The impact of using data warehouse on Manpower Employment Decision Support Systems

By

Muftah A. El-Morabet Onaiba

Supervisor: Dr. Faraj A. El-Mouadib

A thesis submitted to the Department of computer Science

In partial fulfillment of the requirement for the degree of

Master of science

Al-Tahaddi University

Faculty of Science

Department of Computer science

Sirte, G. S. P. L. A. J.

Academic year 2007/2000

التاريخ 23/7/27
الموافق
الرقم الإشاري 1/534/ج

Faculty of Science
Department of Computer Science

Title of Thesis

The Impact of Using Data Warehouse on Manpower Employment Decision Support Systems

By

Muftah A. Onaiba

Approved by:

Dr.Faraj A. El-Mouadib
(Supervisor) .....

Dr. Zakaria Suliman Zubi
(External examiner) .....

DrIdris S. El-Feghi
(Internal examiner) .....

Countersigned by:

Dr. Ahmed Farag Mhgoub

(Dean of faculty of science)

Abstract

Providing data for supporting decision-making process is the main role of Data Warehouse (DW) because it is a repository system of data either detailed or aggregated. The concept of DW has added new benefits by improving and expanding the scope, accuracy, and accessibility of data. DW is a process of assembling data from various sources, and stored in one location after some preprocessing operations such as; cleaning, integration, reduction and transformation in order to get an overview of all of the data that concerns some given. DW can change the nature of decision supporting process because it's the link between applications and data (which was scattered in separate database but now is unified).

The objective of this thesis is to conduct a study for data warehouse, and its impact on information systems especially on Decision Support Systems (DSS) and how decision support process needs to be improve data quality. Our case study contains real data. Through this research we improves the importance of the data warehouse, and why it was chosen on the process of decision making and decision support, because it has a high speed and it saves time and effort used to produce reports needed in decision making.

The case study in this work is for one of the most important sectors in the Great Jamahiriya. It is for the Manpower Employment Agency. This case study is for directing national seekers to different jobs in different sectors and companies, etc. in the Great Jamahiriya. Our work deal with several

databases that differ in nature and huge volumes of data from different sources.

We have shown that data warehousing is an essential approach for decision making based on different data sources, it has been applied to real data and have shown to be effective and efficient in (DSS).

LIST OF TABLES

Table		Page
1.1	Features between data dart & DW.	10
2.1	Major distinguish features between OLTP and OLAP	20
4.1	Misurata popularity database	39
4.2	GPCM fact table structure	47
4.3	GPCM dimension tables structure.	48

LIST OF FIGURES

Figure		Page
1.1	Data mining as a process of knowledge discovery	8
2.1	Data Warehouse Architecture	18
2.2	Basic OLTP system	19
3.1	Star schema of data warehouse	29
3.2	Snowflake schema of data warehouse	30
3.3	Fact constellation schema of data warehouse	30
3.4	The concept hierarchy for time dimesion	34
3.5	Roll up and drill down operations	35
3.6	Slice and dice operations	35
4.1	The basic methods for data cleaning	41
4.2	Sex table (DBF file)	42
4.3	Sex table (MDB file)	42
4.4	Sectors table (DBF file)	42
4.5	Sectors table (MDB file)	43
4.6	Knowledge base for solving some inconsistency	43
4.7	GPCM Metadata	44
4.8	Knowledge base to convert flat file to temporary MDB file	44
4.9	A concept hierarchy for attribute address in Misurata city	45
4.10	Data cube aggregation Reduction	46
4.11	A concept hierarchy for attribute address in Jamahiriya	46
4.12	GPCM star schema diagram	48
4.13	GPCM Multidimensional data model	49
4.14	Sample of statistical data for all cities	51
4.15	Sample of statistical of directed applicants for all cities	51
4.16	Total number of applicants according to the marital status, gender in year	52
4.17	Statistical data for Misurata city in the year 2006 – OLTP system	53
4.18	Statistical data for all cities in the year 2006–Data Warehouse System	53
4.19	Statistical data in the year 2006 – Data Warehouse system	54

Figure **Page**

4.20 Graph of resulted Statistical data for applicants who did finish the military service 54
4.21 Graph of resulted Statistical data for applicants who not finished the military service 55
4.22 Total number of directed applicants to deferent sectors for all cities, from the year 2000 to 2006 56
4.23 Total number of searchers, according to their preferred sectors for all cities, from the year 2000 to 2006 57
4.24 Graphical representation for the number of applicants and directed applicants in all cities from the year 2000 to 2006. 57
4.25 Graphical representation for the number of researchers and directed applicants in all cities from the year 2000 to 2006 58

INTRODUCTION

CHAPTER 1

Introduction

Since the use of computers in business world, data collection has become one of the most important issues due to the fact that the available data contain valuable knowledge. Such volumes of data have been stored in many types of databases.

Since the late 1960s, the area of Database Management System (DBMS) has emerged in response to the need of many organizations to manage and benefit from these huge amounts of data, which have been collected and generated by these organizations.

The concept of tabular oriented relational database was introduced in the early 70s by Dr. Ted Cod. The relational database model has received much attention and developments due to its simple mathematical basis (the set theory). Commercially viable, relational database management systems were available in the market by early 80s.

Although, in the early 1980s, most of the commercial database systems were based on relational models, several alternatives in database models were also proposed. One of those alternatives for relational database was the semantic data model. Another alternative is the Object Orientation (OO) model, the purpose behind both the development of semantic data models and the development OO models is to model the real world as closely as possible. In OO data modeling, each real world entity of problem domain is represented by a set of objects with relations and operations. Each object consists of part of objects or sub-objects that relates objects to each other (relation representation).

In the early 1990, relational database management systems were more popular than hierarchal and network database management systems. Some of the increased advantages of the relational database management systems were its functionality and flexibility and the use of cache up in performance. In current database management systems, object oriented techniques become more popular because of its encapsulation of the data and the functions being performed on these data.

Recently, advances in technology have been revealing new applications of database systems, such as pictures, video clip, and sound message; can now be stored by multimedia database. In addition, maps, weather data, and satellite images can be stored and analyzed by Geographic Information System (GIS).

As data size increases, the needs for more control and information retrieval also have increased. These increases have led to the development of Data Warehouses (DW), Data Mining (DM) systems and Knowledge Discovery in Database (KDD) systems. "Data warehouse (DW) and On Line Analytical Processing (OLAP) system are used in many companies to extract and analyze useful information from very large database for decision making. Real-time and active database technologies are used in controlling industrial and manufacturing processes. Furthermore, database search techniques are being applied to the World Wide Web (WWW) to improve the search for information that is needed by users browsing through the internet." [1]. New generation of integrated information systems have been appealed to computer users since the year of 2000.

1.1 Data and database

1.1.1. Data

In isolation, data are collection of raw facts. Data are descriptive qualities in the sense that it characterizes an entity or values of variables. The values of data must have some meaning. Data should be categorized into one of the following categories:

- Tangible things such as; Materials, Vehicles, Computers, Automobiles, and etc.
- Roles such as; Applicants users, Suppliers, Employees, Decision makers and etc.
- Events such as; Orders, Sales, Contracts, Trips and etc.
- Places such as; Store, Agency, Sale offices and etc.

1.1.2. Database

Any large size of data can be called a database with the emphases that these data are related and meaningful. "A database is a collection of related data. By data we mean Known facts that can be recorded and that have implicit meaning."[4].

Consider, for example, the names, telephone numbers, and addresses of people that a person knows (i.e. relatives or friends) which may be kept in a telephone book, index cards or a computer file in form of **ACCESS** or **EXCEL**. This collection can be considered as a database even though it is simple in structure and relatively small in size. Computer database comes as a part of a system known as Database Management Systems (DBMS)."A Database Management Systems (DBMS) is a collection of programs that enables users to create and maintain a database. The DBMS is hence a general purpose-

software system that facilitates the processes of defining, constructing, and manipulating database for various applications."[4].

1.2 Information and information system

1.2.1 Information

Information is the raw data after performing some operations (some processing) on it. The results of such operations on the data should be useful and beneficial.

1.2.2 Information System (IS)

Information system is a system that controls and organizes the data, which is needed by an establishment system in the sense that it maintains the needed of various facts and figures. An IS consists of a number of subsystems where each component carries out certain function and all of them are integrated into one to accomplish the over all task. An information system can be manual, computerized or a combination of both. "Information system is an arrangement of people, activities, data, networks, and technology, that are integrated from the purpose of supporting and improving the day-to-day operations in a business, as well as fulfilling the problem solving and decision making information needs of business managers."[18].

1.3 Decision Support Systems

Decision Support Systems (DSS) are a special type of computerized information systems that are used by cooperation's management in order to facilitate decision-making activities. Decision Support Systems are an interactive software based systems that collect and document information and knowledge to help decision makers gain wider prospective of their business or organization.

1.4 Data mining

1.4.1 Data mining overview

Today's databases size can range into the terabytes, more than 10^{12} bytes of data. These huge amounts of data hide strategic importance information. But when there are so many trees, how do we draw meaningful conclusions about the forest? The most recent answer is data mining, which is being used both to increase revenues and to reduce costs in finding knowledge. The potential returns are enormous. Many organizations worldwide are already using data mining to locate and appeal to higher-value customers, to reconfigure their product offerings to increase sales, and to minimize losses due to error or fraud. To find information from data that is not so obvious, traditional querying techniques can be used for that purpose. For example, in The General People's Committee of Manpower (GPCM), we can use queries to find the number of applicants who appointed at a certain sectors in a certain period of time per popularity or city. The knowledge you we looking for is already hidden in the database but it is hard to find among so much data. On the other hand, Data Mining (DM) takes finding the knowledge one step further by seeking hidden relationships among the data, discovers new trends, consider the reasons for specific events, or expect how your data works. For instance, DM can be used to discover why certain products are sold well with some other specific ones. In addition, it can help you expect the sales for the near future. So we can conclude that data mining aims to enable a corporation to improve its marketing, and sales operations. When we extract gold from sand, we call this process gold mining but not sand mining, by the same analogy we should refer the term "Data Mining" as knowledge mining. Many authors refer to data mining by other names such as; Knowledge Discovery in Databases

(KDD), but others consider data mining as the main step in KDD as depicted in figure-1.1, although the term data mining is the more popular than the KDD. As it has been mentioned in [5], the KDD process consists of an iterative sequence of the following steps:

- Data integration (where multiple data sources may be combined) [1].
- Data selection (where data relevant to the analysis task are retrieved from the database).
- Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)[2].
- Data mining (an essential process where intelligent methods are applied in order to extract data patterns).
- Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures.
- Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

1.4.2 Data mining definition

There are many definitions of data mining, one of the most common definitions describe data mining as, "data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories." [5]. Another definition of data mining is given in [3], states it as; the search for relationships and global

[1] A popular trend in the information industry is to perform data cleaning and data integration as a preprocessing step where the resulting data are stored in a data warehouse.

[2] Sometimes data transformation and consolidation are performed before the data selection process, particularly in the case of data warehousing.

patterns that exist in large databases that are hidden among the vast amount of data, such as the relationship between patient data and their medical diagnosis.

All above definitions and other ones put stress on the discovery of relationships, patterns and trends in vast amount of data.

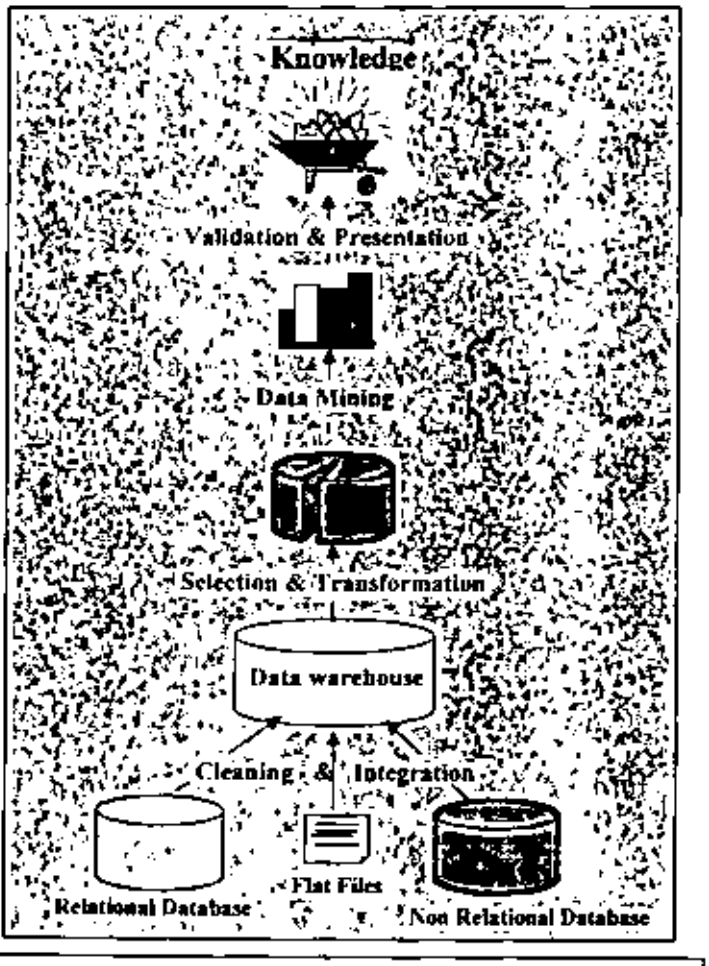


Figure-1.1: Data mining as a process of knowledge discovery

1.5 Data warehouse overview

Data warehouse is a blend of technologies linked together to facilitate an effective integration of operational databases into one environment which allows strategic applications of data. These technologies include relational and multidimensional data management systems, client/server architecture,

metadata modeling, repositories, and graphical user interface. Data warehouse and its practical realization have been developed gradually in the last two decades and have come to its highest degree in the last seven or eight years. The point to be considered as a disadvantage for the slow development is that there are several different definitions to constitute what a DW is?. The idea of DW started in the late 1980s and the main principle was and still today is that the basic business systems are designed to maintain as much information as possible with low rates of failure. The DW is considered to be a repository system for useful data kept by the business systems. DW is designed for easy user access and in turn solves two problems:

1. Enable business decision–makers to get timely access to corporate information.
2. Results consistency when the same data source is used and the reported results are synchronized.

1.6 Data Mart

Data mart is a subset of data warehouse, which holds data that is of a concerned of a specific department or sector of the entire organization. "A data mart is simply a smaller data warehouse. Usually the data in data mart is a subset of the data that is found in an enterprise-wide data warehouse." [9]. For some companies, if the frequent use of the DW is concentrated on a specific part, it will be more beneficial to construct and use a Data mart than DW. Table-1.1 illustrates a comparison between Data mart and Data warehouse.

