Adaptation System for Data Retrieval and Information Translation from Semantically Heterogeneous, Autonomous Data Sources

a thesis submitted
for partial fulfillment of the requirement
for the degree of

Doctor of Philosophy (Ph. D)

by

Majid Zaman

Post Graduate Department of Computer Science,
Faculty of Applied Science and Technology,
University of Kashmir, Srinagar, J&K, India - 190006

under the supervision of

Dr. S. M. K. Quadri

in

Computer Science

August, 2012

Declaration

This is to certify that the thesis entitled "Adaptation System for Data Retrieval and Information Translation from Semantically Heterogeneous, Autonomous Data Sources", submitted by Majid Zaman, in the Post Graduate Department of Computer Science, University of Kashmir, Srinagar for the award of the Doctor of Philosophy (Ph. D.) in the area of Computer Science, is a record of an original research work carried out by him under my supervision and guidance. The thesis has fulfilled all the requirements as per the regulations of the University of Kashmir and in my opinion has reached the standards required for the submission. The results embodied in this thesis have not been submitted to any other University or Institute, for the award of any Degree or Diploma.

(Dr. S. M. K Quadri)
Supervisor and Head
Department of Computer Science,
University of Kashmir,
Srinagar 190006

Dated: August 10, 2012

Dedication

Verily, when he intends a thing, His command is, “Be”, and it is!

So glory to him in whose hand is the dominion of all things:

And to him will ye be all brought back.

This research is humbly dedicated to…

Allah, The Almighty

Acknowledgements

Alhamdulillah, all praises to Allah for the strengths and His blessings in completing this thesis.

I would like to thank all the people who have been supporting me during the past years while I have been pursuing my Ph. D. Programme. In any case activities of this nature always require not only support but also encouragement from different quarters.

I really fail to get words for expressing my gratitude to **Dr. S. M. K Quadri,** Director and Head, Post Graduate Department of Computer Science University of Kashmir, Srinagar who honoured me with his willingness to act as the guide. He made this work possible by encouraging me regularly to follow a research direction in which I am most interested. Despite her heavy workload, he has taken specific care to maintain the technical expertise.

I am indebted to **Prof. Mushtaq Ahmad Peer,** former Director and Head, Department of Computer Science, University of Kashmir, for his encouragement to pursue my research and technical support which he has provided during my research work.

Words are inadequate to express my gratitude to Late **Er (Dr.). Mehrajud Din,** Former Director, Information Technology and Support Systems, University of Kashmir, Srinagar for encouraging me to pursue higher studies. He has been consistently encouraging me to pursue the research and has been extending all the possible requisite support.

Words are inadequate to express my gratitude to my colleague and friend **Er. Muheet Ahmed Butt,** Scientist, Directorate of Information Technology, University of Kashmir who has been always at my back to support and without his moral support and encouragement, this work would have definitely been affected.

I am also grateful for the feedback I received from other colleagues of the University of Kashmir in particular Er. Manzoor Ahmad, Dr Javed Pervez and Mr. Sajad M. Khan of the P.G. Department of Computer Science and Mohd Rafi Khan from the Department of Management Studies.

I would like to extend my sincere thanks to Mr. Mohammad Shafi Mir, Section Officer, P.G Department of Computer Science, University of Kashmir for providing help and support at official level which ensure timely completion of this research.

Giving the highest place of honour along with warm feelings, I place on record my heartfelt gratitude as well as reverence to my parents and my wife.

Majid Zaman

Contents

List of Figures

List of Tables

List of Algorithms

Glossary

Data

Is Defined as the collection of unorganized facts for figures

Information

Is defined as the collection of Organized Data.

Information System

A combination of hardware, software, infrastructure and trained personnelorganized to facilitate planning, control, coordination, and decision making in an organization.

Encryption

Scramblingsensitive information so that it becomes unreadable to everyone except the intended recipient. Encryption is performed by hardware/software devices which use a series of mathematical operations (encryption algorithm) to generate encrypted data called cypher text. The recipient, who must have an 'electronic key' to decrypt the data, uses a similar device to convert cypher text back to its original (readable) form called clear text.

