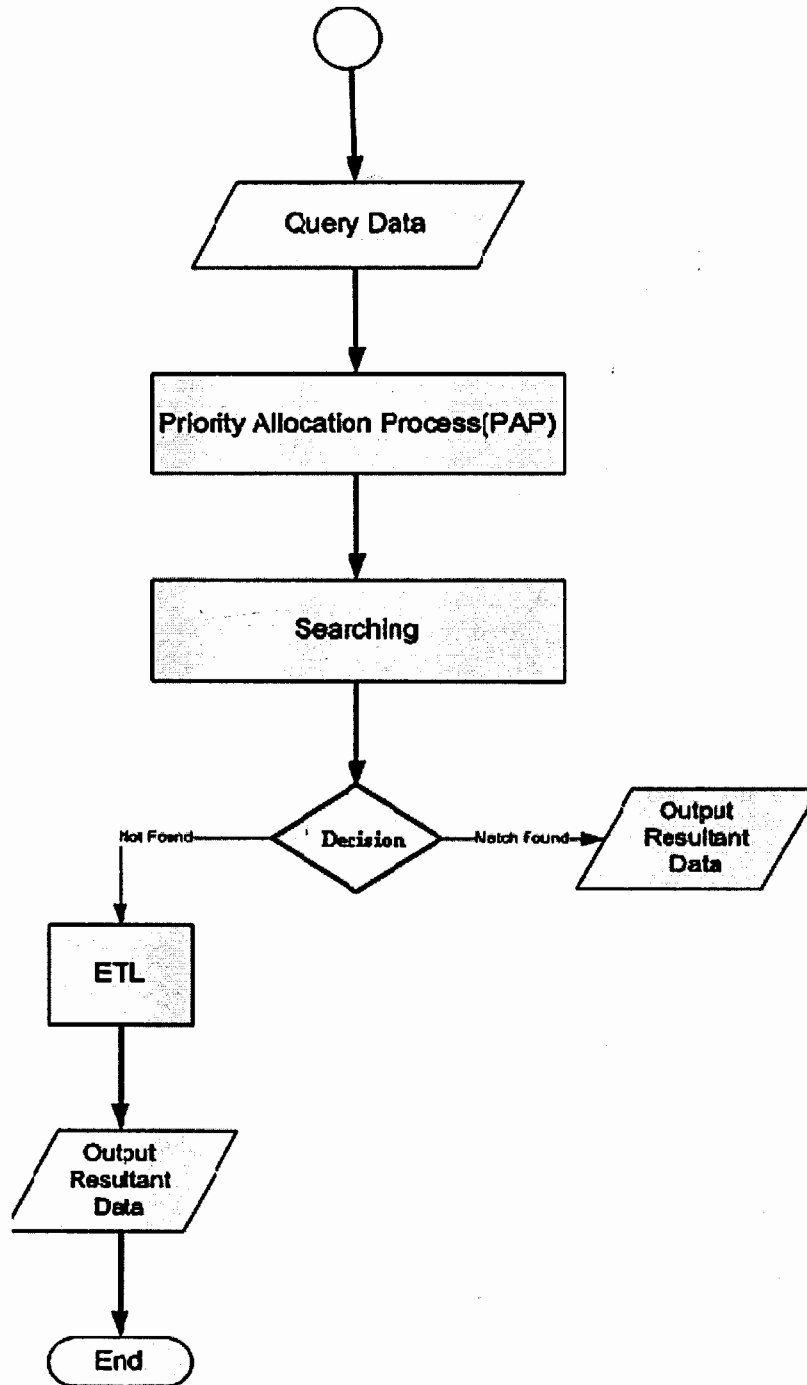


Figure 5-2 Flow of a Data in Proposed Architecture.

## 5.5 The Pseudo Code

The following is the pseudo reflecting the data flow control.

**Begin**

Let  $Q$  be the set of unordered queries

Let  $Q_{Thresh}$  be the set of filtered queries less than threshold value

Let  $Q_i$  be the set of remaining queries filtered out from  $Q$

    For each set of query in  $Q_{Thresh}$

        Sort  $Q_{Thresh}$  on the basis of priority levels

    End

For each set of query in  $Q_i$

    Sort  $Q_i$  on the basis of priority

    Sub sort on the basis of processing/time

    End

End

## 5.6 Summary

In this chapter, proposed algorithm has been applied on educational institute university level data. Initially user with assigned priority level interacts with data warehouse through presentation layer. The proposed algorithm (PAP) is between data warehouse and presentation layer. The calculations will takes place accordingly. The user gets the desired result in an organized manner through whole process.

## **6 TESTING AND PERFORMANCE EVALUATION**

## 6 Testing and Performance Evaluation

### 6.1 Introduction

In this chapter the detailed theoretical and practical example through figures and snap shots are discussed.

A simulation of the proposed algorithm is shown, also the environments in which the work is tested and results generated are discussed here. The C# is a software tool that is used for coding the program and windows Xp as the operating system. The .Net Framework is required to execute a program for simulation.

The computer system Pentium “III” and Pentium “IV” might be used for performing testing of proposed work. There are no hard and fast hardware requirements. The tested environment consists of 507MB of RAM and processor 1133MHz Pentium III. The screen shots and dry run in detail are also covered.

### 6.2 Test Scenarios

#### 6.2.1 Description

The test scenario is applied to a case study related to an educational institute. Here an example in detail is given below.

#### **Classes of users**

There are three main categories of users: at Highest Level, Medium Level and Low Level.

In this research work as a case study, Computer Science Department, COMSATS University is chosen. According to decision making level three categories are identified. The queries emanating at each level are processed and analyzed according to proposed algorithm.

The research has identified the requirements and conditions at each Highest Level, Medium Level and Low level to generate the results; where decisions are required to generate or calculate results. The study concluded three major categories that are involved in decisional level are discussed below.

The Classes of users are separately placed into defined classes and colors are also used to distinguish their category. Users from multiple directions are interacting with data warehouse by putting multiple queries. The users according to their class are assigned with priority levels to interact with system accordingly.

The **users** are identified with **U**, **Queries** with **Q**, and **Time** tag with **T**.

**Class "A" Users → Pink**



Higher Priority Level (H)

**Class "B" Users → Blue**



Medium Priority Level (M)

**Class "C" Users → Yellow**

Lower Priority Level (L)

Table related input data.

Priority Level	User (U)	Set of Query (Q)	Execution Time (T)
0		Q1.1,Q1.2,Q1.3	50sec,2min,1min
0		Q1.4,Q1.5	2min,40sec
1		Q2.1	2min
1		Q2.2,Q2.3	40sec,1min
1		Q2.4,Q2.5	10sec,20sec
2	U6	Q3.1,Q3.2,Q3.3	20sec,10sec,30sec
2	U7	Q3.4	5min
2	U8	Q3.5,Q3.6	1min,50sec

Table 1 Contents of Buffer it Holds to Execute Queries

Threshold value is 30sec

Buffer condition before processing (Q)

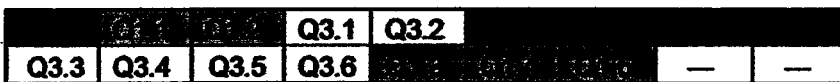


Figure 6-1 Present Condition before Processing

Buffer condition after answering already processed queries (Q)