Property	Data mart	Data warehouse
Scope	Department wide.	Enterprise wide.
Data sources	Few sources	Many sources.
Implementation time	Weeks to Months	Months to years.
Modeling	Star or snowflake schema.	Constellation schema.
Subject	Few selected subjects.	The entire enterprise subjects.

Table-1.1: Features comparison between Data mart and Data warehouse.

Data mart can be built before or after building a data warehouse. Also it is possible to build any of them without the other, but building Data mart has some several advantages:

- Quicker query answering because the data is small in size.
- The development of data mart takes less time.
- Data mart users are less in number than DW users.

1.7 Purpose of data warehouse

The aim of **DW** is to introduce the precise information and have it ready in time. It is the procedure of combining data from different sources of an organization for decision–making purposes. The **DW** is considered to be as an environment. It is an architectural construct for an information repository system for information that provides users with current and historical information which are not provided by traditional operational data stores. Actually, the DW is considered to be organization's cornerstone to carry out effective and efficient information processing. The use of DW within an organization gives it the ability to share the discovery and exploration of important business trends and dependences that might be gone unnoticed. The purpose for the creation and use of DW is analytical and decision-support.

1.8 Thesis objectives

The objective of this work is to study the impact of using **DW** on employment in the General People's Committee of Manpower (**GPCM**) in the view of the quality of the data. The **GPCM** data will be taken up as the case study of this work to demonstrate the reasons and effects of using data warehouse instead of using the traditional databases on decision–making and decision support.

Our **DW** will collect data scattered over several database from different sources in Libya in order to compare the performance and time. The **GPCM** branch of each popularity has its own center, which receives the applicants' data that is processed by a local computerized system.

1.9 Thesis outline

This work is organized as follows:

In this chapter we have introduced the concepts of data and database, Information and Information system, **DSSs**, data mining overview, data warehouse, and data mart.

In the next chapter, an extended overview of data warehouse will be given. The definition, building, characteristics and architecture, also will be discussed. The differences between On Line Analytical Processing (**OLAP**); and On Line Transaction Processing (**OLTP**) will be also mentioned in this chapter. Benefits and drawbacks of data warehouse will be presented at the end of the chapter.

In chapter 3, we will explain the data warehouse design phases, and the business modeling checklist will be elaborated on. Moreover that the multi-dimensional model, data warehouse models, data warehouse components, will be presented, also types of data in **DW** tables, updating data in data

warehouse, will be further explained. Concept hierarchies, and OLAP operations, will be explained with an example.

In chapter 4, the case study of this thesis will be presented. The proposed data warehouse system design and implementation will be further discussed and how this system deal with some real problems which are evident in one of the Libyan national projects namely; General People's Committee of Manpower. The concentration will be particularly on decision support reports produced by the proposed DW system in order to prove that the data warehouse has more flexibility, and higher performance than traditional database systems. In this chapter several statistical charts and tables resulted from our DW system will be used to support our discussion.

Finally in the last chapter a conclusion will be presented and future work trends will be suggested.

BACKGROUND

CHAPTER 2

Background

In this chapter, we review the basic aspects and architecture of data warehouse. The data warehouse contains granular corporate data which can be used for many different purposes.

2.1 Data warehouse definition

The necessary data to support decision-making may be located in several sources with different hardware and software configuration, so it is difficult to access such data in an integrated fashion. The scattered sources of data have posed high costs and reduction of effectiveness in the process of data, therefore the centralized database that represents the concept of data warehouse is considered as solution to this problem since it collects data from different sources, and organizes them to be easily accessible by Decision Support System applications in order to facilitate On Line Analytical Processing (OLAP).

There are many definitions to what a data warehouse is. One of them is: "A data warehouse is a subject-oriented, integrated, time variant and nonvolatile collection of data in support of management's decision making process." [5]. Another definition given in [16] states that; a data warehouse is a structure that links information from two or more databases and brings it into a central repository system to perform some data integration, cleanup, and summarization.

2.2 Building and developing a data warehouse

Due to the fact that analytical systems such as data warehouse receive data from multiple sources and in this process we must identify the sources from

which the data is to be extracted. As it has been mentioned in [9], the simplification of data extraction is accomplished via:

- Choosing the most accessible data sources available.
- Choosing the data with the highest degree of integrity.
- Limit the risk and complexity involved in extracting from multiple sources.

After establishing the data sources, data transformations can be performed to move the data to the data warehousing system. It is important to take into consideration that building a data warehouse is to improve the information quality.

The difference between data in data warehouse and data in operational systems is that the first one is read only, while in the second one, operations such as: addition, deletion, update and backup are needed to keep it up to date. Another difference is that the data in operational systems are the source from which the data is fed to the data warehouse.

The main reason for developing data warehouse is to collect data from various sources and integrate them into a common store which will be analyzed to support decision making within an organization. The following steps are needed in the phase of data extraction:

1. The data must be extracted from multiple heterogeneous sources i.e. relational databases or other form of data stores.
2. Data must be checked for consistency within the data warehouse.
3. The data should be validated by cleansing it.
4. The data should be formatted in such a way as to be fitted into the data model of the data warehouse.

5. The data warehouse must be loaded with data. Monitoring tools to load as well as methods to recover from incomplete or incorrect loads are required.

In the process of building a data warehouse we can start by building small project which aims at fulfilling all the user requirements. If this small project satisfies its aim then we can build a larger project.

2.3 Data warehouse characteristics

Based on the definition mentioned above, the four distinctive features, subject-oriented, integration, time variant and nonvolatile distinguish data warehouse from any other data repository. Here, we would like to explain these features in detail.

1. First, data warehouse is subject-oriented means that the data is organized by subject, such as applicants and directed applicants rather than by applications. This reflects the difference between data warehouse and operational systems. In the former the data is for the purpose of decision-support and in the latter, the data is for the day-to-day applications. In operational systems, data is organized to support a particular application, but in data warehouse the focus is on the analysis of data for decision makers. The data in data warehouse could come from multiple different sources such as relational database, non-relational database and other repository systems.
2. The second feature of data warehouse is the integration, which means that all data from the different sources must be collected together into one coherent unique data store known as data warehouse. In this process some preprocessing techniques are applied to remove noise, correct

inconsistencies, and avoid redundancies. The integration is needed to make the data in a globally accepted manner to be easily accessible by DSS applications, even when original source of the data is structured differently.

3. The third feature of data warehouse is time variant. The data in data warehouse has the element of time associated with it (time series data) such as year, quarter or month. This variation of time is very important in order to perform business analysis. So, the data can be analyzed in the form of past and present trends, and accordingly the results could be compared either in the present or in the past.
4. The fourth feature is nonvolatile, which means that data in data warehouse a static nature, because the data is read only. After the initial loading of the data, new data can be added but not as replacement of the original data. This is called refreshment of the data is for the purpose of keeping the data in data warehouse up to date and to preserve the historical nature. The only operations that are preformed on the data in a data warehouse are loading of additional data and querying.

2.4 Data warehouse architecture

"Many researchers and practitioners share the understanding that a data warehouse (DW) architecture can be formally understood as layers of materialized views on top of each other. DW architecture exhibits various layers of data in which data from one layer are derived from data of the lower layer." [17]. Here, we will illustrate the layers that constitute a data warehouse and as depicted in figure-2.1. **The lowest layer** of the data warehouse architecture is called the data sources layer, which usually consists of the

operational databases. This layer may consist of structured, unstructured or semi-structured data stored in files or other storage system. The data in this layer is extracted to create the data warehouse.

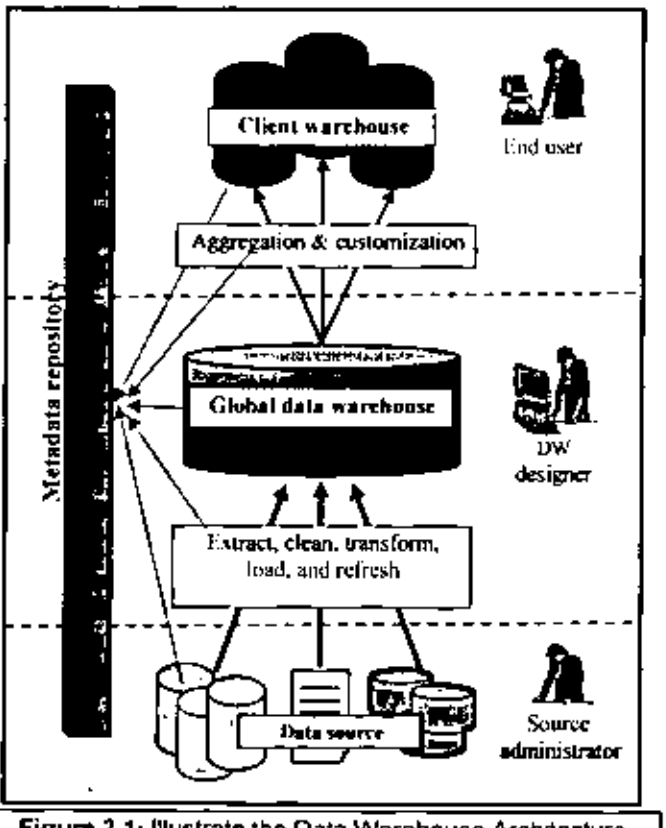


Figure 2.1: Illustrate the Data Warehouse Architecture

The middle layer of the architecture is the global data warehouse. In this layer a historical record of data is stored after being resulted from some operations such as: transformation, integration, and aggregation of detailed data found in the data sources. The data warehouse is populated with clean and homogeneous data.

The top layer is also called client warehouse which contains directly derived data from the global warehouse. This layer is of different kinds such as data marts or the OLAP databases which use multidimensional data structures, or relational database systems.

"All the data warehouse components, processes and data should be tracked and administered from a metadata repository. The metadata repository serves as an aid both to the administrator and the designer of a data warehouse. Indeed, the data warehouse is a very complex system. The volume of recorded data is vast and the processes employed for its extraction, transformation, cleansing, storage and aggregation are numerous. They are sensitive to changes and time-varying. The metadata repository service as a map road which provides a trace of all design choices and a history of changes performed on its architecture and components."[17].

2.5 On Line Transaction Processing (OLTP) vs. On Line Analytical Processing (OLAP).

The purpose of On Line Transaction Processing (OLTP) systems is to allow high concurrency between uses which make it possible for many users to access the same data at the same time. As the name implies, these systems allow transactions to be processed against the data. In other words, these systems control the changes of the data due to some operations such as: insertion, update and deletion during business processes. Figure-2.2 depicts a basic OLTP system:

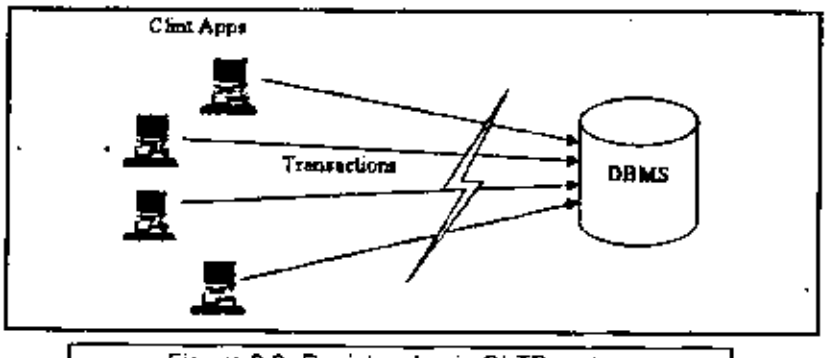


Figure 2.2: Depicts a basic OLTP system.

The figure shows that numerous client applications can access the database to get the needed pieces of information. The broken lines between the client

applications and the DBMS symbolized that these connections can physically be implemented in many different ways.

On Line Analytical Processing (OLAP) is a software technology that allows users to analyze and view data from multiple points of view easily and quickly. OLAP provides dynamic and multi-dimensional support to executives and managers who need to understand different aspects of the data. Activities that are supported by OLAP system are:

- Analyzing financial trends.
- Creating slices of data.
- Finding new relationships among the data.
- Drilling down into sales statistics.
- Doing calculations through different dimensions where each category of data (i.e. product, location, sales numbers, time period, etc.) is considered a dimension.

The OLTP and OLAP have major distinguishing features between them. Table-2.1, demonstrates the distinguishing features between OLTP and OLAP.

Feature	OLTP	OLAP
Users and system oriented	Customer-oriented and is used for transaction and query processing by clerks, clients.	Market-oriented and is used for data analysis by knowledge workers
View	Focuses on current data within enterprise or department without referring to historical data.	Deal with information that originates from different organizations.
DB design	Application-oriented and Entity-Relationship (ER) data model design.	Subject-oriented and star or snowflake model.
Data contents	Manages current data	Manages large amounts of historical data.
Function	Day to day operations	Decision support.
Access	Read/write.	Are mostly read only operations.

Table 2.1: Major distinguish features between OLTP and OLAP

2.6 The benefits of data warehouse

The core benefits of using data warehouse include several aspects that can be summarized as follows:

- Historical information for comparative and competitive analysis as a result of collecting data from several sources over a long period of time.
- Enhancement of the data quality and completeness of it.
- The existence of another data back up source from the original data gives the possibility of recovering from data losses.
- The ability to analyze and execute business decisions based on data collected from multiple sources.
- A data warehouse does not store data values by itself. Data values are collected from different scattered sources.
- Data warehouse can provide savings by the reduction of fraud losses, as a result of intensive analysis.

In addition to the previous benefits, we can increase the benefits from a data warehouse by the use of the proper tools and user training in analysis of the results.

2.7 The drawbacks of data warehouse

Due to the fact that the development of data warehouse projects are much harder to complete successfully than the development of traditional data processing systems. These difficulties are considered to be the drawbacks of data warehouse and these are:

- Analyzing and designing a data warehouse is fundamentally different and more difficult from analyzing and designing OLTP systems.
- The data warehouse tool environment is more complex than the traditional data processing tool environment.

- Data warehouse projects are not business system related, but they are much more related to the analytical system within an organization for decision making.
- The problem of keeping the warehouse in synchronization with the production system is more difficult.
- Data warehouse projects are also more complex and more expensive to develop in comparison to traditional database projects.
- If the data brought over is not large enough then the data warehouse development can lead to failure.

For all of these reasons, data warehouse projects fail much more frequently than traditional systems development projects. To avoid parts of the problems, the project should be done on a small sample of the organization before trying on a large project.