Dimension

One of the aspects, attributes, elements, or factors that make up an entity, item, phenomenon, or situation.

Key

In cryptography, a key is a piece of information (a parameter) that determines the functional output of a cryptographic algorithm or cipher

Data Mining

Sifting through very large amounts of data for useful information. Data mining usesartificial intelligencetechniques, neural networks, and advanced statistical tools (such as cluster analysis) to reveal trends, patterns, and relationships, which might otherwise have remained undetected. In contrast to an expert system (which drawsinferences from the given data on the basis of a given set of rules) data mining attempts to discover hidden rules underlying the data. Also called data surfing.

Legacy System

Obsoletecomputer system that may still be in use because its data cannot be changed to newer or standardformats, or its application programs cannot be upgraded.

Data Mart

Scaled-down and simplified version of a data warehouse, more suitable for smaller organizations.

Data Warehouse

Massive database (typically housed on a cluster of servers, or a mini or mainframe computer)

	serving as a centralized repository of all data generated by all departments and units of a large organization. Advanced data miningsoftware is required to extract meaningful information from a data warehouse. The term was coined by the US consultant W. H. Inmon.
Normalization	Databaseprogramming: Process of reducing a complexdata structure into its simplest and most stableform by eliminating redundantattributes, keys, and relationships.
Decision Support System	Computer system designed to provide assistance in determining and evaluating alternative courses of action. A DSS (1) acquires data from the mass of routinetransactions of a firm, (2) analyzes it with advanced statistical techniques to extract meaningful information, and (3) narrows down the range of choices by applying rules based on decision theory. Its objective is facilitation of 'what if' analysis and not replacement of a manager'sjudgment.

Abstract

There is decent amount of standardization as far as World-Wide Web is concerned, while Google is universal access tool to search and determine source of the information user requires there is still no such tool that can be implemented at enterprise level where there are multitude of data sources and organization users are still facing difficulty in accessing data available on the intranet of the organization and not on the WWW, in order to access such data users within the organizations need to know a lot including location, access techniques etc while still data consistency & redundancy is beyond the scope of common organization user/s.

Data Retrieval is still a pervasive challenge faced in applications that need to query across multiple autonomous and heterogeneous data sources. Data integration & extraction is crucial in every enterprises that own a multitude of data sources, where data sets are being produced independently by various departments of the organization, for better cooperation among various departments within the organization, data retrieval technique that cuts across heterogeneous data sources and is independent of size and access techniques required to access data sources as well as does not require user to have any sort of knowledge that includes location, query language etc, is need of hour.

Solution making use of Knowledge base where in users of the organization irrespective of their technical ability, data source knowledge and location can search heterogeneous data sources including legacy data sources of organization and retrieve information, also taking into consideration user attributes like his/her location, work profile, designation etc so as to make search more relevant and results more precise

The basic standard of data integration is to combine (integrate) designated information sources from a particular domain, in a way that a whole new data Source is generated. The end user, when querying for data, has the impression of interacting with one single system, which presents him a combined logical view of the data available. The first efforts to address information integration issues in enterprises where based principally on data warehousing methods, however proposed architecture, provides the user with a combined view of all data, on which queries can be posed. This particular schema is not intended to store any data; it is purely a logical schema.

20th century resulted in accumulation of two things-wires and data, while both brought enormous success to organization in specific and information technology in general, 21st century is all about management. Industry realized need to get rid of wires and integrate/manage/transform data present everywhere around us. Fiber & Wi-Fi is replacement to wires; however data integration/management is still challenge at large because of varying underlying structure, format, operatingsystem etc. proposed method of data transformation at application level without having to modifying underlying structure of data storage.

Once the data has become machine readable in an Information System many issues related to its security arise. Although the information systems provide users greater access than ever to vast information resources, however they are equally subject to threats that jeopardize the privacy and confidentiality of sensitive information, the integrity of data, and the availability of critical information system resources. Protecting information and the resources that process and maintain information is critical to the continuity of operations. Security of information resources must include controls and safeguards to offset possible threats as well as controls to ensure timeliness, availability, integrity, confidentiality, etc.