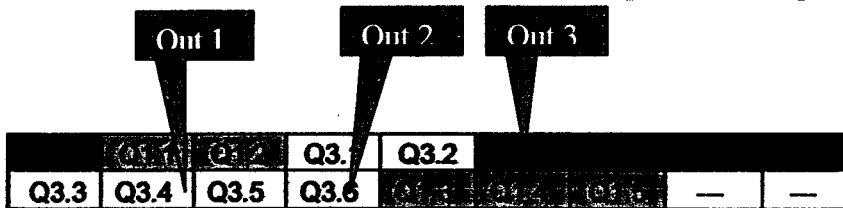


Figure 6-2 Processed Queries



Figure 6-3 Processed

Buffer condition after processing queries

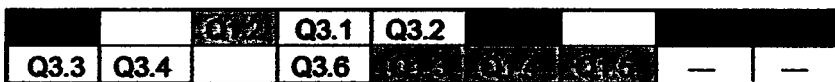


Figure 6- Empty Space in Buffer After Process

**Buffer updated condition**

	Q3.1	Q3.2				Q3.3	Q3.4
Q3.6	Q3.1	Q3.2	—	—	—	—	—

Figure 6-4 Current State of Buffer

**Buffer condition after filter queries less than “Threshold value”**

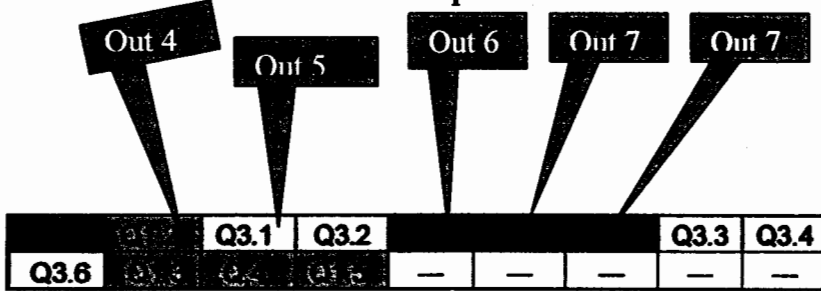


Figure 6-5 Pointing Queries that are Going to Process

1	2	3	4	5
---	---	---	---	---

Figure 6-6 Processed Queries  
**Buffer updated condition**

							Q3.4
Q3.6	Q3.1	Q3.2	Q3.3	Q3.4	—	—	—

Figure 6-7 Empty Space in Buffer After Process

**The queries left in buffer are**

			Q3.4	Q3.6	Q3.3	Q3.2	Q3.1	—
—	—	—	—	—	—	—	—	—

Figure 6-8 Buffer After Answering Already Processed Queries

**Buffer condition after sort at 1<sup>st</sup> time on priority Basis**

Q3.2	Q3.3	Q3.4	Q3.1			Q3.4	Q3.6	—
—	—	—	—	—	—	—	—	—

Figure 6-9 Buffer Condition After Applying Sorting On Queries

**The buffer rearrange after priority order**

0	0	0	1			2	2	—
—	—	—	—	—	—	—	—	—

Figure 6-10 Buffer According To Priority Levels



Buffer condition after sort at 2<sup>nd</sup> time on processing/Time Basis

Q1.5	Q1.6	Q1.7	Q1.8	Q1.9	Q1.10	Q3.6	Q3.4	—
—	—	—	—	—	—	—	—	—

Figure 6-11 Buffer Condition After Apply Sorting On Processing/Time Bases

Buffer rearrange according to processing/time

Q1.5	Q1.6	Q1.7	Q1.8	Q1.9	Q1.10	50sec	5min	—
—	—	—	—	—	—	—	—	—

Figure 6-12 Time Wise Sorting

Buffer condition during execution of queries

Q1.5	Q1.6	Q1.7	Q1.8	Q1.9	Q1.10	Q3.6	Q3.4	—
—	—	—	—	—	—	—	—	—

Figure 6-13 Buffer Condition During Execution of Queries



Figure 6-14 Processed Query

Buffer condition after processing Q5

Q1.5	Q1.6	Q1.7	Q1.8	Q1.9	Q1.10	Q3.6	Q3.4	—
—	—	—	—	—	—	—	—	—

Figure 6-15 Buffer Condition After Processing Q5



Figure 6-16 The Processed Query

Buffer condition after processing Q3

Q1.5	Q1.6	Q1.7	Q1.8	Q1.9	Q1.10	Q3.6	Q3.4	—
—	—	—	—	—	—	—	—	—

Figure 6-17 Buffer Condition After Processing Q3



Figure 6-18 Processed Query

Buffer condition after processing Q2

			Q2			Q3.6	Q3.4	
—	—	—	—	—	—	—	—	—

Figure 6-19 Buffer Condition After Processing Q2



Figure 6-20 The Processed Query

Buffer condition after processing Q4

						Q3.6	Q3.4	
—	—	—	—	—	—	—	—	—

Figure 6-21 Buffer Condition After Processing Q4



Figure 6-22 The Processed Query

Buffer condition after processing Q4

						Q6	Q4	
—	—	—	—	—	—	—	—	—

Figure 6-23 Buffer Condition After Processing Q4



Figure 6-24 The Processed Query

Buffer condition after processing Q1

						Q3.6	Q3.4	
—	—	—	—	—	—	—	—	—

Figure 6-25 Buffer Condition After Processing Q1

<b>Q3.6</b>
-------------

Figure 6-26 processed query

Buffer condition after processing Q6

							<b>Q3.4</b>	
--	-	--	--	--	--	--	-	-

Figure 6-27 Buffer Condition After Processing Q6

<b>Q3.4</b>
-------------

Figure 6-28 Processed Query

Buffer condition after processing Q4

-								

Figure 6-29 Buffer Condition After Processing Q4

Buffer condition after execution of all the queries

-								

Figure 6-30 Buffer Condition after Processing All the Queries

Queries are processed for current scenario next bulk of queries will be executed in the same way.

## 6.3 Performance

### Screen shots

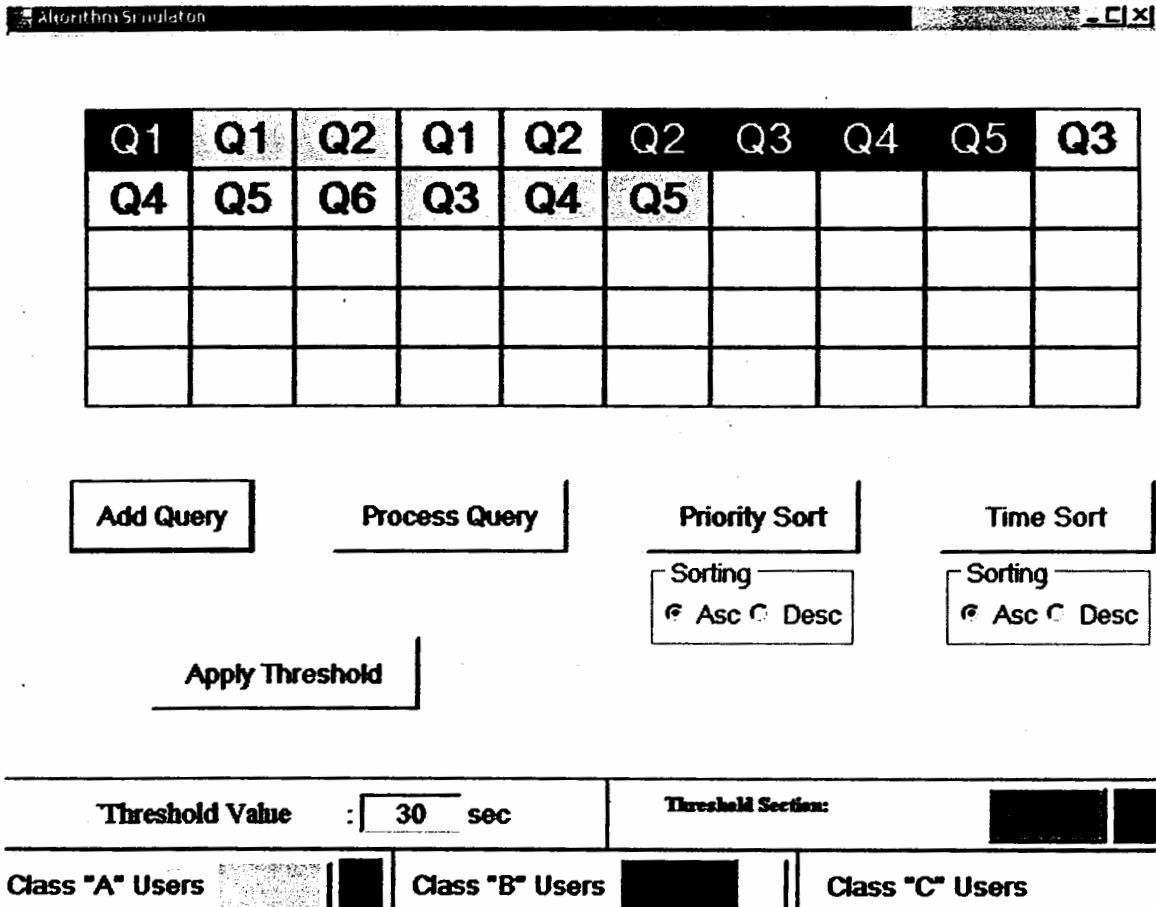


Figure 6-31 Initial Condition of Buffer

The Figure 6-31 depicts the initial condition. The users that are coming from different classes and interacting with system are shown with separate colors. The Class "A" users with pink, Class "B" users with blue similarly Class "C" users with yellow color cell is indicated.

Here the threshold section is also indicated with green color. The figure also depicts that users are interacting randomly and there is no sequence to interact the system.

The sixteen queries are entered here at this stage. This is very initial condition of buffer placing the queries randomly.