DEFINING AND CREATING THE LOGICAL MODEL OF DATA WAREHOUSE

CHAPTER 3

Defining and creating the logical model of data warehouse

3.1 Data warhouse design phases

Recently, a data warehouse designing methods have been established even though many sources could differ in defining certain terms or use alternative syntax. In our case study, we will focus on the major tasks associated with the data warehouse design processes. The major tasks associated with the design of data warehouse are grouped into the following three phases:

- Defining the business model (conceptual model) phase. In this phase strategic analysis is performed in order to identify the general business process very clearly. Within this phase business requirments analysis are identified and documented in addition to the identification of the dimensions for each business process.
- Creating the dimensional model (logical model) phase. In this phase, dimentional model is derived from the business model. The data warehouse schema elements are definitioned in addition to identifying the relationships between these elements and the data sources for the data are recorded.
- Creating the physical model phase. In this phase the dimensional model is transformed into a physical model. This includes the documentation of data element formats, database size and storage planning, and Indexing.

3.2 Business modeling checklist

The business-modeling checklist involves some tasks which are necessary for the creation of a data warehouse. In the following subsection we elaborate on this checklist.

3.2.1 Strategic analysis

This is an essential task to produce a data warehouse design that is achievable and deliverable within an acceptable time frame. The achievement of such task is via the followng steps:

- **Identify** the business processes that are most important to the organization from the standpoint of the decision makers.
- **understand** the business process by drilling down the analytic parameters.
- **Select** the business process that will be implemented in the data warehouse.

This phase requires the business executives to meet with the others who have the ability to give an overall view of the organization to establish a clear and real understanding of the entire business. The discussion in these meetings should focused around some questions such as:

- What are we trying to achieve?
- How do we measure success?
- How often do we measure success?
- What are the objectives of the organization?
- How can we identify problems?
- How can we know if we are doing well or not?
- What is our busines objectives today?

- What is the impact of not achieving some of these objectives on the organization?

3.2.2 Creating the business model

We can build the business model using the following steps:

- The business model is created by defining business analysis requirements for each selected process. This can be accomplished via several meetings with the business managers and business analysts who are directly responsible for the specific business processes.
- Verification of the data sources that we need to support the business analysis requirements and their existence.

3.2.3 Creating metadata document

Metadata is data or information about the data. Metadata allows users as well as technical administrators to track the structure of the data they are using. "Essentially, examining metadata enhances the end user's understanding of the data they are using."[8]. The metadata can help administrators to guarantee data accuracy, integrity and consistency of the data. The results of studying business modeling process are summarized in the metadata document that should be created in the business modeling phase, these information acts as the crucial resources for the design process. The metadata document will finally contain full detailed descriptions of the sources and composition "structure" of the attributes of the data warehouse. The metadata should provide the following:

- Document the design process.
- Document the development process.
- Provide a record of changes.

- Record enhancements over time. Appandix

3.3 Multidimensional model

Multidimensional modeling is a design technique, which is used in data warehouse environments. "The purpose of multidimensional modeling is to design a database system that supports data analysis and reporting requirements."[9]. Multidimensional models can provide hierarchical views of the data by the use of operation such as roll up and drill down. The Roll up operation can be imaged as moving up in a concept hierarchy producing more general concepts from low-level or basic concepts. On the other hand the drill down operation provides the opposite capabilities of the Roll up operation where more general concepts are broken down into their more specific details concepts.

So, multidimensional models take advantage of inherent relationships existing in the data to populate the multidimensional matrices that are referred to as data cubes. If the matrix contains more than three dimensions it's called hypercubes. "Multidimensional databases are modeled to provide the fastest possible response times to complex queries."[9]. In modeling a multidimensional databases, star schema and snowflake schema are used for that purpose.

3.4 Data warehousing models

To design a relational databases Entity-Relationship data modelling technique is used, where a database schema is built from a group of entities or objects and the relationships between them. This modelling technique is appropriate for online transaction processing systems. On the other hand,

data warehouse requires a subject-oriented schema to facilitate online analytical processing. Multidimensional model is the most popular model for data warehouses. This model can be of the form of star schema, snowflake schema or fact constellation schema.

3.4.1 Star schema model

The star schema consists of one large central table known as fact table and a set of smaller tables called dimension tables. In this schema each one of the small tables represent only one dimension in the fact table which contain information specific to the dimensions. "Star schema has gained a widespread acceptance for data warehousing, and has been used in most data warehouse to represent the multidimensional data model." [12]. In Figure 3.1 the star schema contains a central fact table with keys to each of the six dimensions along with three measures: Number of applicants, Number of seekers, Number of directors, these measures are numerical values and calcuted by the standard functions such as SUM and COUNT. To design a data warehouse the best starting point is the star schema model in which relational rules are constructed using primary keys and foreign keys that link the fact table with dimension tables. "One advantage of the star schema is its use of standard database technology to achieve the power of OLTP."[7]. In addition to the ease of implementation, the star schema have the following advantages:

- It involves fewer tables.
- Very simpler to construct queries.
- The performance queries is quicker than in an other schema types, because fewer joins are required to access data values.

In our case study **GPCM**, there are dimension tables such as; city, congress, sector and education level.

3.4.2 Snowflake schema model

As we have mentioned above, the star schema consists of a single fact table and one table (unnormalized) for each of the dimensions. The snowflake model is an extension of the star schema model where some dimension tables are normalized. "It provides a refinement of the star schema model, where normalization operations are applied on some dimension table." [12]. For example, the congress dimension of figure-3.1, is normalized to show districts of each congress, as depicted in figure-3.2.

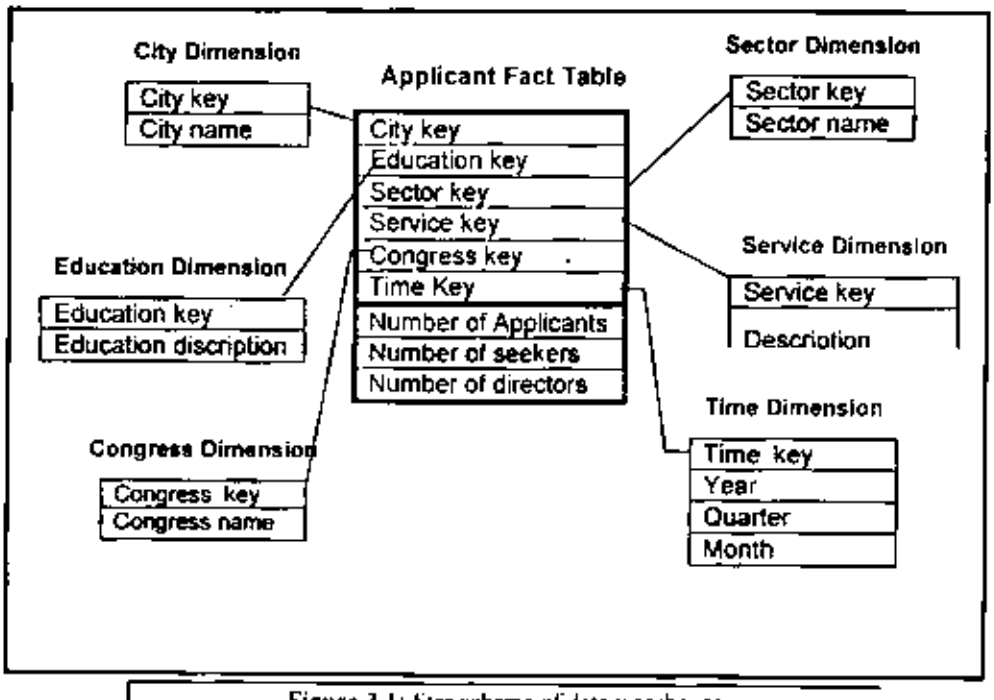


Figure-3.1: Star schema of data warehouse

3.4.3 Fact constellation model

Some complex applications may require complex structure, which may be accomplished by multiple fact tables and share dimension tables. Therefore, the answer to such problem would be to use fact constellation model. The

fact constellation model consists of a set of fact tables that share some dimension tables. The structure can be imagined as a star schema structure surrounded by other star schemas, giving shape of collection of star schemas. This type of structure is also called galaxy schema. Figure-3.3 depicts an example of fact constellation schema.

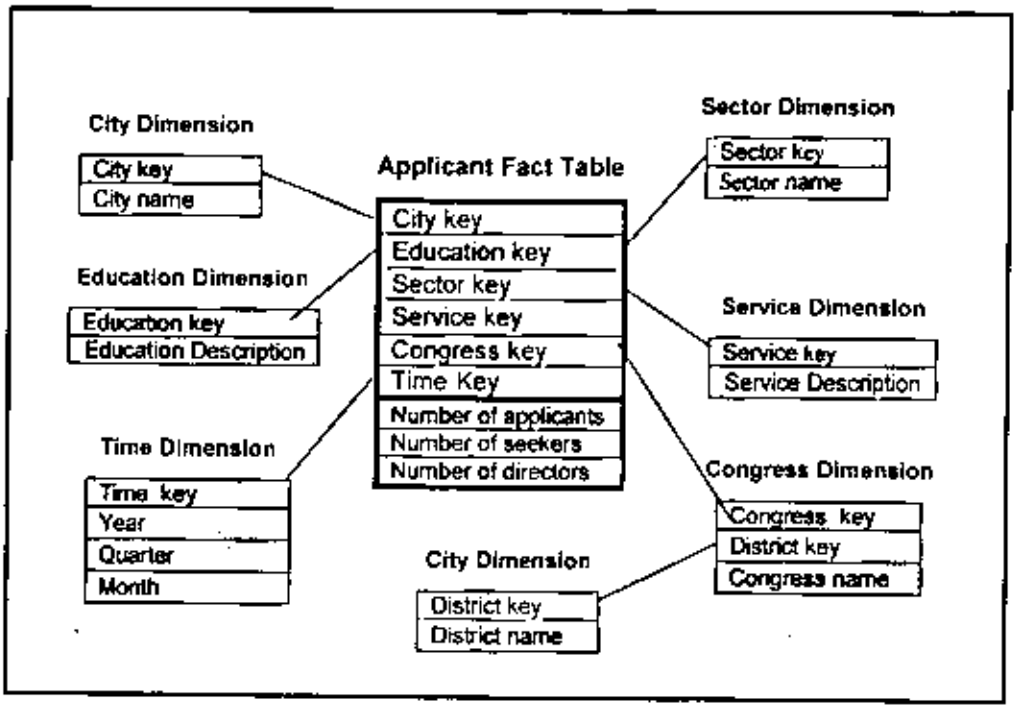


Figure-3.2: Snowflake schema of data warehouse

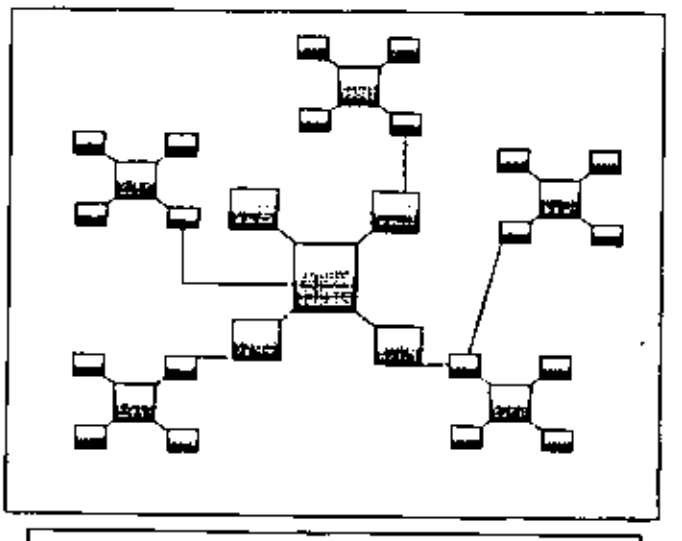


Figure 3.3: Fact constellation schema of DW

In our case study, we have chosen the star schema model because it satisfies and fits well our application.

3.5 Data Warehouse components

Before the process of designing a data warehouse, we should become familiar with the basic components of such system. These components include fact table(s), dimension, measures, and dimensions. Here, we elaborate on them.

1. **Dimension tables**: Dimension tables are relational database tables with a key that is used to joine it to a fact table. Dimension tables provide analytical perspective for all business processes. We have to ensure that each dimension table has the following charactersistics:
    - Has primary key.
    - Has one-to-many relationship with the fact table.
    - Consist of columns with highly descriptive information, which describe data that is stored in the fact table.
    - Flexible structure to allow some change.

In our General People's Committee of Manpower (GPCM), data warehouse includes dimensions such as city, congress, sector, education and so on. These dimensions allow the agency to keep track of information such as; quarterly or yearly direction of applicants, in which city and to which sector and so on.

2. **Fact table**: Fact table is a large central table of the star schema that contains most of the data with no redundancy; it contains keys to each of the related dimension tables and measures. Measures should be additive and numeric, because these values are the basic form on which

calculations can be performed. The dimensional keys of the dimension table are combined to form the multipart keys of the fact table. For instance, the city, congress, service, and education-level keys together form the multipart key. We should consider the following business points during the creation of a fact table:

- It's important to describe the characteristics and decide whether to include precalculated data.
- Keep in mind the size of the fact table and its impact on query performance.
- We have to choose the data grain (detail) to be stored in the fact table. The grain that you chose determines the facts that you can use in the database.
- Descrptive data, that in dimension tables.

For instance, a branch manager of the General People's Committee of Manpower may ask a question such as; *Which sector receives the highest number of citizens applications for employment?* To answer this query highly summarized data is required. On the other hand, the top management of the General People's Committee of Manpower may ask, *what was the total number of applicants that have been directed to a specific sector, in a specific city, in a specific year?* To answer this question we need more detailed data than the one needed by the branch manager.

3. **Measures:** are a numerical data in fact table typically represent the values resulted from some numerical functions such as count, sum, ect.
4. **Dimension:** is a business entity, such as city, sector, and applicant.

3.6 Types of data in data warehouse tables

There are various types of data in data warehouse tables such as:

- A numerical measure data (is also called fact data) that is calculated at each point in the cube space and it is quantitative nature.
- Derived data is data that has been derived or created from fact data.
- Metadata is data about the data

3.7 Updating data in data warehouse

Data in data warehouse is refreshed from time to time to keep the data up-to-date by adding new data snapshots. Business users determine the refresh rate cycle but dimension data is not refreshed in the same way as fact data. However, sometimes changes occur in the dimension data that must be updated. For instance, in our General People's Committee of Manpower data warehouse a dimension contains information about the different sectors may require some changes due to the changes that may occur in the real life. This process requires careful scheduling and execution.

3.8 Hierarchies

Within each dimension, there is an implicit grouping of values that can be organized into multiple levels. Such grouping (natural or imposed) is called Concept Hierarchy (CH). A concept hierarchy is some kind of mapping from low-level concepts (basic data values) to higher-level concepts (more general values). For example the grouping of time dimension from days to weeks, months, quarters, or years is depicted in figure-3.4.