Information technology (IT) security encompasses the total infrastructure for maintenance and delivery of information, including physical computer hardware, supporting equipment, communication systems, and logical processes defined by software, procedures etc.

Data security is proposed using encryption and University Registration System is taken as case study in understanding the system. For the access control of the system, the built in Database Control features are used however the user operations are monitored and controlled by introducing the process of encryption. We propose a model for system security involving the available database security features and the encryption technique. The model is designed specifically for preventing un-authorized modification of the data by its users which enjoy different levels of authorizations.

1.1 Introduction

With almost every enterprise ranging from small medium to large deploying database applications for efficient storage and retrieval of process specific data, the databases have grown out in volume and have much more data. The optimum utilization of this data will happen only when every end user can get the data, which he/she needs, and when needed. But still the process of culling out useful information from database is in the purview of the designers, programmers or managers who are well versed with database specific query languages. These languages scare away naive end users who are left at the mercy of the programmers who provide them with limited predefined queries to extract information from the database. But the scope of these queries is generally restricted to routine information and any new queries have to be designed by programmers and handed over to end-users. Moreover the end users are not at all aware of the underlying schema, which can help them design queries if at all they know some querying language. As data storage techniques improved over the years organizations rapidly opted for the change and not only collected large volumes of data but in different sources and formats (Fig 1.1) e.g. .txt, .xml, .dbf, .html etc.

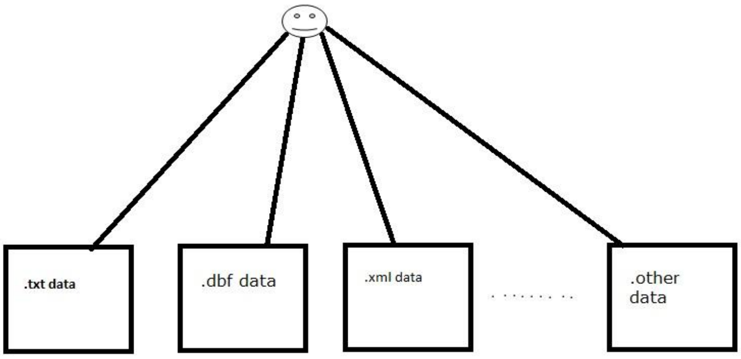


Fig 1.1: Generic Data Distribution in an Organization

Querying relational databases requires users to be aware of the underlying database schema and also the knowledge of the structured query language specific to the database. Most of the business database needs lots of reports to be derived/extracted from the underlying data for analysis and decision making purposes. This is accomplished by the use of pre-designed formats that accepts specific input and generates output in the specified format. There is little flexibility in viewing results in different order or seeing further details if required. This approach is fine when the information required by the user is served by the pre-designed forms and serves the purpose well. But if the user wants to see the reports based on relationship between different types of entities for which there is no direct correlation, then the form based approach doesn’t works well or alternatively it would require designing numerous formats.

1.2 Motivation

With use of popular search engines for information retrieval, the keywords based search has become defacto standard of extracting information from the wide sea of information. This is done by specifying a string of keywords and relevant documents ranked in the order of relevance are served to the user. Since a lot of information in the organizations is stored in the databases (and not as HTML documents), it is important to provide similar search paradigm for databases where users can query information without the knowledge of underlying database schema and query languages, data is stored in different database formats where data storage and retrieval techniques are different and no universal generic rules are applicable.

1.3 Goals & Objectives

Data is not only stored in databases but also in different file formats (Fig 1.1)e.g. .txt, .html, .xml etc. were data retrieval rules are altogether different and vary from format to format. Retrieving data from files requires users to be aware of the underlying file schema and also the knowledge of the retrieval method used, specific to the file format.