Algorithm Simulation

Q1	Q1	Q2	Q1	Q2	Q2	Q3	Q4	Q5	Q3
Q4	Q5	Q6	Q3	Q4	Q5				

Add Query

Process Query

Priority Sort

Time Sort

Sorting  
 Asc  Desc

Sorting  
 Asc  Desc

Apply Threshold

Threshold Value :  sec

Threshold Section: ████████ ████████

Class "A" Users ████████ ████████

Class "B" Users ████████ ████████

Class "C" Users ████████ ████████

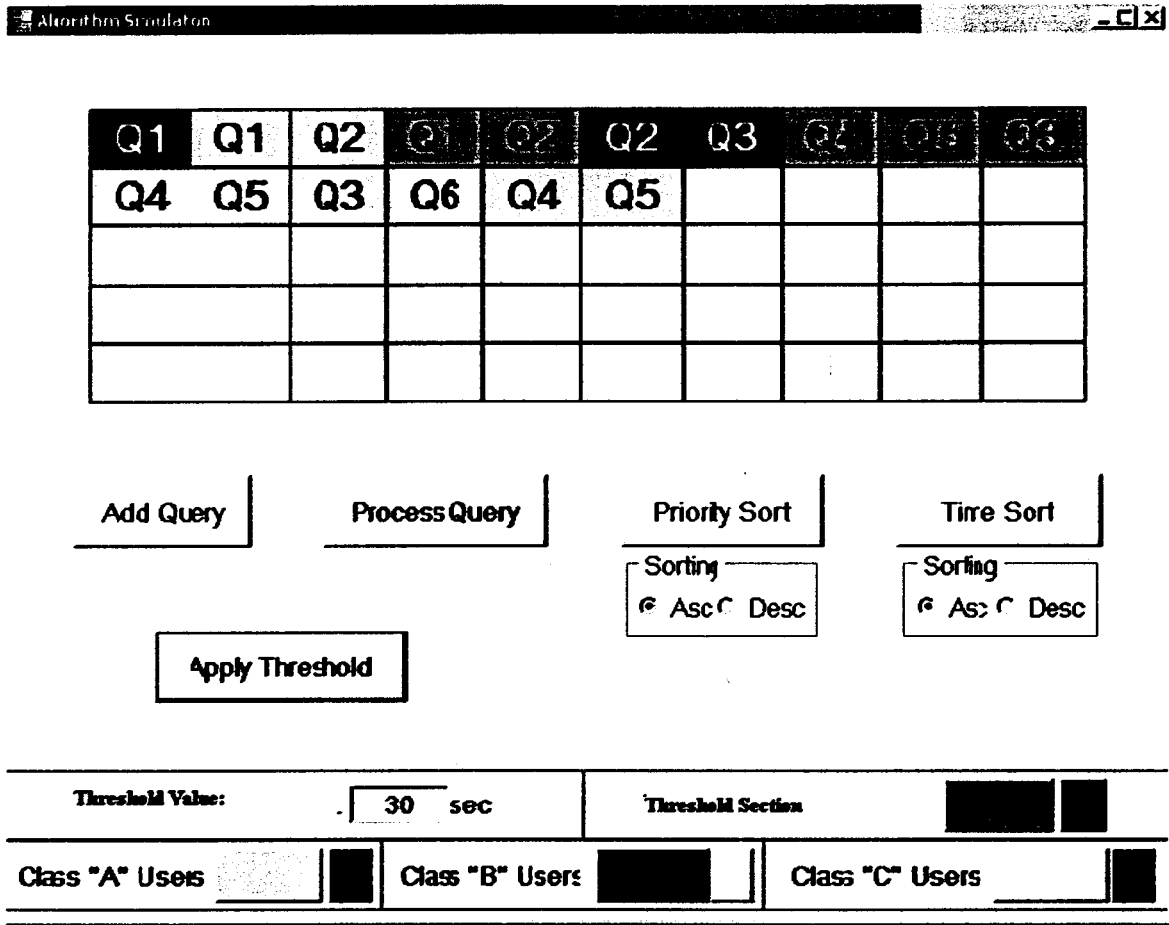


Figure 6-32 After Applying Threshold.

The Figure 6-32 shows the starting of algorithm. The boxes that are highlighted with green identified the received values that are less than or equal to threshold. As soon as buffer receives values the algorithm starts working. The threshold is the least time the query takes to process. Here threshold is taken as 30 sec. as it's visible with label Threshold value 30 sec. The five queries that are highlighted with green have least processing time. Initially processor will flush out the queries that has processing time less than or equal to threshold.

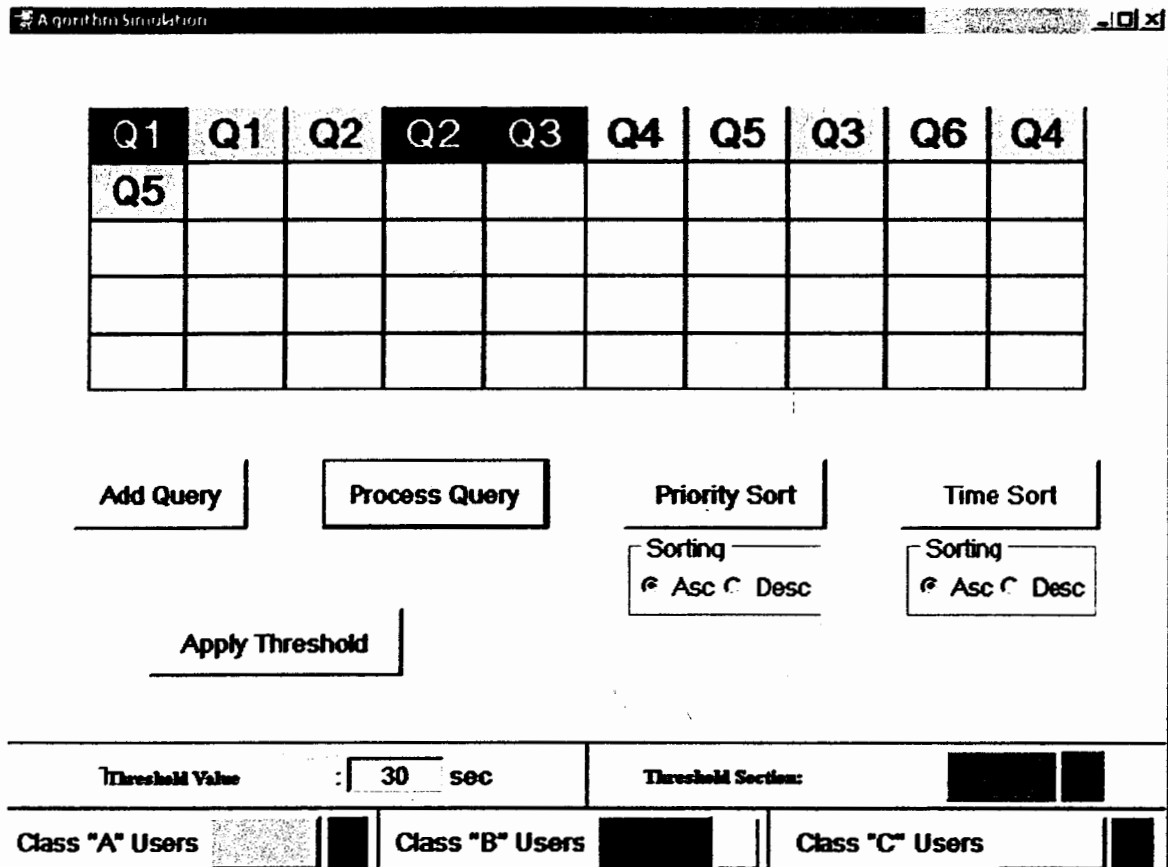


Figure 6-33 After Execution of Values Equivalent or Less than Threshold

This Figure 6-33 indicates eleven remaining queries. The remaining queries that are visible here have higher processing time. The queries are still not prioritized it will happen in next step.

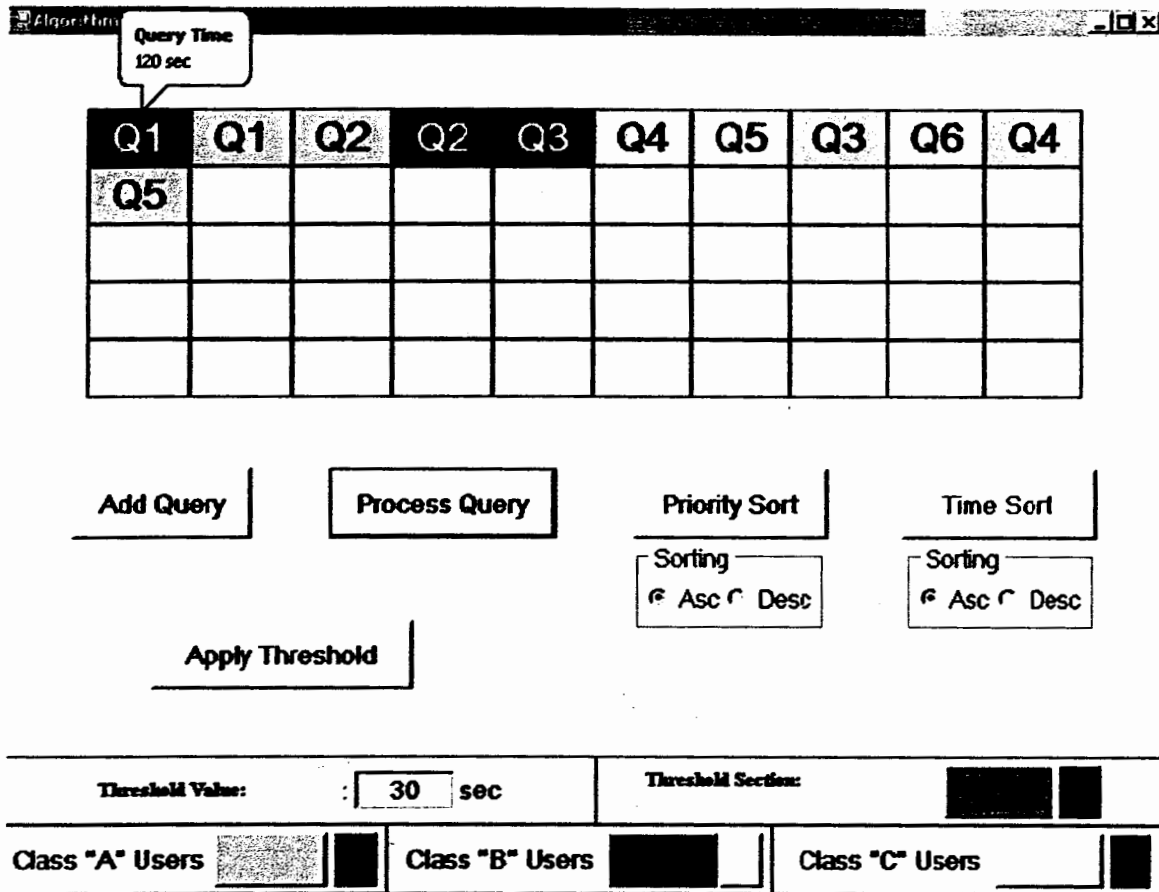


Figure 6-34 Before Applying Priority

In Figure 6-34 the first query processing time is shown as its 120 sec. The queries will change their position as soon as priority sort is applied and buffer will be organized accordingly.