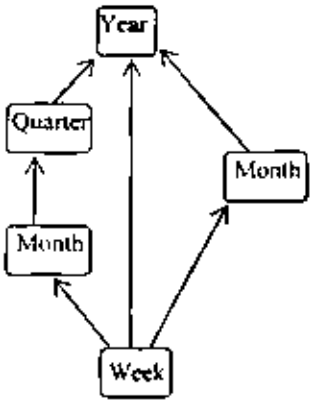


Figure 3.4: A concept hierarchy for time dimension.

3.9 Roll up and drill down operations

The Roll up and drill down are typical OLAP operations for multidimensional data. They refer to the analysis of data to greater or lesser detail. The drill down operation is to start from more general data to a more specific one, which resembles going from higher level in the concept hierarchy to a lower one. For example, using time hierarchy, we may start our analysis at the quarter level and drill down to month level, as depicted in figure-3.5. The Roll up operation is the reverse of the drill down operation, which means that we move up in the hierarchy to get the data in terms that are more general. For example, by using the time hierarchy, we may start our analysis at the quarter level and roll up to year level, as depicted in figure-3.5. The roll up operation is used to reduce the dimensionality of the data.

3.10 Slice and dice operations

The slice and dice operations are also typical OLAP operations for multidimensional data. As defined in [5], the slice operation performs selection on one dimension of the data cube. This operation can be imaged as taken horizontal or vertical slice of the data cub resulting in a subcube.

The dice operation defines a subcube by performing a selection on two or more dimensions. Figure-3.6, illustrates the slice and dice operations from our case study.

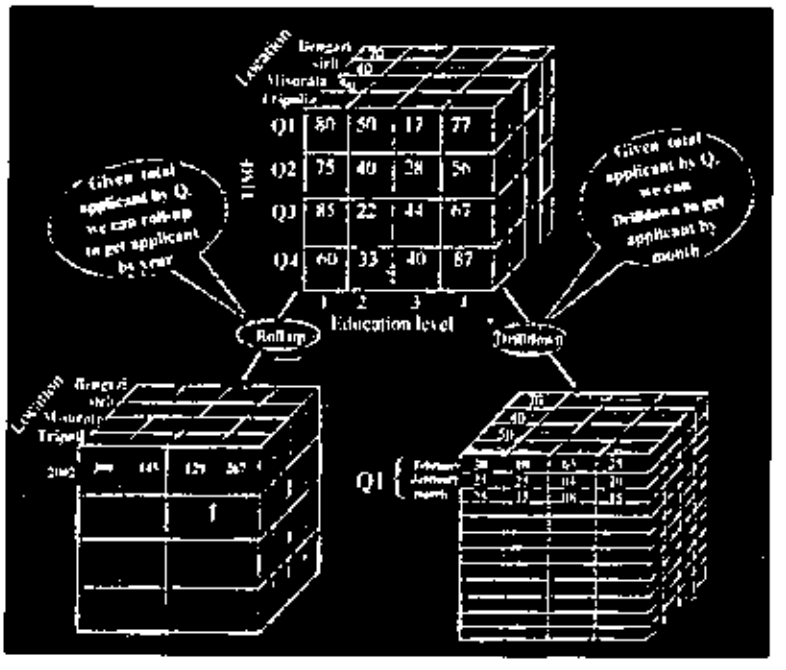


Figure-3.5: An illustration of roll up and drill down operations.

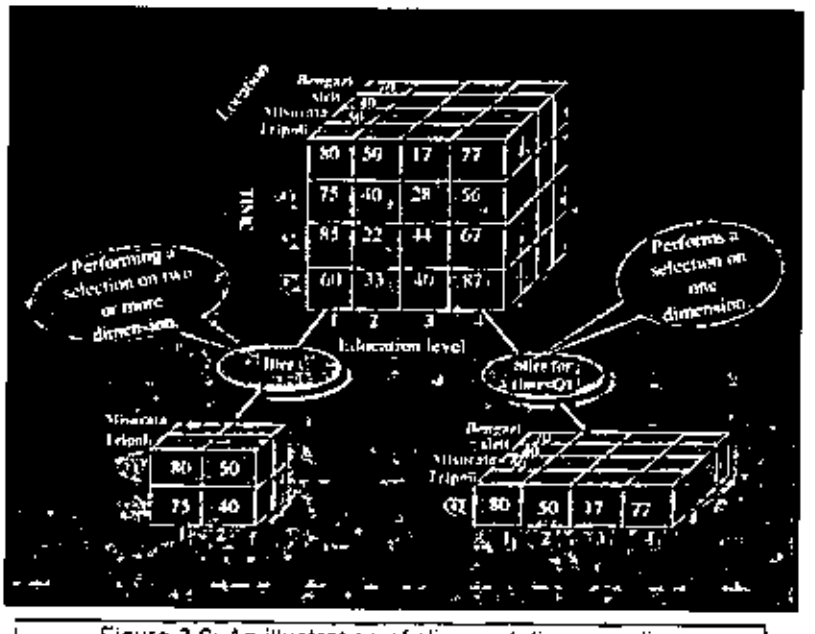


Figure-3.6: An illustration of slice and dice operations.

SYSTEM DESIGN AND IMPLEMENTATION

Chapter 4

System design and Implementation

Nowadays the unemployment problem in many developed countries has become one of the most important issues. In our country, the General People's Committee of Manpower (GPCM) is trying to avoid this problem by introducing a new project, which offers many jobs for the youth in accordance to their different educational levels. The first step in this project is to announce the new jobs in national and local newspapers, the national Jamahiriya broadcast radio station and in the local broadcasts radio stations. The job seekers apply for the suitable job according to their specialization (major), education level and the jobs offered in their city of residence. The applicants have to apply for job in any of the (GPCM) centers in each city. The General People's Committee of Manpower is doing its best in manually form to collect the reports produced by the computerized systems in each city. However each city in Jamahiriya has its own computer system to help in un employing people, but there are two main problems with these systems which are:

1. The reporting module takes a lot of time to extract some statistical reports about the applicants.
2. The system in each city works in an individual manner with its own database that differs from the other databases in other cities, which makes it very difficult to extract, and collect all the statistical reports needed for all cities together.

More details of the current system problems will be discussed later in this chapter. In addition to the above mentioned problems, the current OLTP system

takes too much time to extract the needed reports. Therefore, we need to collect all cities' data into one database.

For all the above-mentioned reasons, and to supply the decision makers with the information they need, building a data warehouse will be the best-sought solution.

4.1 Old system problems

According to the currently used information system in the General People's Committee of Manpower (GPCM) and based on our study of a sample of three centers in Tripoli, Misurata and sirite popularities, we realized that each center has its own separated database that differs in structure, coding, data types and field lengths, from the ones used in other centers. For example in Tripoli the files used are flat files data format, in Misurata a relational database is used, while in sirite non-relational database is used. Therefore we have realized that the current system of the General People's Committee of Manpower has the following drawbacks:

1. The administration of data is complex.
2. There is data inconsistency.
3. Solving inconsistencies is expensive and too slow.
4. The needed informational data for the decision support takes long time to acquire.
5. The decision support needs an informational data not a raw data.
6. The traditional decision support environment has failed to provide complete, accurate, integrated, and timely information to the secretary of manpower

These deficiencies motivated us to investigate the idea of using data warehouse in the decision support sector of the General People's Committee of Manpower.

4.2 OLTP system structure

Each local Secretary of Manpower in each city in Libya has its own database, each of which contains various database tables. As an example, table-4.1 depicts Misurata popularity database tables.

Table Name	Table Description
Applicant	This table contains personal data of the applicant such as; ID-number, name address, sex.
Specialty	This table contains the information related to the specialty (major) of the applicant.
Job-Group	Contains the attributes that describe the job-group specified in each city
Sector	Contains the attributes that describe the sector to which, applicants were directed to.
Moahel	Contains the attributes that describe the education qualification of the applicant.
Education-level	Contains the attributes that describe the education level of applicant, like primary preparatory, secondary, ...
Mothamer	Contains the attributes that describe the congresses in each city.
Service	Contains the attributes that are related to the military service record for males or national service for females.

Table 4.1: Misurata popularity database

4.3 The process of data warehouse design

In order to build the required data warehouse, we started by collecting information from scattered databases in the above mentioned three cites targeted in our study and store the data under a unified schema. As we will illustrate bellow, we constructed data warehouse via a process of data cleaning, transformation, integration, reduction, loading, and periodic data refreshing.

The design of the relation database requires the use of Entity Relationship (ER) model, which is appropriate for On-Line Transactional Processing (OLTP). Data warehouse design requires subject schema that is appropriate for On Line

Analytical Processing (OLAP). "The most popular data model for data warehouse is a multidimensional model. Such model can exist in the form of star schema, a snowflake schema, or a fact constellation schema." [5]. According to [11], most data warehouses use a star schema to represent the multidimensional data model, so we adopt this type of schema in the design of the General People's Committee of Manpower data warehouse. The details of the designed star schema will be discussed later in this chapter.

4.3.1 Data preprocessing

Typically, the Manpower databases, as an example of real-world databases that are highly susceptible to noise, missing, and inconsistent data due to their normally huge size. So it is better to improve the quality of the data by preprocessing and in turn this will improve the mining task.

During the implementation of the GPCM data warehouse, we were faced with two main problems. The first problem is the redundancy of the applicant's records, where applicants have applied in more than one city. We have to overcome this problem by the use of cleansing procedure in our data warehouse tools that deletes any redundant records. The second problem is that of entry data discrepancy. Because there are about 25 centers for data entry and each operator has its own style for data entry, for example:

1. The name "محمد" has been written as "محمد" by one operator and written as: "محمـــــد" by another.
2. The name "امنة" has been written as: "امنة" by one operator and written as: "أمنة" or "امنه" by another.
3. The name "عبد الحميد" has been written as: "عبد الحميد" by one operator and written as: "عبدالحميد" by another.
4. There are many ways of writing names in Arabic for such letters as: "ة", "ه", "ي" and "ى".

We have solved the above mentioned problems by some preprocessing procedures in our data warehouse tools to perform cleansing, integration, transformation and data reduction.

4.3.1.1 Data cleaning

There are many possible reasons for noisy data, at most human error occurring at data entry, the data cleaning routines work to clean the data by filling in missing values using a global constant, a most probable value, or using an attribute mean for numeric values. We used a global constant approach, which is replacing all the missing value by the same constant such as a label like "لايوجد". Even though this approach is simple and easy to implement, it is not recommended because it might effect the mining process because it could be mistakenly taken to be as an interesting concept. For example, we have replaced "****", "د" and "---" with "لايوجد". Also some of the attribute values have different values with the same meaning so it is better to unify the values with the same meaning to have the same values. For example, for the attribute military status "SERVICE" we replaced the values ["ضابط مستقيل", "ضابط سابق", "عسكري مستقيل"] with ["احتياطي"], and for the attribute education level "EDU_LEVEL we replaced the values ["امي", "لا يقرا", "لا يجيد القراءة "] , with ["امي"], as shown in figure-4.1.

ARMY	
NAM	NEWNAM
عسكري مستقيل	احتياطي
تعريف ضابط صف	احتياطي
د	لا يوجد
مستقيله	غير معلوم
خدمه	خدمة وطنية
تسويه لوضع	خدمة وطنية
ضابط سابق	احتياطي
بصفة التحاق بالقدري	خدمة وطنية
****	لا يوجد
اعفاء	غير معلوم
تسويه	خدمة وطنية

MOSTWA	
NAM	NEWNAM
لا يقرا	امي
لا يوجد القراءة	امي
دورة	شهادة تدريبية
ثانوية عامة	شهادة ثانوية
خبرة	شهادة خبرة
معهد متوسط	دبلوم متوسط
دبلوم معلمين	دبلوم متوسط
متوسط	دبلوم متوسط
---	لا يوجد
جامعي	شهادة جامعية
امي	امي

Figure 4.1: Some of the basic methods for data cleaning

Data inconsistencies may be the result of data integration of many different sources, different data store types and naming conventions. These inconsistencies may be corrected manually using external references or by the use of some routines designed to help correct such inconsistency. Figures 4.2 to 4.5 depict examples of the inconsistent data recorded for some tables in our case study of different data stores.

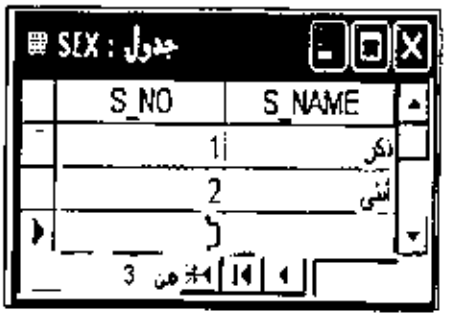


Figure 4.3: Sex table (MDB file)

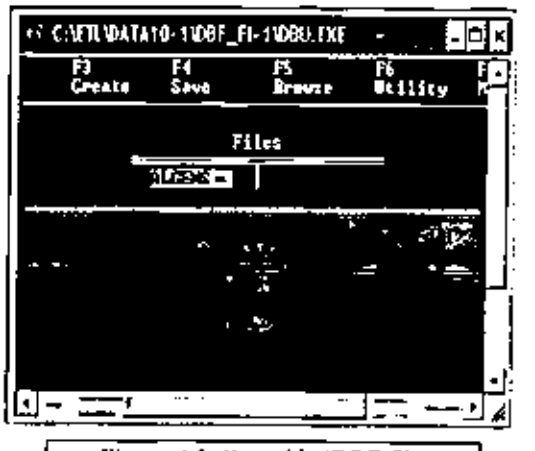


Figure 4.2: Sex table (DBF file)

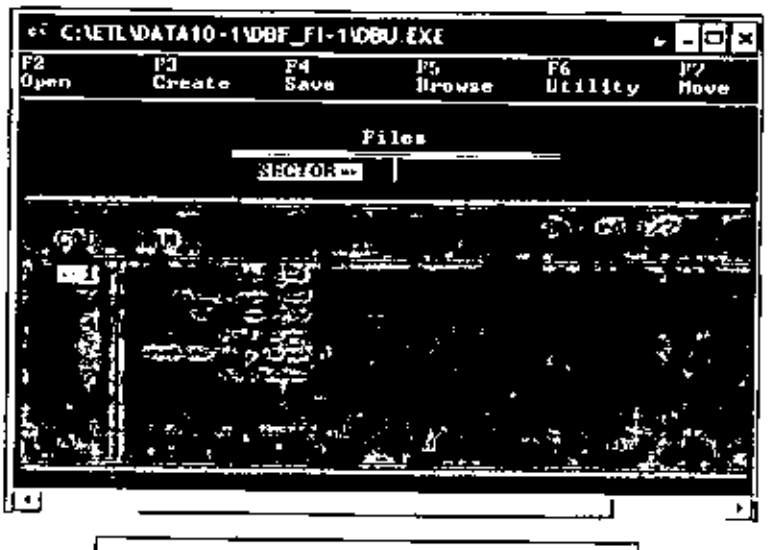


Figure 4.4: Sectors table (DBF file)

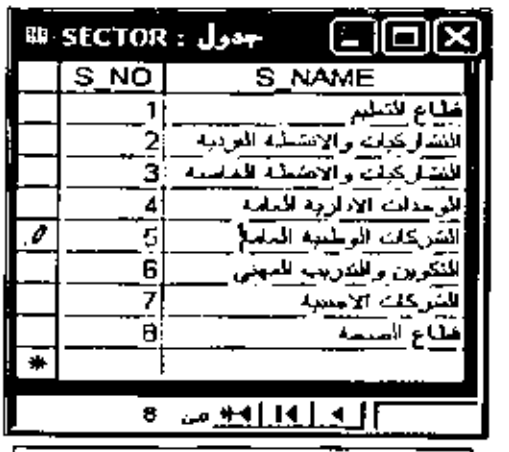


Figure 4.5: Sectors table (MDB file)

These data type store inconsistencies have been corrected by the development of knowledge base that can be used to help avoid errors, as depicted in Figure- 4.6.