User is also required to have data storage knowledge, as to where data of his/her interest is stored. It becomes very complicated as there can be m files and n tables. User is then either supposed to depend on programmer or memories where data of his interest is stored, things get over complicated because of data replication- same data can be stored in n file and n table at the same time. User does not want to depend on programmer neither have patience to memories where data is stored, user expects to give query without specifying where data is stored and in which format it is stored

It’s not only about data retrieval but also about data translation. Data retrieved from databases are all together in different form as compared to data retrieved form files e.g. .txt, .xml etc. Users are interested in having data presented in specific form and not according to their storage pattern, in other words data stored in database will be say in columnar form were as data stored in .txt file will be all together in different form as compared to .html file, irrespective of data storage pattern and data retrieval technique data has to be presented in user desired format, where in user can decide the format himself/herself. Data is presented in user desired format and this technique is not dependent on security, while data has to be integrated from the sources this can lead to security lapse, Algorithm form integration should ensure data is secure and information is not compromised.

1.4 Thesis Outline

Chapter 2 is Data Accumulation and Retrieval where in principle are laid down for integration of data spread across the organization in heterogeneous data sources and algorithm for data retrieval from n heterogeneous data sources is proposed.

Chapter 3 is Data Migration and Information Translation where in issues of data migration from legacy sources is dealt upon and algorithm for information translation is proposed.

Chapter 4 this chapter deals with security, algorithm for encryption/decryption is proposed.

Chapter 5 is Adaptation System where in Artificial Intelligence based Generic Search Optimization for Heterogeneous Data Sources is proposed.

Chapter 6 is Conclusion and future work where in conclusion is made on present and attempt is made to define area of future work.

1.5 Review of Literature

Many organizations extract data from their databases and transform them into HTML pages. These pages are then available over the web. This approach is proper for static data, as it requires the whole process to be continual whenever the database contents change. The other method, uses application-specific programs to run parameterized SQL queries and dynamically create HTML pages comprising required information. Limits of this approach have already been mentioned above.

Open source software teams frequently develop complex software products in frequent-release settings with somewhat light weight processes and project documentation. In this situation project a major challenge for information collection is how to extract the pertinent project management knowledge successfully and efficiently from a varied range of software project data sources, such as artefact versions, bug reports, and discussion forums. In their paper Stefan Biffl, Wikan Danar Sunindyo, Thomas Moser, [Stefan Biffl, 2010] they presented a context and tool support for the semantic integration of data from a multiplicity of data sources to simplifyeffective data collection, even in projects with frequent iterations. Based on data from real-world use cases in open source projects we compare the efficiency of the proposed framework with a old-fashioned data warehouse approach. Major result is that the proposed method can make data gathering for project monitoring about 30% - 50% more effective, in particular, in situations where heterogeneous information sources change during the project.

The World Wide Web (www) aids a huge, extensively distributed, global information facility. Much information presents in the form of a web record which exists in both feature and list pages. Due to the upsurge of online web databases, it is obligatory to get useful required information which is to be structured before presenting the users which is one of the web information extraction (WIE) tasks. Using web query interfaces information is retrieved and is enwrapped in the web pages in the form of data records. There are large numbers of manually constructed, supervised, semi supervised and unsupervised WIE systems are planned and developed. The job of mining records from web pages is done by a software agent called as wrapper. The process of leaning a wrapper from a collection of similar pages is called wrapper induction R. Ashok Kumar, Dr

Y. Rama Devi [R. Ashok Kumar, 2011], have debated the different methods for wrapper induction at record level data.

Tari, L. Tu, P. Hakenberg, J. Chen, Y. Son, T. Gonzalez, G. Bara[Tari, 2010], designate a novel approach for information extraction in which extraction needs are communicated in the form of database queries. Using database queries for information extraction permits generic extraction and minimizes reprocessing of data by execution incremental extraction to find which part of the data is affected by the change of components or goals.