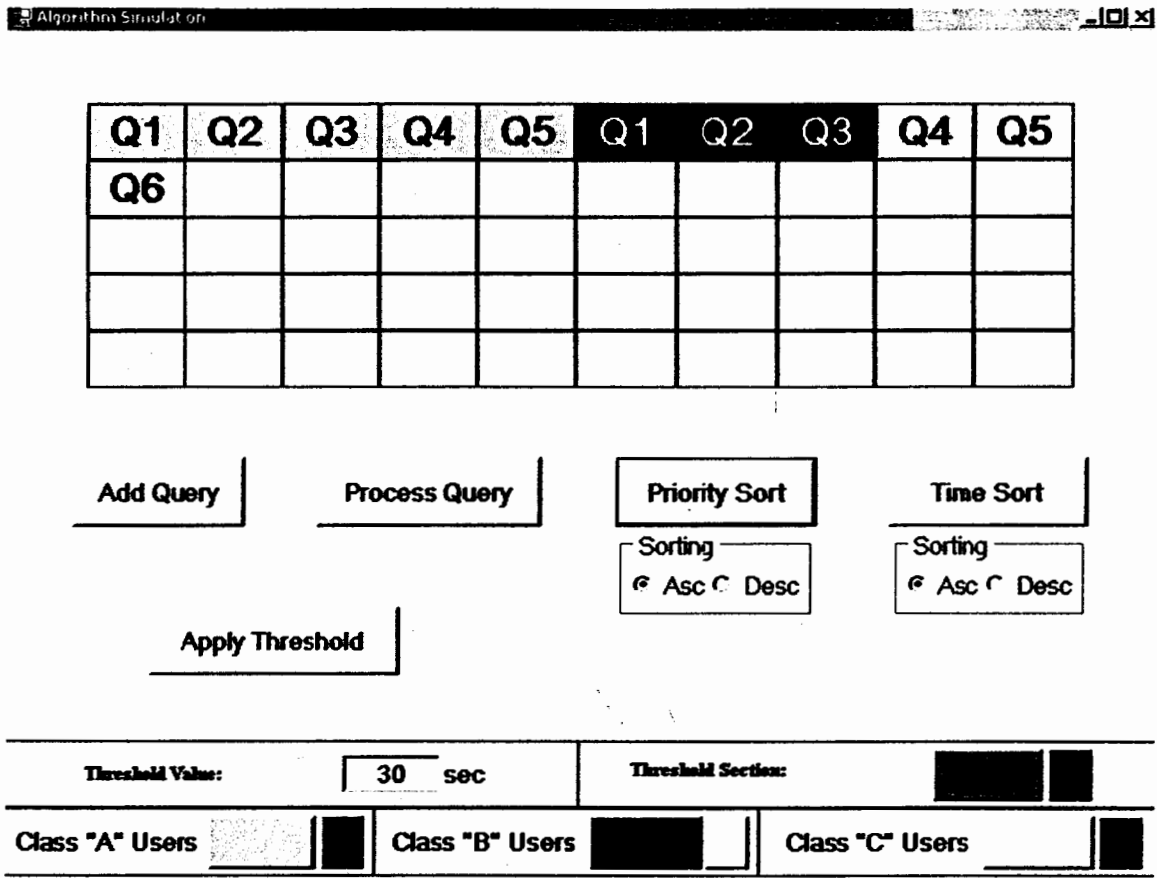


Figure 6-35 After Applying Priority Sort On The Basis Of Classes

In Figure 6-35 all the colors that are from same class are shown together. This indicates that priority sort is being applied here in order to fulfill the algorithm requirement. The users with higher priority level are arranged first depicted with pink then blue and finally lower priority level users queries appear with blue boxes.

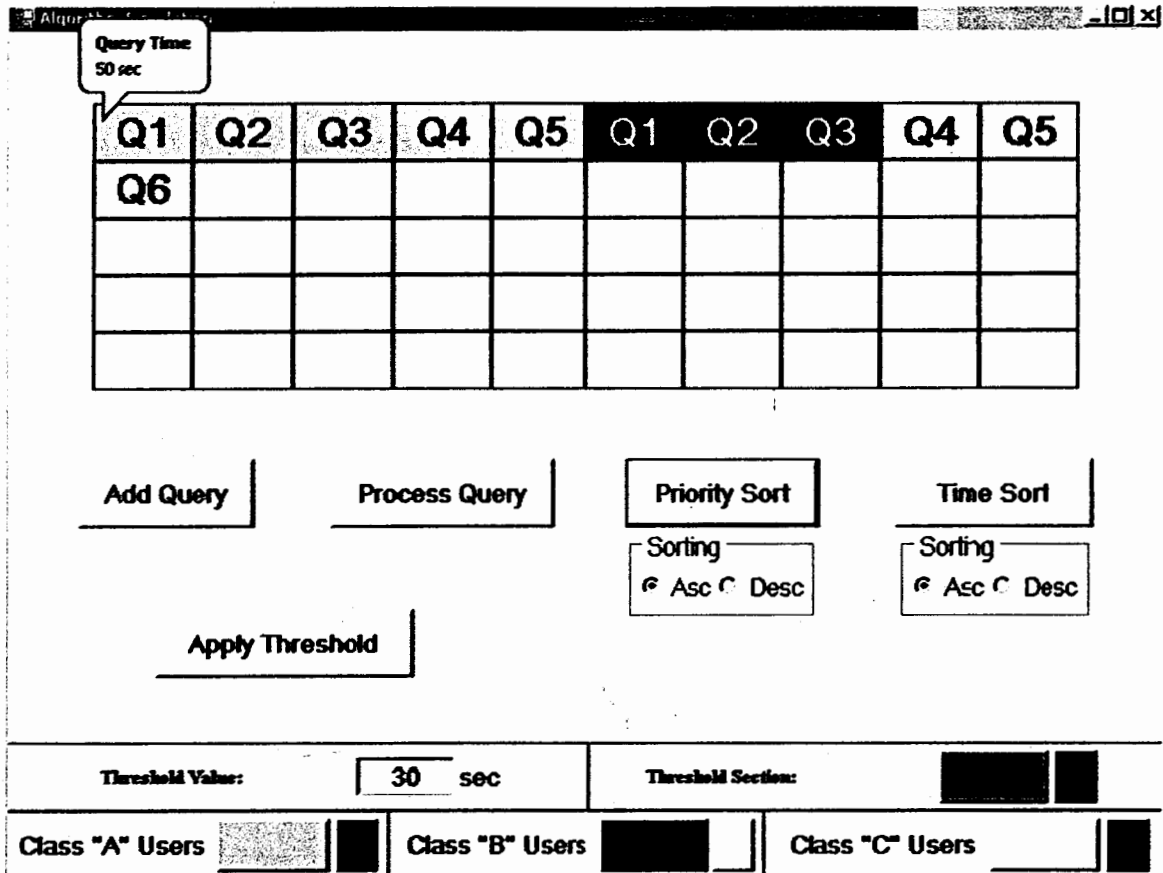


Figure 6-36 Time Of the Very First Query before Applying Time Sort

Figure 6-36 indicates the call out with Query time 50 sec as it may have to change its position according to priority of processing/time.