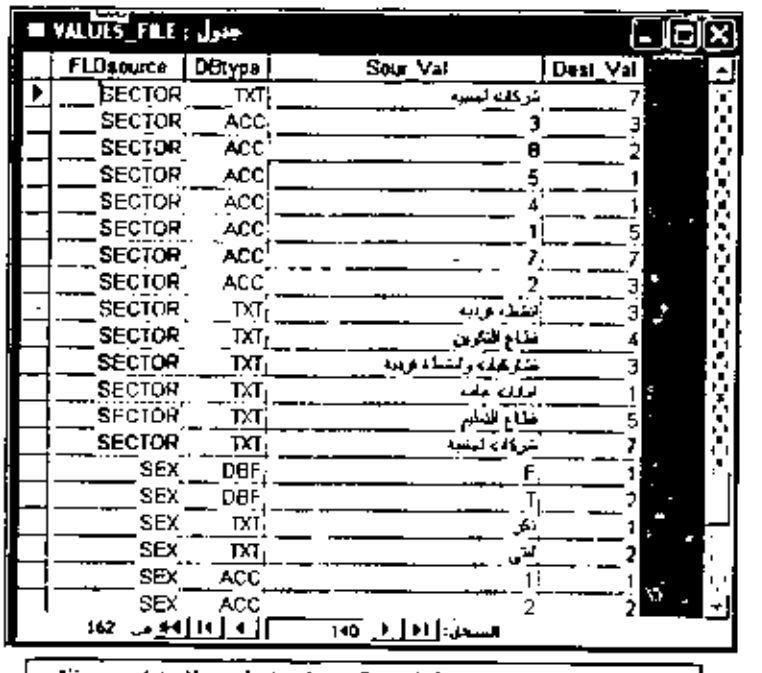
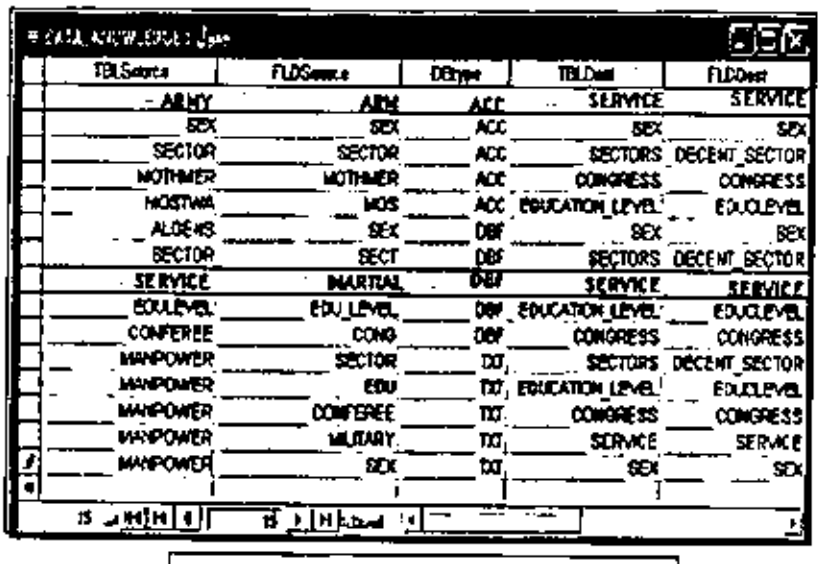


Figure-4.6: Knowledge base for solving some inconsistency

4.3.1.2 Data integration

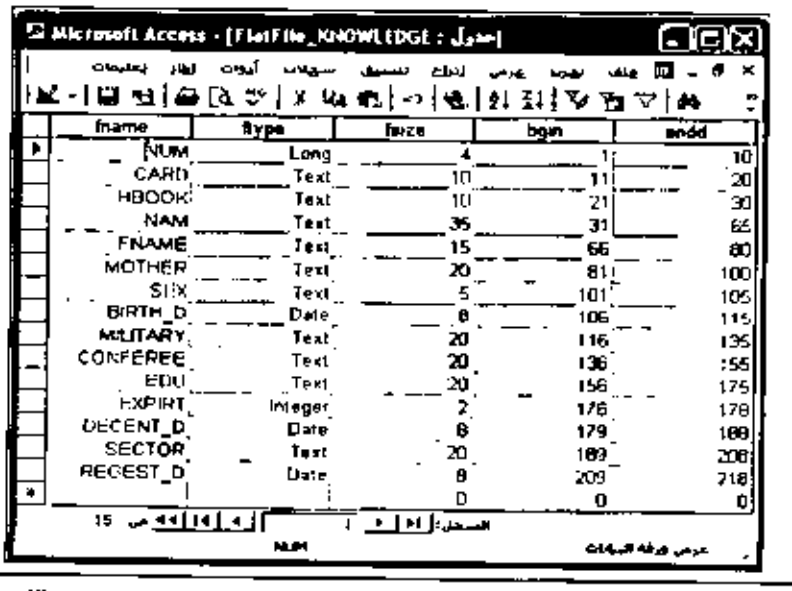
It is normal that the data analysis task involves data integration, which combines data from multiple sources (Text, DBF, MDB files), into a coherent database. As it has been mentioned in [5], this problem is known as the entity identification problem to (match/matched) up different data from different sources.

For example, in our case study, we made sure that the attributes MILITARY in DBF file, and ARM in MDB table refer to the same attribute SERVICE in our data warehouse for the same entity (applicant), by using Knowledge base that helps us to avoid errors in schema integration. Figure-4.7 below illustrates that. Figure-4.8, illustrates the knowledge base that used to convert flat file to temporary MDB file.



TBLSource	FLDSource	DBType	TBLDest	FLDDest
ARMY	ARM	ACC	SERVICE	SERVICE
SEX	SEX	ACC	SEX	SEX
SECTOR	SECTOR	ACC	SECTORS	DECENT_SECTOR
MOTHER	MOTHER	ACC	CONGRESS	CONGRESS
MOSTWA	MOS	ACC	EDUCATION_LEVEL	EDUCLEVEL
ALIENS	SEX	DBF	SEX	SEX
SECTOR	SECT	DBF	SECTORS	DECENT_SECTOR
SERVICE	MARTIAL	DBF	SERVICE	SERVICE
EDULEVEL	EDU_LEVEL	DBF	EDUCATION_LEVEL	EDUCLEVEL
CONFEREE	CONG	DBF	CONGRESS	CONGRESS
MANPOWER	SECTOR	TXT	SECTORS	DECENT_SECTOR
MANPOWER	EDU	TXT	EDUCATION_LEVEL	EDUCLEVEL
MANPOWER	CONFEREE	TXT	CONGRESS	CONGRESS
MANPOWER	MILITARY	TXT	SERVICE	SERVICE
MANPOWER	SEX	TXT	SEX	SEX

Figure 4.7: GPCM Knowledge base.



Microsoft Access - [FlatFile_KNOWLEDGE : Table]

fname	ftype	fsize	bgin	endd
NUM	Long	4	1	10
CARD	Text	10	11	20
HBOOK	Text	10	21	30
NAM	Text	35	31	65
FNAME	Text	15	66	80
MOTHER	Text	20	81	100
SEX	Text	5	101	105
BRTH_D	Date	8	106	113
MILITARY	Text	20	116	135
CONFEREE	Text	20	136	155
EDU	Text	20	156	175
EXPIRT	Integer	2	176	178
DECENT_D	Date	8	179	186
SECTOR	Text	20	189	208
REGEST_D	Date	8	209	216
		0	0	0

Figure 4.8: Knowledge base to convert flat file to temporary MDB file

4.3.1.3 Data transformation

Data transformations involve multiple techniques like, smoothing, aggregation, generalization, normalization and attribute construction. In our implementation, generalization is used to replace low-level data by higher-level concepts by the use of concept hierarchies. For example categorical address in which districts are generalized to high-level concept "congress", as depicted in figure-4.9.

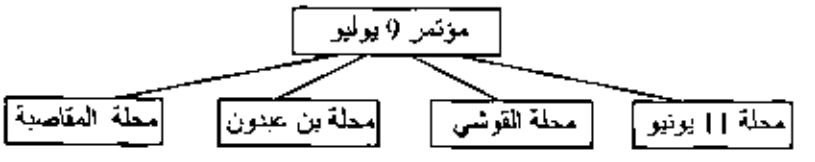


Figure 4.9: A concept hierarchy for attribute address in Misurata city

4.3.1.4 Data reduction

The manpower's data is huge and complex, so it takes along time to be processed, that makes it impractical for data mining. So, it is important to reduce the data size and remove the irrelevant attributes for the mining process, in a way that preserves the integrity of the original data. To reduce the data, number of data reduction techniques could be applied such as; dimension reduction, where the irrelevant or redundant attributes are removed. Another data reduction technique which is known as data cube aggregation might be also used where aggregation operations are applied to the data and the result is stored in a multidimensional data cube. For example, the manpower's database contains data for directed applicants for different jobs per quarter for the years from 2000 to 2006. However, sometimes it's of more interest if the total directed applicant by year rather than by quarter, so the aggregated totals are resulted in a smaller volume without losing any information. This type aggregation is illustrated in figure-4.10.

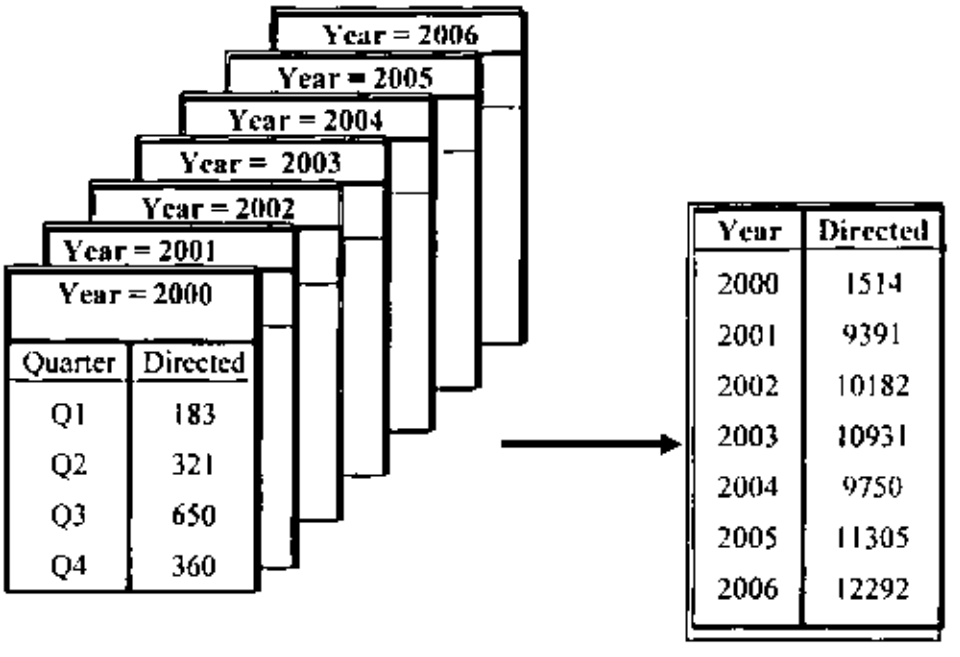


Figure 4.10: Data cube aggregation reduction

Concept hierarchy technique is also used in data reduction. For example in our case study the values of the attributes city, congress, district form a concept hierarchy with multiple levels as depicted in figure-4.11.

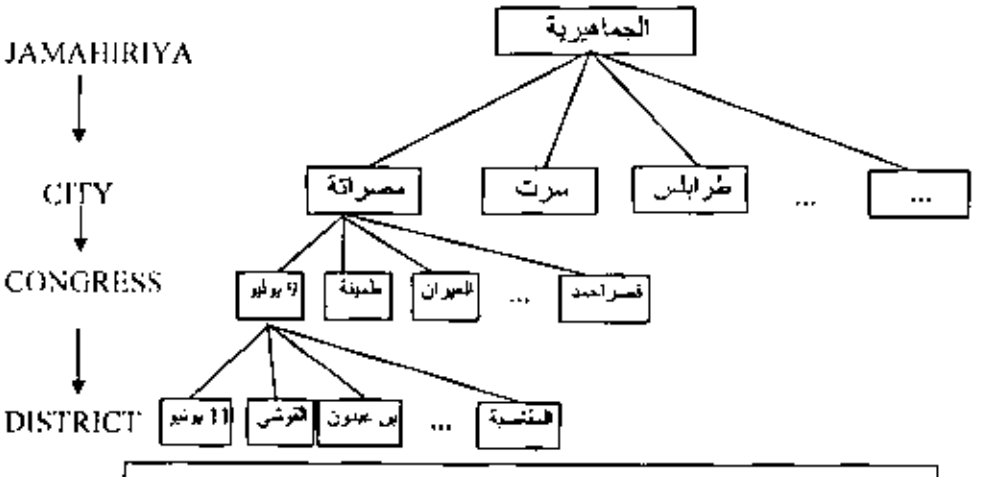


Figure 4.11: A concept hierarchy for attribute address in Jamahiriya

4.3.2 Building GPCM data warehouse

In the process of building the General People's Committee of Manpower data warehouse, it is sufficient to use star schema model, which containts a single fact

table and a number of dimensional tables. Table-4.2 and table-4.3, depicts the structure of the fact table and the dimension tables respectively.