Masermann and Vossen [Masermann,2000][G. Vossen,2000] describe their study where the goal is to offer a simple, schema-independent web interface to relational databases, where queries, containing of a few keywords (as in search engines), can be framed in a declarative fashion. Whereas the standard SQL requires knowledge of tables and their attributes (i.e., schema definitions), they make use of parameters in SQL prototypes as placeholders for relations and attributes. When applied to a specific database, given its schema, they dynamically produce correct SQL statements from the parameters. These queries can then be executed, and results formatted suitably for display to the user. The generation of queries is based on an SQL extension called Reflective SQL for handling technical data. Thus, in their method, user query is first interpreted into Reflective SQL, and the resulting expressions are then executed on the underlying database. They generate SQL queries based on a prototype that matches attributes of relations for the given keyword. The prototype is applied to a data dictionary that gives table names, column names and column types. Their method is quite general and declarative. It only uses schema information, and does not build extra indexes or copy the database data (into a graph, etc.). No re-building of any middle data structures is required in case the database is updated by the application. However, there are certain limitations to their approach. They do not take into explanation other data semantics for making more relevant results. They do not consider ordering results based on some notion of significance. They also do not use vocabulary or any other information to relate keywords to the table attributes, but instead make all possible queries based on only attribute types. Finally, the query outcomes are in a pre-defined format that is difficult to comprehend and is not navigable.

Md. Sumon Shahriar and Jixue Liu [Md. Sumon Shahriar, 2010] proposed how data from diverse source information systems can be changed to a global information system. They also reviewed how constraints in data conversion are used in data integration for the purpose of integrating information systems. Their research was towards the handling of semantics using integrity constraints in data integration from heterogeneous information systems.

Shafer and Agrawal [Shafer, 2000] propose a web-based interface and an application (called Eureka) for communicating exploration of databases, where, instead of forming exact SQL queries, the user browses through the information, places filtering predicates on attributes, or selects ‘example’ records for recovering similar other records. Their interface seamlessly assimilates continuous queryingwith result-browsing. The interface is intuitive and effective in applications (such as ecommerce) where users browse with a purpose. However, it lacks the power and demand of keyword-based search that may go beyond multiple tables. For good performance, they need to preserve complex data structures for caching results on the client side.

Goldman et al [Goldman, 1998] lengthen the textual proximity quest paradigm for searches within databases. A database is viewed as a graph with objects as nodes and relationships as edges. Relationships may be defined built on the construction or meaning of the database. They define contiguity based on the shortest distance between the objects. The prototype lets queries to find objects of interest which are near to another set of objects. The objects found are graded on their proximity.

Bhalotia et al [Bhalotia, 2002] describe an extension to the work by Goldman et al [Goldman, 1998] and give an efficient implementation for searches in relational databases. In their graph representation of a database, tuples are nodes, and edges capture foreign key and primary key relationships. They propose proximity based not only on shortest path, but also on the presence of other paths and in/out degrees of the nodes.

PESTO [PESTO, 1996] by Carey et al is a tool for querying/browsing in object databases. It shows objects as windows on the screen, and facilitates moving through group of objects and navigating along (reference) links amongst them. It supports 'query-in-place', by which filters can be specified for choosing objects as per a specific condition. It offers many useful features like 'synchronous browsing', last-query modification, etc. The tool is schema-aware, and its search metaphor is not keyword-based searching.

Gey et al [Gey, 1999] identify a typical problem met by end-users in searching a database. The users are frequently not aware of how data are classified, categorized, abbreviated, named and represented in the database. They propose use of "vocabulary modules" to tie the gap between the user's usual language, and the database system's metadata and stored data. They propose an agent-based design to develop domain specific vocabularies.

Software infrastructures and applications more and more must deal with information available in a range of different storage engines, accessible through a host of protocols and interfaces; and it is common that the size of the information involved involves streaming-based processing. In their article Marc Van Cappellen, Wouter Cordewiner, Carlo Innocenti [Marc Van,2008], showed how XQuery can influence the XML Data Model to abstract the data physicaldetails and to offer optimized processing allowing the growth of highly scalable and performance data integration solutions.

Toyama and Nagafuji [Toyama, 1988] defines an extension to SQL for making results as an HTML (or, another type of) text. The extension permits defining intra-page and interpage ranked structures for surfing and steering. The user must use SQL and the extensions to retrieve the obligatory database objects. A keyword based search for the database contents is not the objective here.