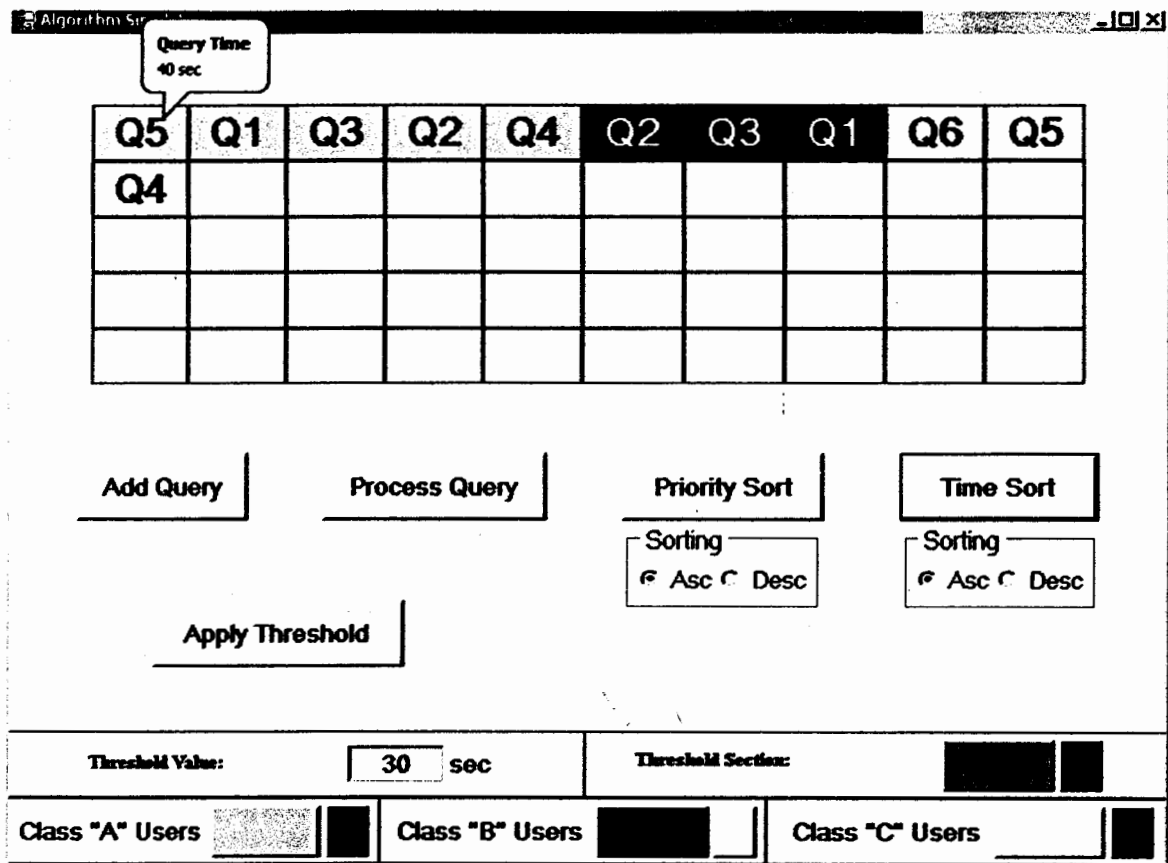


Figure 6-37 After Applying Time Sort

Here Figure 6-37 describes the position of buffer after applying time sort. Now the query with lowest processing time under same color of class is placed first compare to others. As shown query Q5 has processing time 40 seconds which is the lowest in the class of priority. Next query Q1 has processing time 50 and higher.

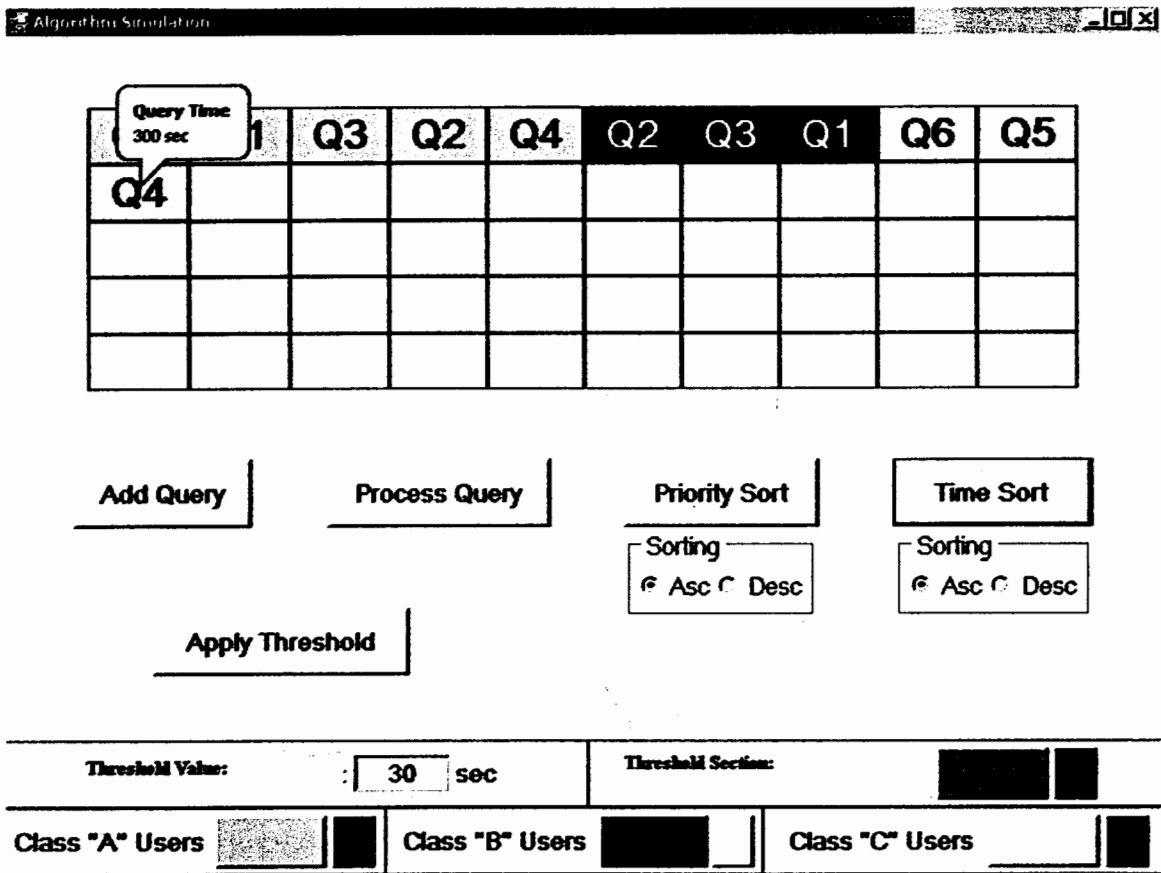


Figure 6-38 Query with High Execution Time Is Placed In Last

This Figure 6-38 is also taken after applying time sort. The query from class “C” that is placed in last of queue has higher reprocessing time. The call out indicated query highest processing time that is 300 sec.

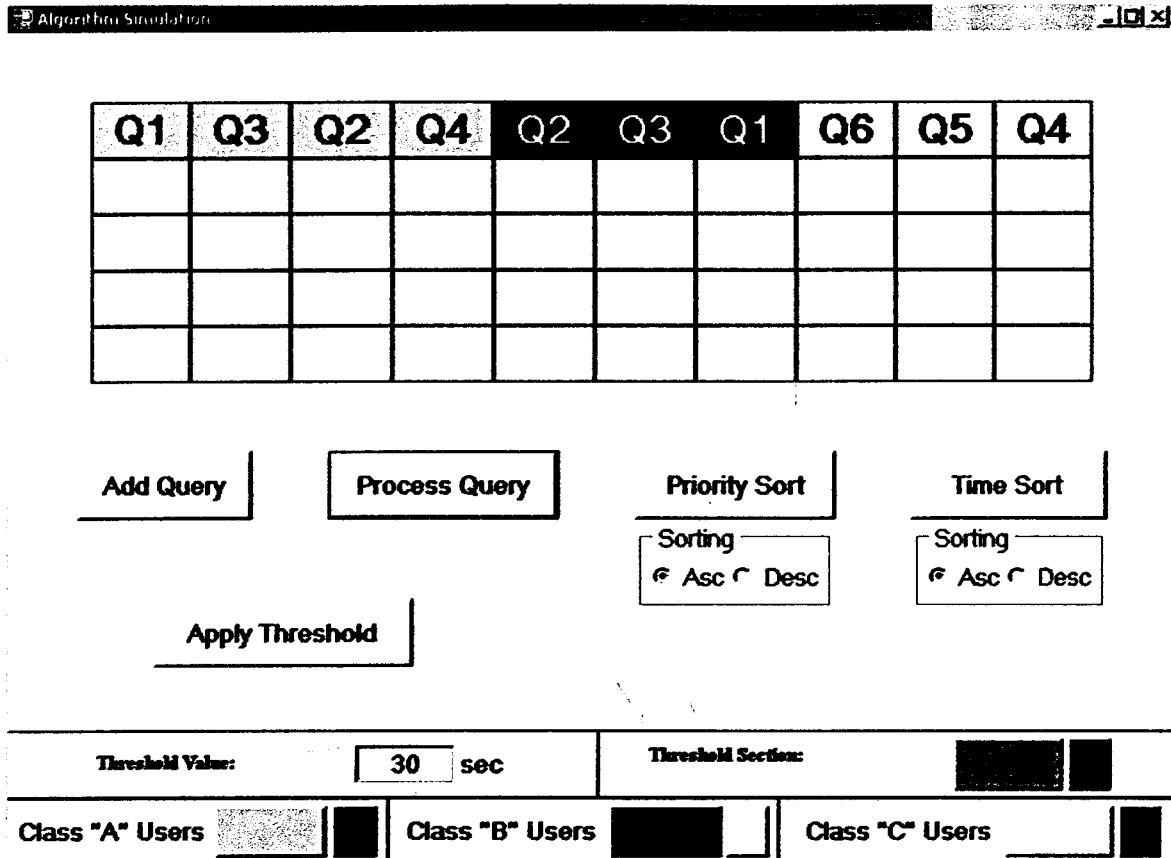


Figure 6-39 During Processing, the Queries Are Processed Accordingly

This Figure 6-39 is same as above but here no popup is shown. It simply indicates the current position of buffer after applying priority and time sorts.