Attribute name	Description	Remark
City_id	Each city has an unique number	key
Sector_id	Each sector has an unique number	key
EduLevel_id	Each education level has an unique number	key
Cong_id	Each congress has an unique number	key
Service_id	Each military service has an unique number	key
Time_id	Each time unit has an unique number	key
Total number of applicant	Number of all applicants	Measure
Number of seekers	Number of all seekers	Measure
Number of directed	Number of all directed	Measure
Education level	Number of applicants for specific education level	Measure
Military service	Number of applicants for a specific military service	Measure

Table-4.2: GPCM fact table structure

The fact table contains a number of measures that constitute the output. These measures are:

- Total number of applicants represents the number of applicants that can be reported accurately in a very short time.
- Number of seekers represents the number of applicants that seek to be employed for each city and/or for a specific period of time.
- Number of directed applicants represents the number of applicants that have been directed to a specific job. The number of directed applicants can be tabled in many ways such as; the number of directed applicants to a specific sector or in a given city or in a given period of time.
- Education level count represents the number of applicants in each education level.
- Military service count represents the number of applicants in each military service category in a given city or for all cities in a given period of time.

Dimension Name	Dimension Description
City	Contains the attributes that describe the city
Descen_Sector	Contains the attributes that describe the sector to which the applicant is directed.
Education_level	Contains the attributes that describe the education level
Congress	Contains the attributes that describe the congress in each city
Service	Contains the attributes that describe the military service for male or social service for female
Time	Contains the attributes that describe the time dimension, each year consists of 4 quarters.

Table-4.3: GPCM dimension tables structure.

The star schema diagram containing the fact table and the dimensional tables is depicted in figure-4.12.

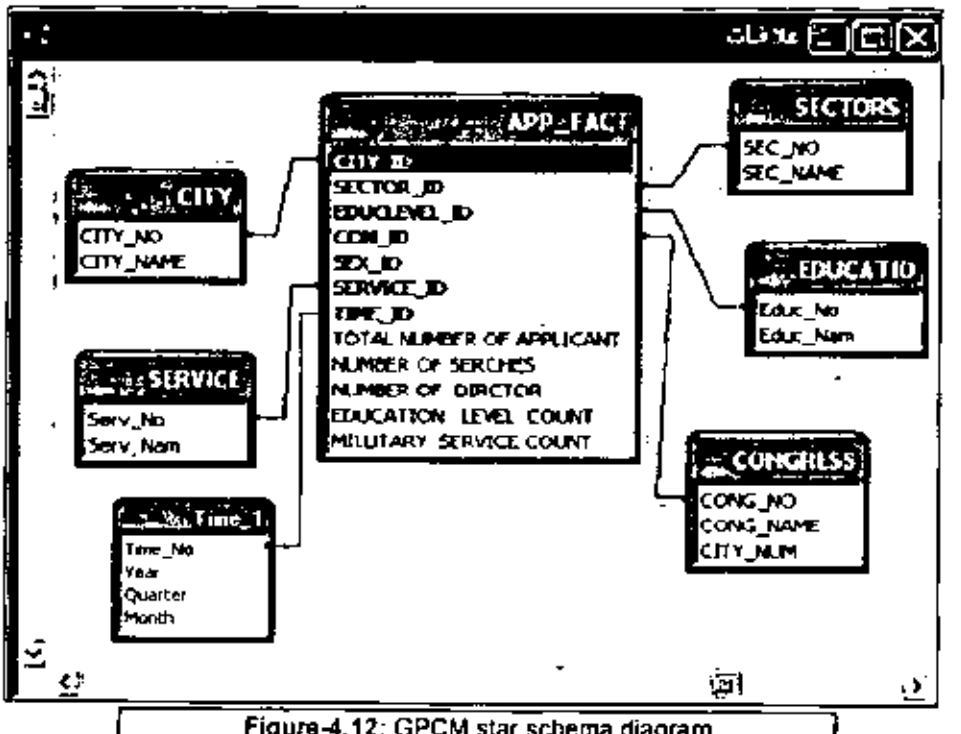


Figure-4.12: GPCM star schema diagram

4.3.3 GPCM multidimensional data model

The GPCM multidimensional data model can be visualized as data cube with several dimensions, as depicted in figure- 4-13. The city dimension consists of the cities in Libya in our case study. The time dimension divides the year into four

quarters (Q1, Q2, Q3 and Q4). The education level dimension is divided into six educational levels.

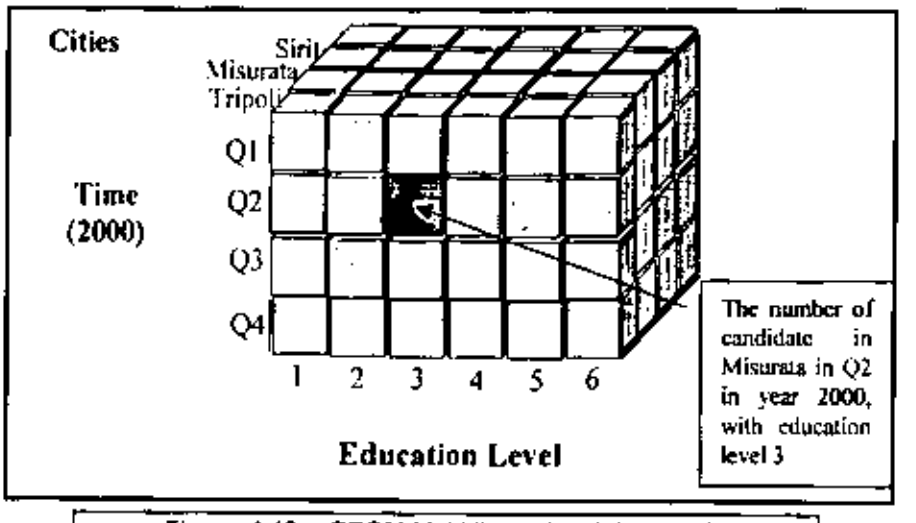


Figure-4.13 : GPCM Multidimensional data model

4.4 Advantages of data warehouse

There are a number of advantages that are beneficial to the use of data warehouse in any corporation and these are:

- **Scalable**

  One of the advantage for which we use data warehouse is to store and manage a huge amount of data, which was stored in several and different databases. The data warehouse can be scalable to contain the huge amount of additional data that may accumulate rapidly through the next few years.

- **Available**

  The availability of up to date data in data warehouse helps the decision makers to get the information they need whenever they want instead of collecting it manually or semi-manual from different cities or locations.

- **Flexible**

  The flexibility in dealing with the data stored in data warehouse system by the use of operations such as; drill down and roll up. By drilling down on a given

dimension, we can see more detailed data and by rolling up, we can view summarized or aggregate data.

- **Integrated**
  A final singular and globally acceptable set of information can be produced from a data warehouse; this information is derived from different heterogeneous sources in different cities or locations.
- **Reliable**
  In order to provide the decision makers with accurate and consistent data, the data in different cities are modulated. The applicants' data in each city is corrected before transporting it to the data warehouse.

4.5 Data warehouse reports

It is more valuable for the GPCM to extract general report for all the cities in Jamahiriya to be used in the decision-making process. In the present system, to collect all the reports from the cities it may take days or even weeks, in addition to the problems of data accuracy because the process is not fully computerized. The use of data warehouse gives the solution for these problems because data is collected completely by computer systems after being cleansed. Therefore, as we have explained previously it is an important step for data warehouse to be refreshed with data from time to time to be ready for extracting knowledge that is helpful to make the right decision whenever it's needed for all cities.

There are many reports that can be extracted from the data warehouse, which cannot be extracted via conventional OLTP systems because it takes a lot of time to create the program needed for that, in addition to the slow performance. Figures-4.14 and 4.15 show a sample of statistical reports for the total number of directed applicants in all cities (Tripoli, Misurata and Sirite) according to educational level, quarter in the different sectors in a given year.

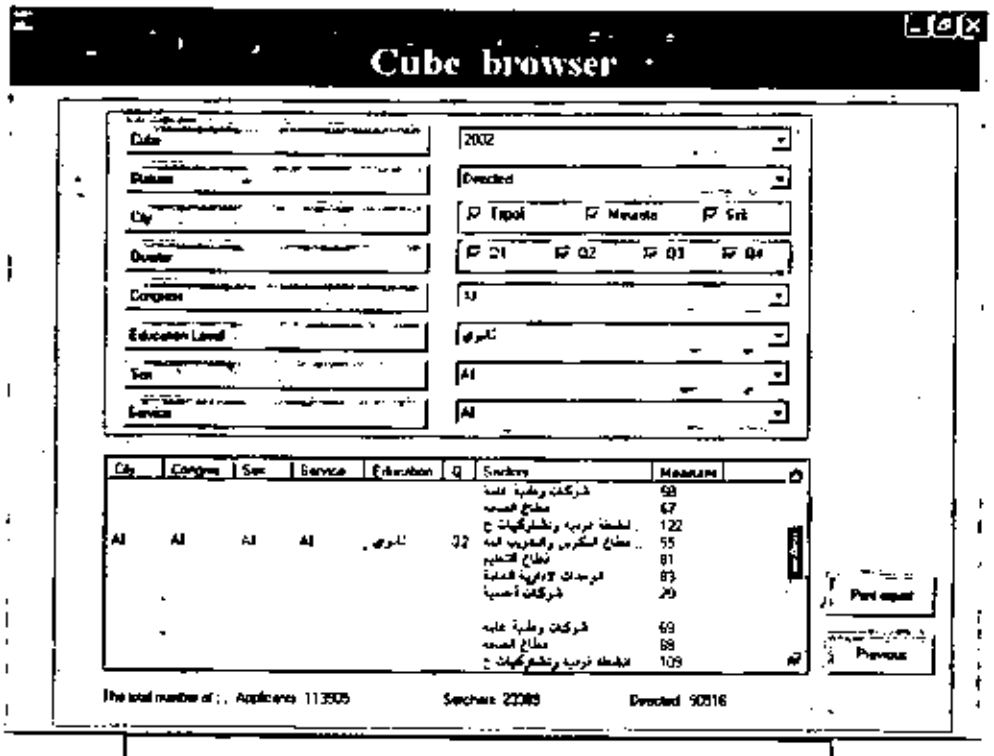


Figure-4.14: Sample of statistical report data for all cities.

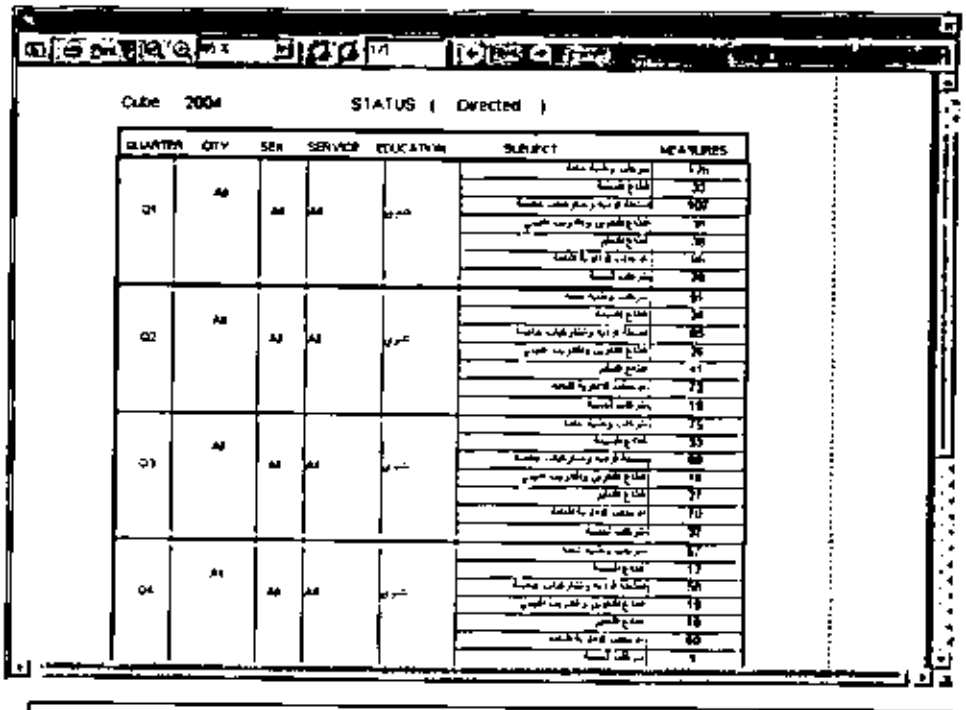


Figure-4.15 : Sample of statistical of directed applicants for all cities

Next, figure-4.16 depicts the total number of applicants in all cities distributed according to the marital status, gender, in the year of 2003.

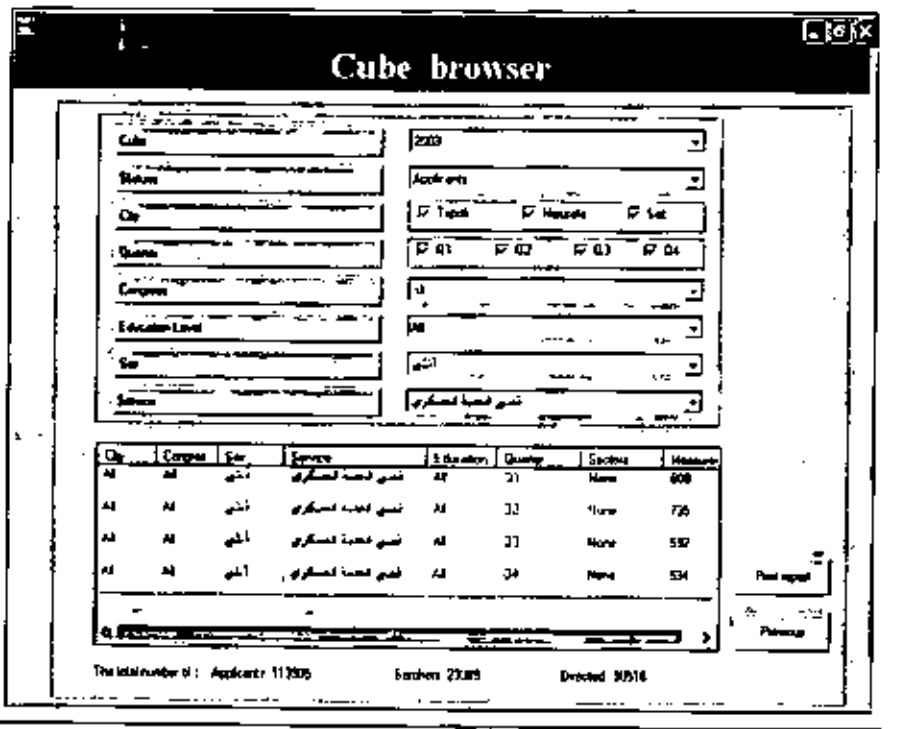


Figure-4.16 ; Total number of applicants according to the marital status, gender in year 2003

Figure-4.17 shows a snapshot of a statistical report from the OLTP system, which calculates the total number of applicants according to different educational levels in the year of 2006, in Misurata city (Misurata city is a sample example for all cities). Because the current information system gets the statistical reports from each city in Jamahiriya individually, that takes a lot of time, to collect them with less confidence.