For a sensor network including autonomous and self-organizing information sources, capable similarity-based search for semantic-rich resources (such as video data) has been deliberated as a challenging task due to the lack of infrastructures and the multiple limitations (such as band-width, storage and energy). While the past study discussed much on routing protocols for sensor networks, few works have been reported on effective data retrieval with respect to enhanced data search cost and fairness across various environment setups. Bo Yang and Manohar Mareboyana [Bo Yang, 2009] their study offered the design of reformist content prediction approaches to ease efficient similarity-based search in sensor networks.

DbSurfer indexes the textual content of each relational tuple as a virtual web page. For querying and navigation, the database foreign-key constraints between tuples are canned as hyperlinks between virtual web pages. Given a keyword query, the scheme computes a ranked set of virtual web pages that match at minimum one keyword. Then a best-first development algorithm finds a ranked set of steering paths originating from the starting web pages. However, even yet the web pages are indexed offline, substantial query time computation is required to expand the starting web page set to find navigation paths satisfying the full query, work that we perform offline.

Three systems, DBXplorer [DBXplorer,2002], BANKS [BANKS,2002], and DISCOVER [DISCOVER,2003], share a alike approach: At query time, given a set of keywords, first find tuples in each relation that comprise at least one of the keywords, usually using secondary indexes. Then use graph-based methods to find tuples among those from the previous step that can be joined together, such that the joined tuple contains all keywords in the query. All three systems use foreign-key relationships as edges in the graph, and point out that their approach could be extended to more general join conditions.

2.1 Introduction

A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources. In addition to a relational database, a data warehouse environment includes an extraction, transportation, transformation, and loading (ETL) solution, an online analytical processing (OLAP) engine, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users [Kimball, 1996][Bauer,2001]. A data warehouse is an asset for an enterprise and exists for the benefit of an entire enterprise including business unit, individual customer, Student etc. Data in a data warehouse does not conform specifically to the preferences of any single enterprise entity. Instead, it is intended to provide data to the entire enterprise in such a way that all members can use the data in the warehouse throughout its lifespan [Jeff Lawyer, 2004].

Data warehousing systems enable enterprise managers to acquire and integrate information from heterogeneous sources and to query very large databases efficiently. Building a data warehouse requires adopting design and implementation techniques completely different from those underlying information systems [Bernardino, 2002]. Issue of information integration is vital part of any data warehouse and the design challenge increases by the variety of heterogeneous data sources present in the system. These data sources have varying conceptual models, semantic heterogeneity which seems to be an unavoidable burden in data integration.

A 'data warehouse' is often termed as repository of an organization's electronically stored data. Data warehouses are designed to facilitate reporting and analysis [Inmon, W. H, 1996]. This classic definition of the data warehouse focuses on data storage. However, the means to retrieve and analyze data, to extract, transform and load data, and to manage dictionary data are also considered essential components of a data warehousing system [Larry, Greenfield, 1997]. These operations depend more on the way the data is stored.

There are two leading approaches to storing data in a data warehouse

i. Dimensional approach and
ii. Normalized approach

In the dimensional approach, transaction data are partitioned into "facts", which are generally numeric transaction data, and "dimensions", which are the reference information that gives context to the facts [Kimball, Ralph, 1996]. A key advantage of a dimensional approach is that the data warehouse is easier for the user to understand and to use. The retrieval of data from the data warehouse also tends to operate very quickly. The main disadvantages of the dimensional approach are:

iii. in order to maintain the integrity of facts and dimensions, loading of data from different operational systems is complicated, and
iv. it is difficult to modify the data warehouse structure if the organization adopting the dimensional approach changes the way in which it does business.

In the normalized approach, the data in the data warehouse are stored following, to a degree, the Codd normalization rule. Tables are grouped together by subject areas that reflect general data categories. The main advantage of this approach is that it is very easy to add information into the database. A disadvantage of this approach is that because of the number of tables involved, it can be difficult for both users to join data from different sources into meaningful information and then access the information without a precise understanding of the sources of data and of the data structure of the data warehouse.

These approaches are not exact opposites of each other. Dimensional approaches can involve normalizing data to a degree [LIN Yu, 2003]. In this paper we have implemented a Star Schema Model of a Data Warehouse of an Central Automation of Examination System catering many colleges, Departments, Courses, Subjects, Subject Groups, Marks and tried to prepare results notifications at various levels which will enable us to build a build a Decision Support Database for future analysis.