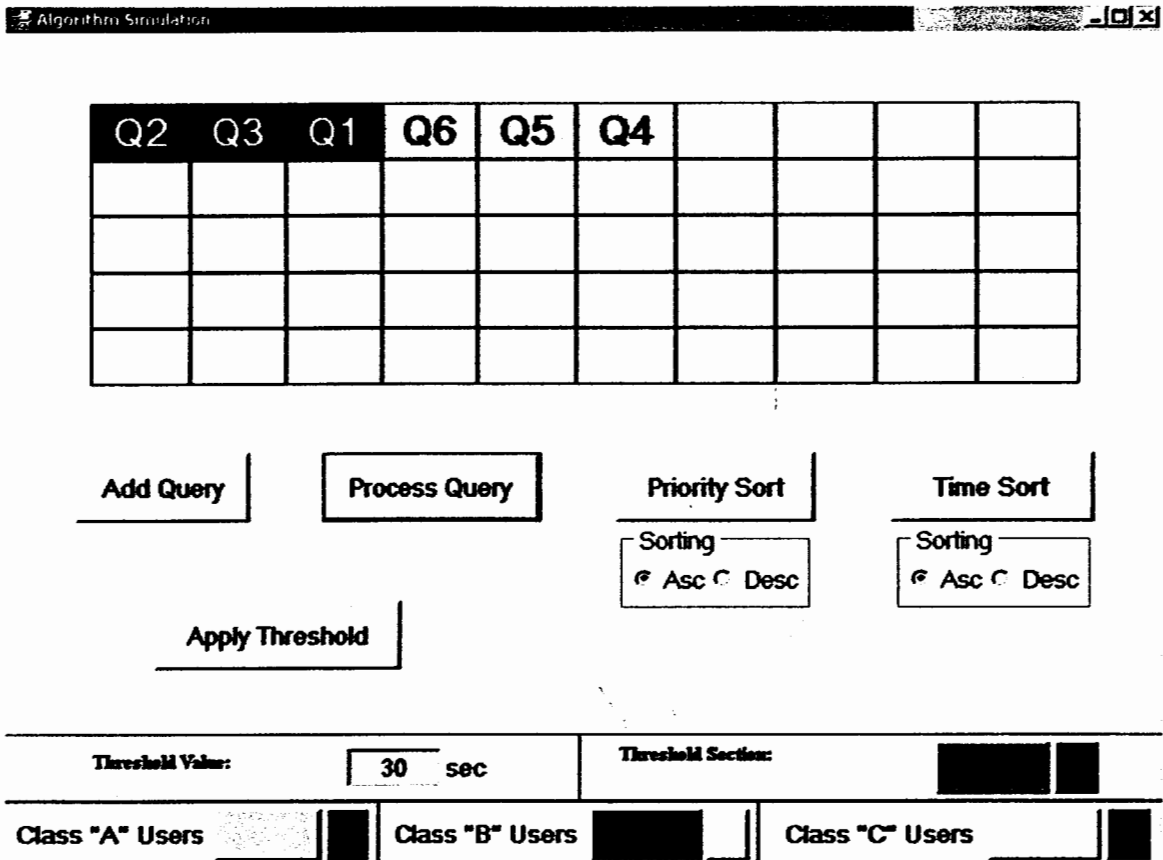


Figure 6-40 queries From Class 'A' Are Answered

In Figure 6-40 the buffer is left up with six queries. The figure depicts the queries with higher priority level are processed first. The queries under the same class "A" are processed according to their processing time. The query with low processing time under the same class is answered first.

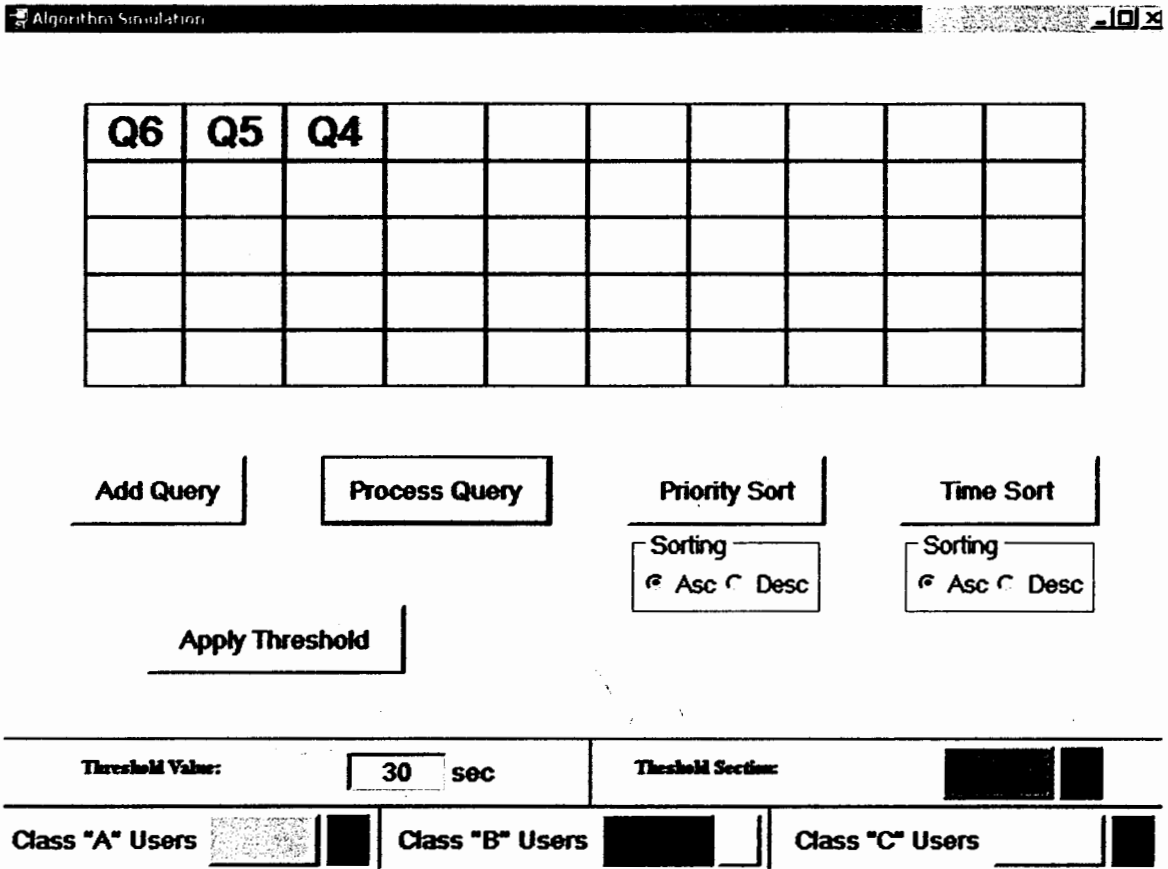


Figure 6-41 Queries From Class 'B' Are Answered

In this Figure 6-41 the remaining three queries are shown. The queries that are from Class "B" having blue color are processed as they belong to class "B" second highest priority.