العدد الكلي	الإناث			الذكور			المؤهل العلمي
	المجموع	منسبين	باحثين	المجموع	منسبين	باحثين	
561	221	125	96	340	239	101	أمي
657	256	155	101	401	289	112	ابتدائي
563	185	120	065	478	256	222	إعدادي
940	278	101	177	562	290	272	ثانوي
1304	582	293	289	722	402	320	جامعي
263	162	115	47	201	103	98	عالي

Figure 4.17: Statistical data for Misurata city in the year 2006 – OLTP system

Figure-4.18 shows a statistical report produced by our data warehouse system. In this report the total number of applicants in all cities is calculated according to different educational levels in the year of 2006.

العدد الكلي	الإناث			الذكور			المؤهل العلمي
	المجموع	منسبين	باحثين	المجموع	منسبين	باحثين	
2508	1047	456	591	1461	860	601	أمي
2569	961	554	407	1608	896	712	ابتدائي
3476	1556	655	601	1920	998	922	إعدادي
4978	2095	1276	819	2883	1711	1172	ثانوي
7555	3198	2212	986	4357	2437	1920	جامعي
974	207	101	106	767	469	298	عالي

Figure-4.18 Statistical data for all cities in the year 2006-Data Warehouse system

In figures-4.19 to 4.21, the resulted statistical information which represent the total number of applicants in all cities is tabled according to the military service and sex, from the year 2000 to the year 2006.

المجموع الكلي	لم يقضي الخدمة الوطنية			قضى الخدمة الوطنية			المدينة
	المجموع	انثى	ذكر	المجموع	انثى	ذكر	
58021	19404	5052	14352	38617	10187	28430	طرابلس
24705	8141	2888	5253	16564	7103	9461	مصراتة
18047	8012	2747	5265	10035	3100	6935	سرت

Figure-4.19: Statistical data in the year 2006 – Data Warehouse system

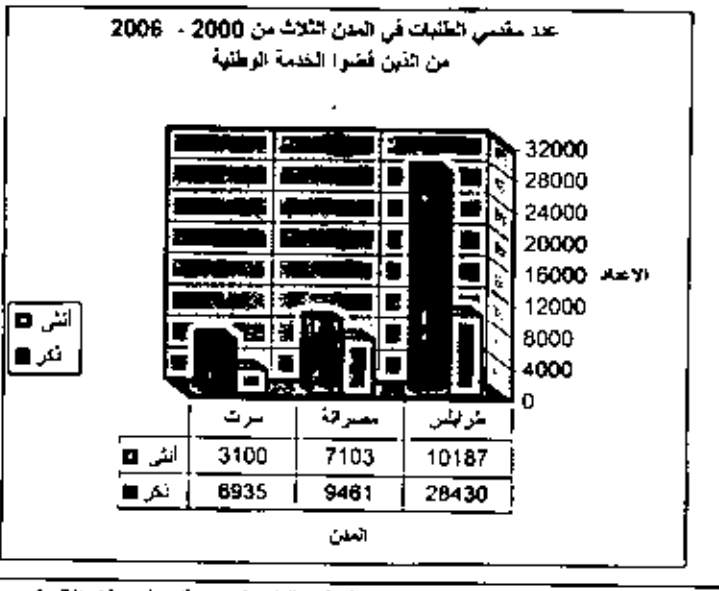


Figure-4.20: Graph of resulted Statistical data for applicants who did finish the military service

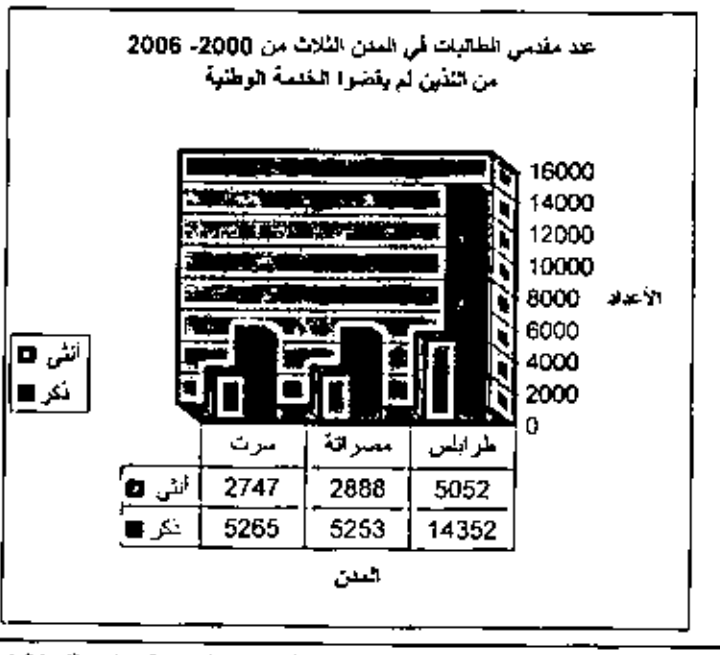


Figure-4.21: Graph of resulted Statistical data for applicants who not finished the military service

Figure-4.22 depicts the total number of directed applicants to different sectors in all cities from the year 2000 to 2006. Figure-4.24 represents the number of applicants and directed applicants in all cities from the year 2000 to 2006. We can notice that the line representing the number of applicants' lies above that for the directed applicants, which is a logical result for the applicant count to be greater than or equal to the directed applicant.

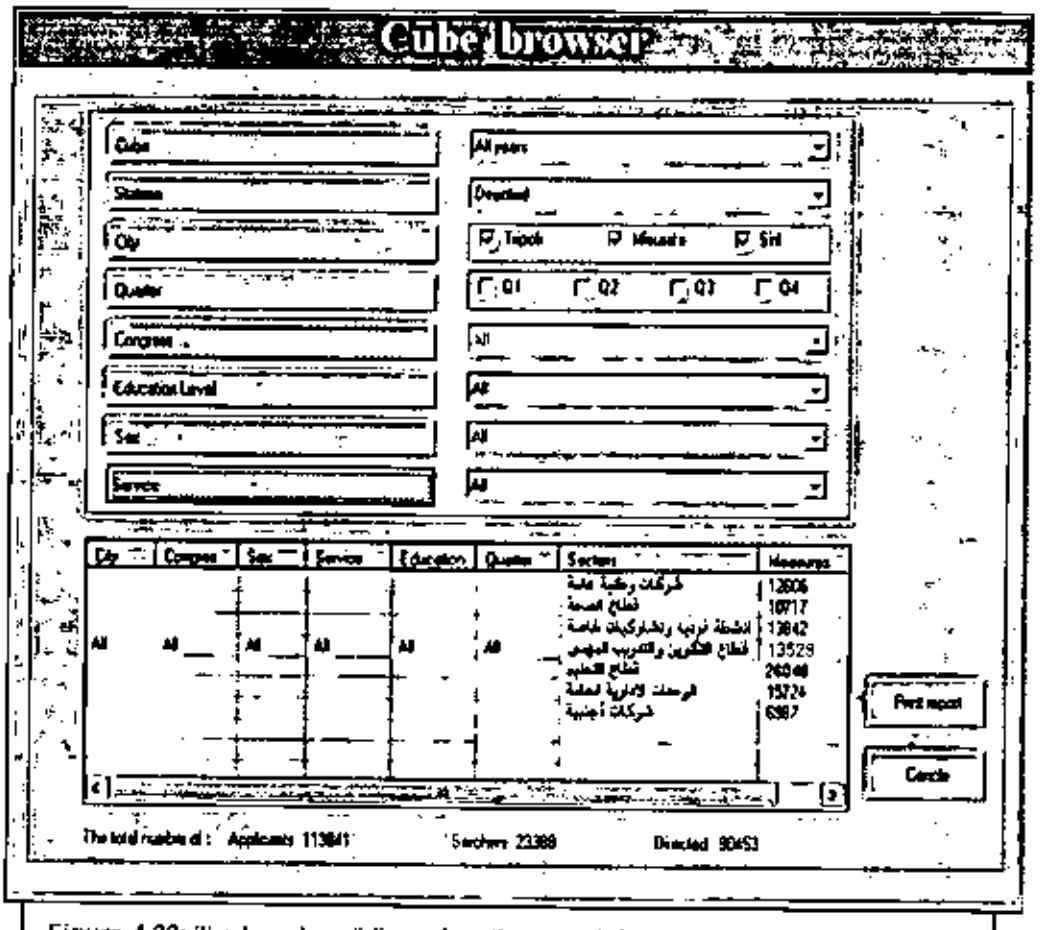


Figure-4.22: Total number of directed applicants to deferent sectors for all cities, from the year 2000 to 2006.

Next figure-4.23 depicts the total number of job seekers, according to their preferred sectors in all cities fro the year of 2000 to the year of 2006. Figure-4.25 represents a comparison between seekers for jobs and the directed applicants to different sectors from the year 2000 to 2006 in all cities.

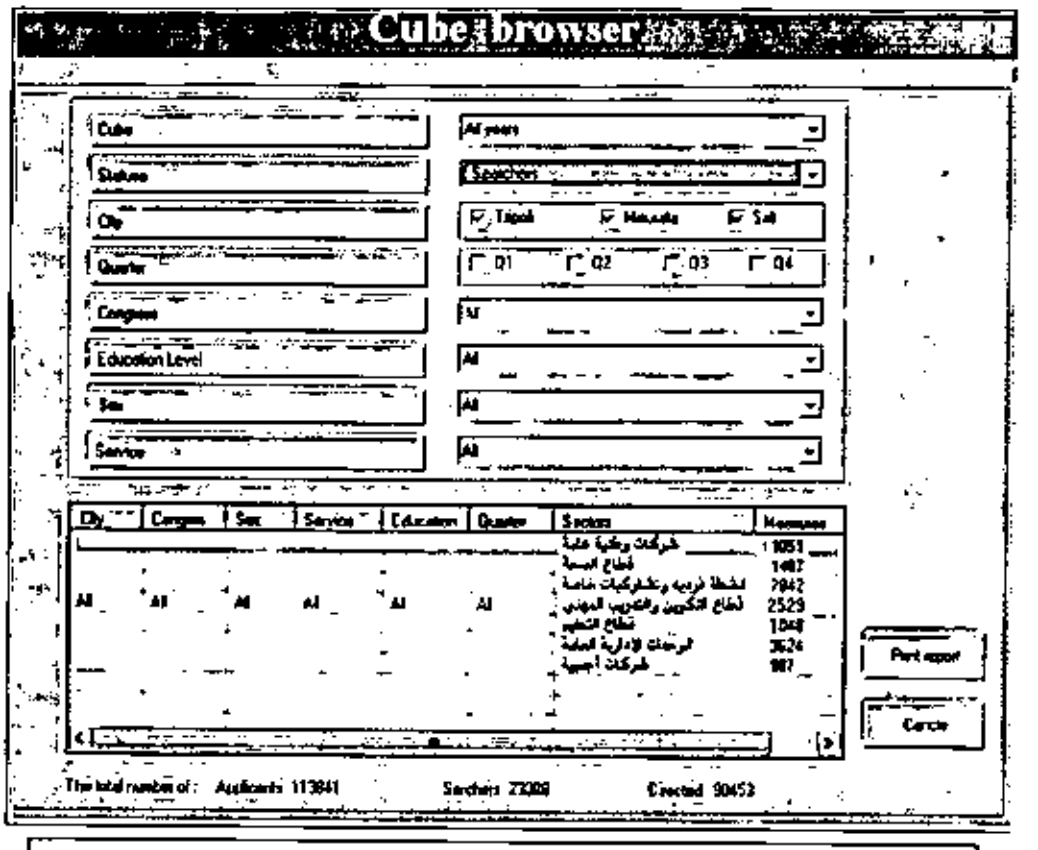


Figure-4.23: Total number of seekers, according to their preferred sectors for all cities, from the year 2000 to 2006.