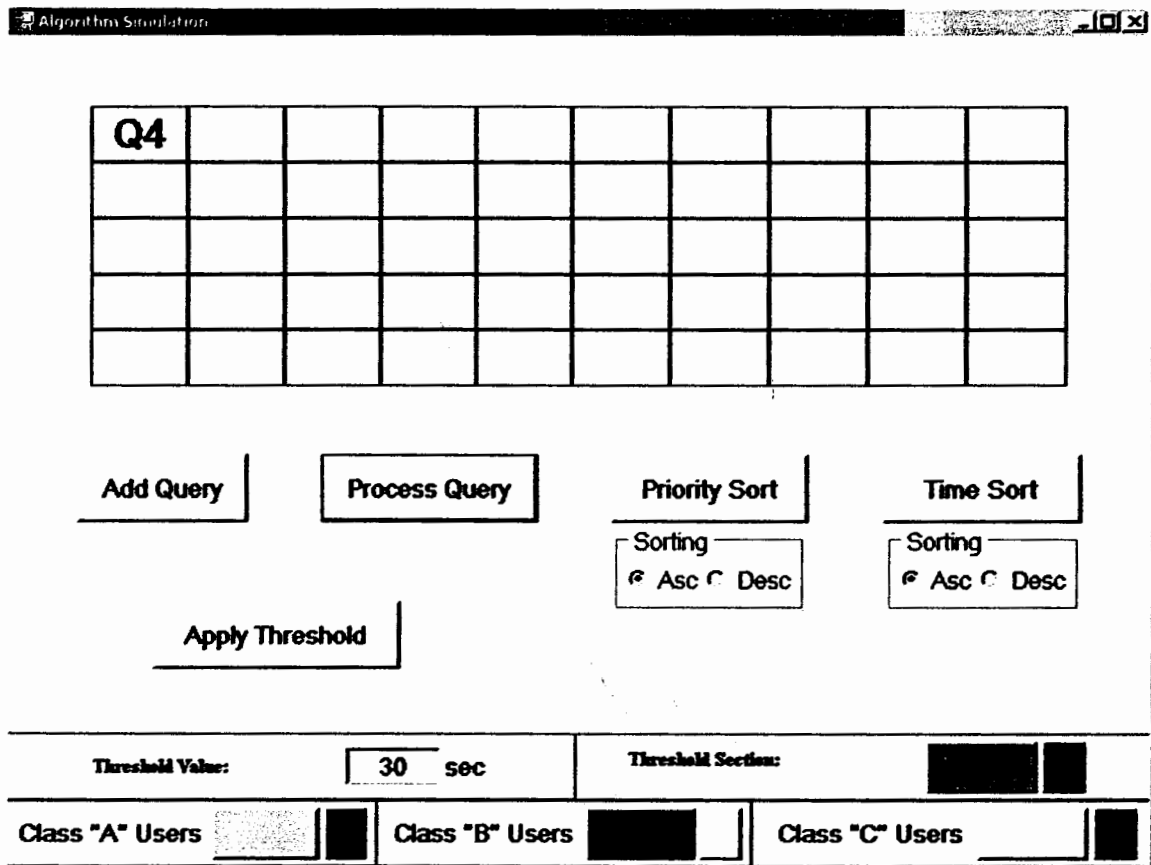


Figure 6-42 Queries from Class 'C' Are Processing

The Figure 6-42 shows only one query as here the buffer has to process Class "C" queries. So from remaining three Class "C" queries this is the query that is going to process in last due to higher processing time 300 sec.



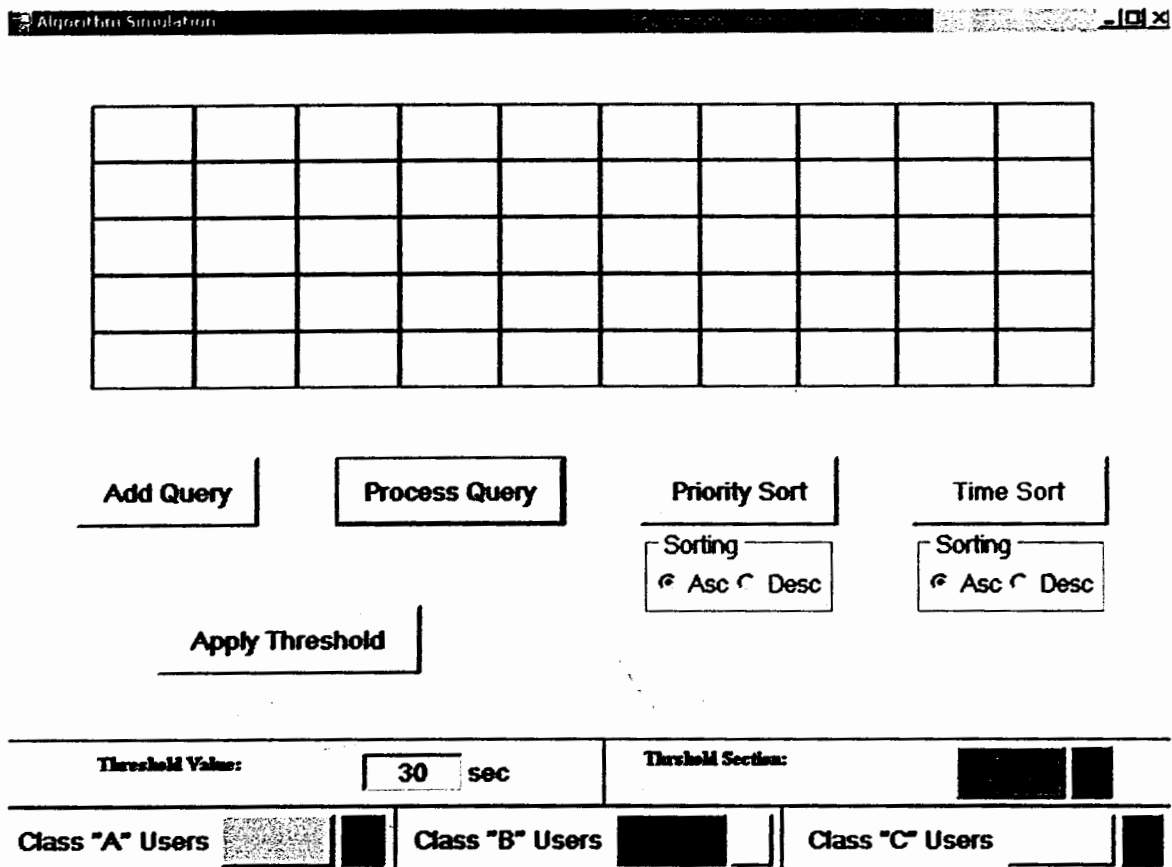


Figure 6-43 All The Queries Are Processed, Buffer Is Empty.

The Figure 6-43 depicts the final condition of buffer where all queries have been processed according to their priority of classes and processing/time. Empty buffer indicates all queries are answered successfully according to proposed algorithm.

## 6.4 Summary

In this chapter the case study is briefly described. Detailed examples are shown with graphics, the results before and after compilation are also discussed. Here the simulation of proposed priority algorithm is also elaborately covered. The simulation tool and its application are also described. The screen shots are taken and pasted here to show the internal working of proposed work.

## **7 CONCLUSION AND OUTLOOK**

## 7 Conclusion and Outlook

This chapter is about the present working and future recommendations.

### 7.1 Introduction

Data warehouse is basically used for decision making purposes. The different manipulation techniques are used to access data in a data warehouse. The data from multiple diverse sources are gathered to data warehouse, to ensure consistency, data cleaning and integration techniques are applied. The variety of tools like OLAP, ETL, data mining and data visualization provide the way to make multidimensional analysis on an applied query. The query tools provide powerful decision making that leads to beneficial outcome. In data warehouse the related historical data is stored. The fresh data is provided as a supplement, as data is not overwritten. The data in a data warehouse is extracted from different operational sources and loaded to a data warehouse after applying cleansing techniques. So the data in a data warehouse is relevant and clean.

To make successful decisions the organization of queries is also very important. The architecture of a data warehouse plays an important role for accurate decision. The users from different classes may also interact with data in a data warehouse. The users that belong to different classes are tackled accordingly. The proposed work presented a new model for a data warehouse. The model's main focus is to organize the queries issued by multiple classes of users and process them according to their assigned priority level.

The proposed architecture introduced a new phase (PAP) Priority allocation processes is also described in detail in previous chapters. The algorithm and simulation is also done consequently. The objective to propose such architecture is to organize multiple users in order and processes the initiated queries.

## 7.2 Achievements

The data manipulation takes place in appropriate order. The users are assigned priority levels. The users with higher priority level dealt with first, so they may not suffer long wait. The users with any one class but have low complexity of query is also preferred. A threshold is maintained in order to tackle the queries that have low processing time and they are from any one of the defined class. The time is allocated to each class of user to generate queries. The processing times of queries are automatically provided to processor according to their complexity i.e. (joins) the queries with more joins are more complex. Query is answered back to each user in an organized way.

## 7.3 Improvements

The concept of a data warehouse is new concept. The typical architecture of a data warehouse is at an abstract level. In reference paper [3] the web warehouse architecture is discussed also. The discussed papers are thoroughly studied in order to use their concept and build a new improved architecture. In this proposed work the queries are organized in well structured way. The set of users that belongs to different hierarchical background are treated according to assigned priority levels. The data is gathered from heterogeneous sources are gathered and distributed on related data marts.

## 7.4 Future Recommendations

The Priority Allocation Process (PAP) under Priority Allocation Layer (PAL) is new concept in this era and is next step from data warehouse architecture. The simulation of proposed algorithm is carried out in C# and through case study Step by Step processing of algorithm is also shown in previous chapter. In future a complete team of software designers, funds and resources are required to practically apply the

proposed algorithm. Also some new phases such as Query Optimized Layer (QOL) can be added in order to make the architecture more users friendly. The research study may continue to work for query optimization and user friendly environment.

## **7.5 Summary**

The last chapter seven is about the conclusion and future work. In this chapter the summary of over all proposed research study is discussed. The future work is also described here. In future the implementation of this simulation may be done with team work and related resources.

## **References and Bibliography**

- [1].Qian Zhou, XIAO Qing, “The Study on Data Warehouse Modeling and OLAP for Highway Management” International Conference on Measuring Technology and Mechatronics Automation 2009.
- [2].Qian Zhou, Qing Xiao, “A Study of Building Data Warehouse Based on Making Use of Its System Structure and Data Model”, 2009 International Conference on Measuring Technology and Mechatronics Automation.
- [3].Saif Ur Rehman Malik, Maqbool Uddin Shaikh, “Web Warehouse: Towards Efficient Distributed Business Management”.
- [4].Henning Baars and Hans-George Kemper, “Management Support with Structured and Unstructured Data An Integrated Business Intelligence Framework”, *Information System Management*, Vol.25,No.2,2008, pp.132-148.
- [5].Michel Schneider, “A general model for the design of data warehouses”, *International Journal of Production Economics*, Vol.112, No.1, 2008, pp.309-325.
- [6].Karen C. Davis, Yeol Song, “Data Warehousing and OLAP”, *Journal of Database Management*, Vol.17, No.1, 2006, pp.1-3.
- [7].Jane Zhao ,“Designing Distributed Data Warehouses and OLAP Systems” , *Journal of Systems and Software* May 2005, Volume 79 , Issue 5

- [8]. Hua-long Zhao, "Application of OLAP to the Analysis of the Curriculum Chosen by Students", *Anti-counterfeiting, Security and Identification*, 2008. ASID 2008.2nd International Conference in Guiyang.
- [9]. Nilakanta. Sree, Scheibe. Kevin, "Dimensional issues in agricultural data warehouse design", *Computers and electronics in agriculture*, Vol.60, No.2, 2008, pp.263-278.
- [10]. Timon C. Du, Jacqueline Wong, "Designing Data Warehouses for Supply Chain Management", 2004 Proceedings of the IEEE International Conference on E-Commerce Technology.
- [11]. Nazih SelmouneZaia Alimazighi "A decisional tool for quality improvement in Higher Education", 3<sup>rd</sup> International Conference on Information & Communication Technologies, Vol.4.
- [12]. Jan Chmiel, Tadeusz Morzy, Robert Wrembel, "Multiversion join index for multiversion data warehouse", *Information and Software Technology*, Vol.51, No.1, 2009, pp.98-108.
- [13]. ZHOU Qian, SUN Li-jun, "The Architecture and Design Strategy for Data Warehouse of Highway Management", 2009 Second International Conference on Intelligent Computation Technology and Automation.
- [14]. Thomas Connolly, Carolyn Begg, *Database Systems: A Practical Approach to Design, Implementation and Management*, 4th Edition, Addison-Wesley, 2003
- [15]. Data Warehouse Evolution Framework, <http://syrcondis.citforum.ru/2007/13.pdf>, Accessed Data: Oct 11, 2008



- [16]. Z. Bellahsene, Schema, "Evolution in Data Warehouses", *Knowledge and Information Systems*, Springer-Verlag, pp 283-304, 2002
- [17]. S. Chen, X. Zhang, E.A. Rundensteiner, "A Compensation-Based Approach for Materialized View Maintenance in Distributed Environments", In Computer Science Technical Report, Worcester Polytechnic Institute, Worcester, MA, USA, 2004
- [18]. E.A. Rundensteiner, A. Koeller, X. Zhang, "Maintaining Data Warehouses over Changing Information Sources", *Communications of the ACM*, Volume, 43, New York, NY, USA, pp 57-62, 2000
- [19]. Data Warehousing and OLAP, [www.cs.uh.edu/~ceick/6340/dw-olap.ppt](http://www.cs.uh.edu/~ceick/6340/dw-olap.ppt), Accessed Data: Nov 25, 2008
- [20]. Online Analytical Processing (OLAP) and Data Warehousing, [academic2.bellevue.edu/~jwright/CIS605/Lesson10/OLAP.ppt](http://academic2.bellevue.edu/~jwright/CIS605/Lesson10/OLAP.ppt) Accessed Data: Dec 5, 2008
- [21]. Jeffrey A. Hoffer, Mary B. Prescott, Fred R. McFadden, *Modern database management*, Sixth Edition, Pearson Education Publishers, Singapore
- [22]. Data Cleaning, <http://research.microsoft.com/DMX/DataCleaning/>, Accessed Data: Sep 19, 2008
- [23]. S. Chaudhuri, U. Dayal, "An Overview of Data Warehousing and OLAP Technology", *SIGMOD Record*, volume.26, Number.1, pp. 65-74, 1997.
- [24]. Data Visualization, <http://web.cs.wpi.edu/~matt/courses/cs563/talks/datavis.html>, Accessed Data: Oct 16, 2008.
- [25]. A, Abdullah "Data warehousing" Virtual University of Pakistan'2002.
- [26]. Uma Sekaram, "*Research Methods for Business*," 3rd Edition, United State of America, John Wiley and Sons Inc, 2000.
- [27]. Qualitative and Quantities Research Methods, [www.ics.heacademy.ac.uk/events/presentations/271\\_slides3Oct.ppt](http://www.ics.heacademy.ac.uk/events/presentations/271_slides3Oct.ppt), Accessed Date: Nov 24, 2008.

- [28]. Norman K. Denzin, Yvonna S. Lincoln, "*The SAGE Handbook of Qualitative Research*," Sage Publications, 2005.
- [29]. Qualitative Research: An Introduction,  
[www.staff.ncl.ac.uk/david.harvey/AEF801/MBQualMeth.ppt](http://www.staff.ncl.ac.uk/david.harvey/AEF801/MBQualMeth.ppt), Accessed Date:  
Nov 25 2008.
- [30]. Qualitative Research Methodology  
[https://idenet.bth.se/servlet/download/news/26148/winter\\_070329\\_qualitative\\_research\\_methodology.pdf](https://idenet.bth.se/servlet/download/news/26148/winter_070329_qualitative_research_methodology.pdf), Accessed Date: Dec 5, 2008.
- [31]. Research Methodologies,  
[www.cs.uvic.ca/~mstorey/teaching/infovis/course\\_notes/researchmethods.pdf](http://www.cs.uvic.ca/~mstorey/teaching/infovis/course_notes/researchmethods.pdf),  
Accessed Date: Nov 27, 2008.
- [32]. Sourav S. Bhowmick, Wee-Keong Ng, and Ee-Peng Lim, "Information Coupling in Web Database," Springer-Verlag Berlin Heidelberg, pp. 92-106, 1998.
- [33]. S. S. Bhowmick, S. K. Madria, W.-K. Ng., and E.-P. Lim, "Web Warehousing: Design and Issues," Springer-Verlag Berlin Heidelberg, pp. 93-105, 1999.
- [34]. W.K.Ng, K.-P.-Lim, C.T.Huang, S.Bhowmick, F.Q.Qin, "Web Warehousing: An Algebra for Web Information," *In Proceedings of the IEEE International Conference on Advances in Digital Libraries*, Santa Barbara, California, 1998.

- [35]. S.S. Bhowmick, W.-K. Ng and E.-P. Lim, "Information Coupling in Web Databases," In *Proceedings of the 17th International Conference on Conceptual Modelling (ER'98)*, Singapore, 1998.
- [36]. S. S. Bhowmick, S. K. Madria, W.-K. Ng and E.-P. Lim, "Data Visualization in a Web Warehouse," Springer-Verlag Berlin Heidelberg, p p.68–80, 1999.
- [37]. S. S. Bhowmick, S. K. Madria, W.-K. Ng, and E.-P. Lim, "Web Warehousing: Design and Issues," Springer-Verlag Berlin Heidelberg, p p . 93– 105, 1999.
- [38]. W. K. Ng, E.-P. Lim, C. T. Huang, S. Bhowmick and F. Q. Qin. "Web Warehousing: An Algebra for Web Information," In *Proceedings of IEEE International Conference on Advances in Digital Libraries (ADL'98)*, Santa Barbara, California, April 22–24, 1998.
- [39]. Sourav S Bhowmick, Ang Kho Kiong and Sanjay Madria, "Formulating Disjunctive Coupling Queries in a Web Warehouse," Elsevier Science, 2002.
- [40]. Mukesh Mohania, Yahiko Kambayashi, A. Min Tjoa, Roland Wagner, and Ladjel Bellatreche, "Trends in Database Research," Springer-Verlag Berlin Heidelberg, pp. 984–988, 2001.

## Web pages

Data Warehouse Evolution Framework, <http://syrcondis.citforum.ru/2007/13.pdf>,  
Accessed Data: Oct 11, 2008

Data Warehousing and OLAP, [www.cs.uh.edu/~ceick/6340/dw-olap.ppt](http://www.cs.uh.edu/~ceick/6340/dw-olap.ppt), Accessed  
Data: Nov 25, 2008

Data Visualization, <http://web.cs.wpi.edu/~matt/courses/cs563/talks/datavis.html>,  
Accessed Data: Oct 16, 2008.

Data Cleaning, <http://research.microsoft.com/DMX/DataCleaning/>, Accessed Data: Sep  
19, 2008

Online Analytical Processing (OLAP) and Data Warehousing,  
[academic2.bellevue.edu/~jwright/CIS605/Lesson10/OLAP.ppt](http://academic2.bellevue.edu/~jwright/CIS605/Lesson10/OLAP.ppt) Accessed Data: Dec 5,  
2008

# **Acronyms**

## List of Acronyms

<b>ETL</b>	<b>Extract Transform &amp; Load</b>
<b>OLAP</b>	<b>On-line Analytical Processing</b>
<b>OLTP</b>	<b>Online-Transaction Processing</b>
<b>DDW</b>	<b>Distributed Data Warehouse</b>
<b>DWH</b>	<b>Data Warehouse</b>
<b>DW</b>	<b>Data Warehouse</b>
<b>SQL</b>	<b>Standard Query Language</b>
<b>PAP</b>	<b>Priority Allocation Process</b>
<b>PQ</b>	<b>Prioritized Queue</b>
<b>RAM</b>	<b>Random Access Memory</b>
<b>PL</b>	<b>Presentation Layer</b>