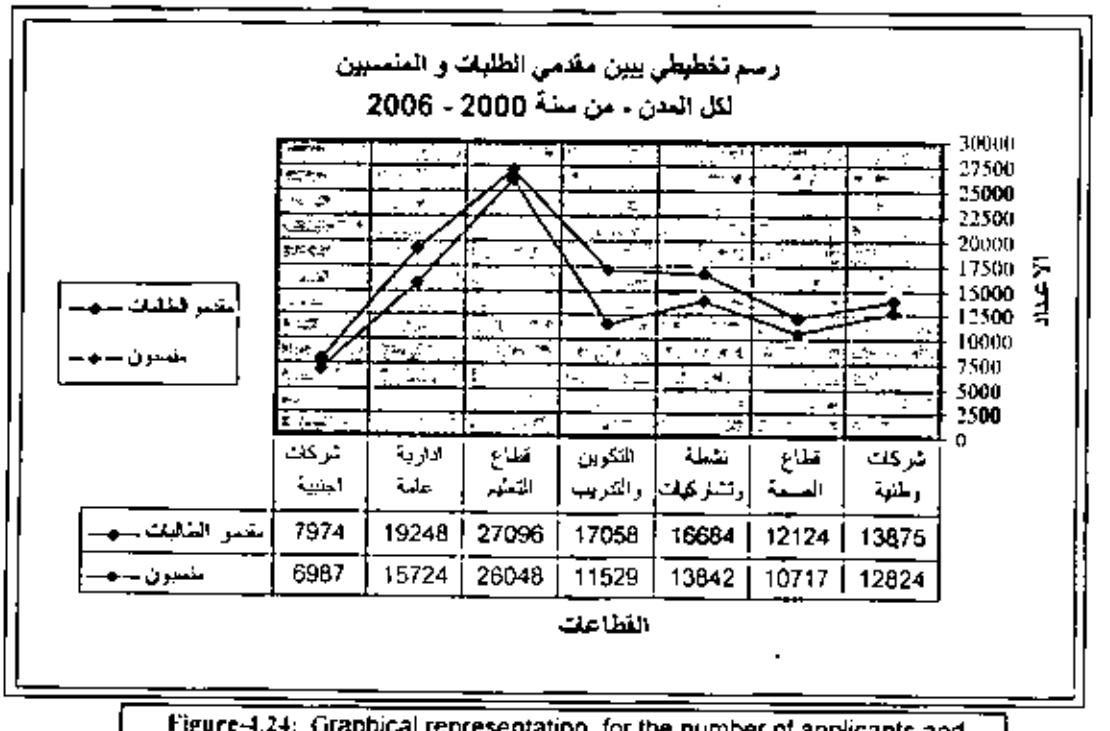
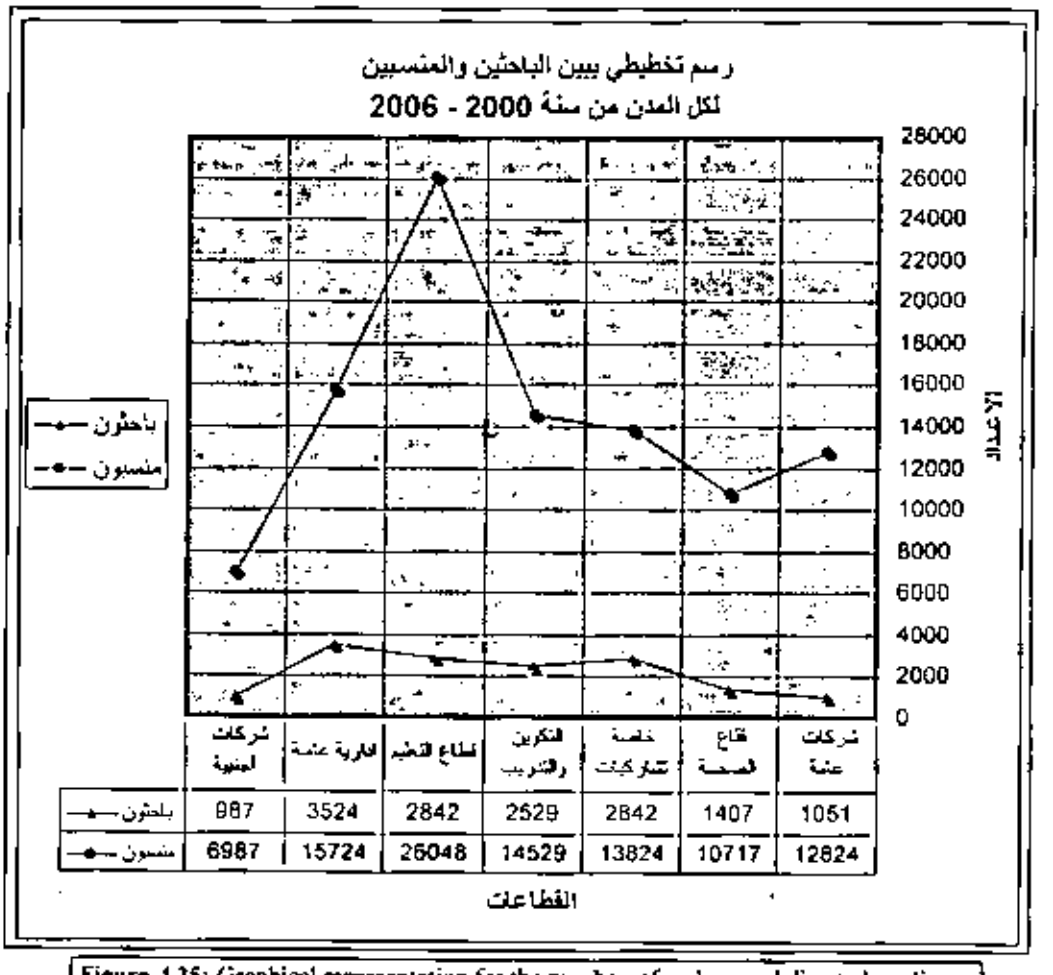


Figure-4.24: Graphical representation for the number of applicants and directed applicants in all cities from the year 2000 to 2006



	شركات أجنبية	قابلية عامة	قطاع التعليم	التكوين والتدريب	خاصة شركات	قطاع الصحة	شركات عامة
باحثون	987	3524	2842	2529	2842	1407	1051
منسبون	6987	15724	26048	14529	13824	10717	12824

Figure-4.25: Graphical representation for the number of seekers and directed applicants in all cities from the year 2000 to 2006

The previous graphical representations and reports are the results of our data warehouse system which will help the decision makers in taking the right decisions in the right time with no delay according to the new information available.

CONCLUSION AND FUTURE

Chapter 5

Conclusion and Future

5.1 Conclusion

"As our world is now in its information era, a huge amount of data is accumulated everyday. A real universal challenge is to find actionable knowledge from a large amount of data. Data mining is an emerging research direction to meet this challenge. Many kinds of knowledge (patterns) can be mined from various data", [15]. The objectives of this work are to study the impact of using data warehouse on a huge amount of data in Manpower Employment Decision Support System in which the data quality are concerned. In our research, we compared the data warehouse with the OLTP system to know the issues and impacts of using the data warehouse system on the information system. In fact, we found that we cannot completely replace the information system with the data warehouse but the data warehouse is just for helping the information system in the process of decision-making. We measured the reporting performance between the two systems and found that the data warehouse is more powerful than the ordinary old system. The old system has its features, which the data warehouse cannot do it (as an example add a record, update a record or delete a record), this is because it is not its functions. We cannot use the old system instead of the data warehouse especially in decision support reports or multidimensional reports, which need allot of time from the Information System staff, and at the end, these reports are static reports not dynamic ones.

We have accomplished the objectives of the study by focusing on the benefits gained from using data warehouse, and why it is more powerful than the use of traditional databases in decision-making. The case study tackled the impact of using the data warehouse on employing people in People's Committee of Manpower .We have analyzed and designed it by V.B studio 6.0, and SQL services programming languages, in order to compare the performance and time. We tested our system with real, synthetic and scattered data that was collected from different sources in Libya with different types (access database, DBF files and flat file) and sizes (15.56 MB, 6.85 MB and 28.1 MB).

The problems that face the old system users, which we mentioned in section 4.1 motivated us to use data warehouse system in the decision support organization. The advantages of this system were shown in section 4.4, where the resulted reports and figures emphasize those advantages and the ability of such system to help decision making which is not easy to achieve using the old system. The reports and figures produced by the presented system show how clear the overall situation of the employment process. So decision makers can take the right decisions in the right time with no delay according to the new information available.

5.2 Future research trends

This work mainly concerned with the data warehouse construction (i.e. preprocessing of data and the construction of data warehouse model), so extension of this work to be of more value and further help decision makers is to add knowledge extraction and discovery techniques such as classification, association rules, and sequential analysis, and using data mining for

applications such as: customer profiling, market-basket analysis, and fraud detection.

Also the internet environment might be considered to extend this system for online updating of data warehouse contents. For the development of such system its better to consider a programming language that directly support the data warehouse operations (like Oracle).

So we recommend the current information system in General People's committee of Manpower to include data warehouse system especially in the decision support reports.

References

1 Al-saiad Gergawi, Data Integration in database systems, faculty of engineering, Tanta University, Egypt, 2004, pp. 1.

2 Matthias Jarke, Yannis Vassiliou, Data Warehouse Quality. 2nd Conference on Information Quality. Massachusetts Institute of Technology, Cambridge, 1997, p. 4.

3 Amel Bakry Abd El Alem. MSC, Data Mining For Improving Data Capabilities, Department of Computer & information science, Institute of Statistical Studies and Research(ISSR), Cairo University, Egypt, 2001, p. 9.

4 Ramez Elmasri, Shamkant Navathe, Fundamentals Of Database Systems, The Benjamin/Cummings, Inc, California, 1989, pp. 3-4.

5 Jawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Simon Fraser University, Canada, Morgan Kaufmann, 2000, pp. 7-9.

6 Robert Vieirra, professional SQL Server 2000 Programming, A John Wiley & Sons, Inc, Indiana, 2000, pp. 891-910.

7 Micael J. A. Berry, Gordon S. Lin off, Data mining techniques, for marketing, sales, and customer relationship management, second edition, published by Wiley Inc, Indiana, U.S.A, 2004, pp. 505-508.

8 Karttih J., white paper, Data Mirror Benefits of Transformational Data Integration, Data mining corporation, Toronto, Canada, 2002, pp.10-11.

9 Lori oviatt, Margo Crandall, Designing and implementing of data warehouse using Microsoft SQL service™ 7.0 delivery guide, Microsoft Corporation, 1999, pp. 7-10.

10 Halim Habib Hanna, MSC, Data Modeling in database and conversion between models, Institute of Statistical Studies and Research (ISSR), Cairo University, Egypt, 1994, pp. 1-3.

11 Yin Jenny Tam, Data cube Its Implementation and Application in OLTP, Simon Fraser University, Canada, Department Computer Science, 1998, pp. 12-15.

12 Ahmed masheri El Diab, Studying Performance of Data Mining of real Database Using The Classification Technique, Faculty of Engineering, Cairo University Giza, Egypt, June 2003, pp, 37-36

13 M. Berry and G. Linoff, Introduction to Data Mining and Knowledge Discovery, Third Edition by Two Crows Corporation, U.S.A, 1999, PP. 1-3.

14 Daniel T. Larose, Discovering knowledge in data, Central Connecticut State University, A John Wily & Sons, Inc., Canada, 2004, pp. 1-5.

15 Jian Pei, Pattern-Growth Methods For Frequent Pattern Mining, Simon Fraser University, Canada, June 13, 2002, pp. 143.

16 Olivia Parr Rud, Data Mining Cookbook, a John Wiley & Sons, inc, New York, 2001, pp. 31-34.

17 Panos Vassiliadis, Mokrane Bouzeghoub, Christoph Quix, Towards Quality-Oriented Data Warehouse Usage and Evolution, DWQ: Foundations of Data Warehouse Quality, (CAISE 99), Heidelberg, Germany, 1997, pp 1-3.

18 Karl Aberer, Klemens Hemm, A methodology for building a data warehouse in a scientific environment, first IFCIS international conference on cooperation information systems, Brussels Belgium, 1996. pp, 2-7

الملخص العربي

إن توفير البيانات لدعم اتخاذ القرار هو الوظيفة الرئيسية لمستودع البيانات، و على الرغم من ظهور نظم المعلومات التقليدية التي تتعامل مع البيانات التي يحتاجها متخذي القرار، إلا أن مستودع البيانات يضيف لمثل هذه النظم مزايا متعددة من خلال تحسين و توسيع مجال و دقة و سرعة الوصول للبيانات.

إن مستودع البيانات هو عملية وليس نتيجة لأن بياناته جمعت من مصادر مختلفة و وحفظت في المستودع بعد إجراء العديد من عمليات المعالجة الأولية مثل التنظيف و التكامل و التوحيد و التحويل و الاختزال لغرض الحصول على نظرة واحدة لجزء من العمل أو العمل ككل.

من الطبيعي جدا أن مستودع البيانات قد يغير من طبيعة دعم القرار، فهو حلقة الوصل بين التطبيقات والبيانات (التي كانت متفرقة ومنعزلة ولكنها الآن متحدة في مستودع واحد).

الهدف من هذه الرسالة هو دراسة موضوع مستودع البيانات، وتأثيره على نظم المعلومات، وخاصة على عملية اتخاذ القرار. وكيف أن عملية دعم اتخاذ القرار تحتاج إلى تحسين كفاءة و جودة البيانات. ولقد قمنا باستعراض تطبيق عملي يتضمن بيانات حقيقية، ومن خلال هذا البحث برهنا على أهمية مستودع البيانات، ولماذا يتم اختياره في نظم دعم اتخاذ القرار، وذلك لأنه يتميز بالسرعة الشديدة وتوفير الوقت والجهد المطلوبين لإعداد التقارير المستخدمة في دعم القرار.

ومن خلال هذا البحث أيضا، اتضح تأثير استخدام مستودع البيانات على نظم المعلومات في دعم القرار و يكمن هذا في سرعة ودقة التقارير المستخلصة.

هذه الدراسة بنيت على أحد أهم القطاعات في الجماهيرية العظمى، وهو قطاع القوى العاملة والتدريب وذلك بتنسيب الوطنيين للعمل في مختلف القطاعات والشركات داخل البلاد، والذي يتعامل مع قواعد بيانات متعددة ذات طبيعة مختلفة وأحجام هائلة، ومصادر متعددة.

وفي نهاية بحثنا هذا برهنا أن مستودع البيانات هو تطور جديد وميزة كبرى في نظم دعم القرار.

تحتوي الرسالة على خمسة فصول و هي:

الفصل الأول : وهو مقدمة الرسالة، يتضمن مقدمة عن البيانات وقواعد البيانات ونظم المعلومات و نظم دعم القرار و نظرة عامة على تنقيب البيانات ومستودع البيانات، وكذلك أهم أهداف الرسالة وخلاصتها.

الفصل الثاني: يحتوي على الخلفية اللازمة للبحث، فهو يعرض تعريفات لمستودع البيانات، و بناءه و تطويره ووصف خصائصه، و مميزاته والبنية الأساسية له. ومقارنة بين نظامي المعاملات اليومية والمعاملات التحليلية، و وصف العوائق التي تعتري مستودع البيانات.

الفصل الثالث: يعرض مراحل تصميم مستودع البيانات و مقارنة جميع النماذج الخاصة ببناء مستودع البيانات واختيار الأنسب في موضوع الدراسة.

الفصل الرابع: يتم في هذا الفصل تصميم مستودع البيانات وبيان كيفية تعامله مع المشاكل الناجمة على استخدام قواعد بيانات تقليدية في مشروع تنسيب الوطنيين إلى الوظائف المختلفة، ولقد تم عرض مجموعة من التقارير والرسوم البيانية الموضحة لنتائج استخدام مستودع البيانات ومقارنتها مع تقارير الأنظمة العاملة.

الفصل الخامس: وهو يعرض الاستنتاجات ونظرة مستقبلية لإمكانية الاستفادة من دعم نظم المعلومات الحالية ونظم اتخاذ القرار بمستودعات البيانات .

كليــــة العلــــوم

قســــم الحاســــوب

عنــوان البحــث

تأثيـــر استخــدام مستــودع البيانــــات علـــى أنظمــة دعــم اتخــاذ القــــرار فـــي مجــال القــوى العاملــــة

مقـــــدمة مــــن الطالب

مفتــام علـــي المرابـــط أعنيبـــة

* * لجنــــة المناقشــــة :

1 – د. فــرج عبد القــادر المــؤدب ....

(مشــــــرفاً)

2 – زكــريا سليمـــــان الـــزوبي ....

(ممتحنـــاً داخليــاً)

3 – د. ادريــس سامـــي الفقيــه ....

(ممتحنـــاً خارجيـــاً)

يعتمـــــد

د. أحمد فرج المحجوب

أمين اللجنة الشعبية لكلية العلوم