2.2 Data Sources

Various Data Sources are debated in the context of research below

2.2.1 Data Mart & Data Warehouse

A data mart is a simple form of a data warehouse that is focused on a single subject (or functional area), such as Examination, Registration, orAccountsetc[Kimball,1996][http://docs.oracle.com/html/E10312_01/dm_concepts.htm, 2012]. Data Marts are often built and controlled by a single department within an organization. Given their single-subject focus, data marts usually draw data from only a few sources. The sources could be internal operational systems, a central data warehouse, or external data. A data warehouse, unlike a data mart, deals with multiple subject areas and is typically implemented and controlled by a central organizational unit such as the corporate Information Technology (IT) group. Often, it is called a central or enterprise data warehouse. Typically, a data warehouse assembles data from multiple source systems.

Nothing in these basic definitions limits the size of a data mart or the complexity of the decision-support data that it contains. Nevertheless, data marts are typically smaller and less complex than data warehouses; hence, they are typically easier to build and maintain [http://docs.oracle.com/html/E10312_01/dm_concepts.htm, 2012].

Each Data Mart can contain different combinations of tables, columns and rows from the Enterprise Data Warehouse [Date, 1995][W. J. Labio, 1997]. For example, a business unit or user group that doesn't require a lot of historical data might only need transactions from the current calendar year in the database. The Personnel Department might need to see all details about employees, whereas data such as "salary" or "home address" might not be appropriate for a Data Mart that focuses on Sales [Jorge Bernardino, 2002].

2.2.2 OLAP and Data Mining

On-line Analytical Processing (OLAP) is the analytical capabilities provided by the data warehouse or data mart. One can view granular data or various aggregations of data for business analyses using graphical-user-friendly tools [CAI Yong, 2003][YUAN Hong, 1999]. Data warehouse and data marts exist to answer questions and find business opportunities. There are many ways to analyze data using procedures such as Proc decodeMks, Proc getResult, Proc fmaster, Proc rollidx, Proc Tabulate.

Finally, data mining is the name given to newer statistical techniques used to explore voluminous data stores. These techniques include decision trees and neural networks. These methods, like neural networks, can sometimes handle co-linearity better than the older statistical techniques.

2.2.3 Distributed and Decentralized Data

Data storage is organized in a completely decentralized manner and information retrieval might involve querying multiple data sources. In large enterprises for example, where decisions are usually made based on data observations, each department might keep its own database system (HRM, Finance, Sales etc). Therefore accumulated information for the whole enterprise requires data combinations from various sources. The decentralization is further enhanced if we also take into account possible partners, vendors or competitors, whose data might be of company's interest. Another area, where this decentralization worsens information retrieval is large scale scientific projects. Here not only the volume of data, but also the complexity has to be taken into account. Scientists nowadays, besides profound domain knowledge, require access to data and results provided by others. Therefore, querying individually different data sources leads to significant inefficiency in their work. Finally, for an effective search in Enterprise, user is required to look up for information in multiple data sources and collect the data individually.

2.2.4 Heterogeneity of Data Sources

In addition to the decentralization, the effectiveness of information retrieval is further worsened by the variety of heterogeneity present in the data sources. In each of these sources, data is organized using a different system (operating systems, SQL Vendor Implementations etc), based on different [ETH Group, 2010] conceptual models, and on different formats "system level heterogeneity" is considered to be nowadays much easier than before (e.g. via ODBC/JDBC connections on relational databases), much interest is laid on the so called “semantic heterogeneity”, which appears every time there is a more than one way to structure a body of data. Semantic heterogeneity seems to be an unavoidable burden in data sharing and manipulation, since people tend to model their data according to their own understanding of the reality. This of course is fundamentally different for each individual. In that sense, heterogeneity is to be found in data models, conceptual schemas, and of course the mind of the users [ETH Group, 2010].

2.3 Schémas in Data Warehouse

A schema is a collection of database objects, including tables, views, indexes, and synonyms. There is a variety of ways of arranging schema objects in the schema models designed for data warehousing. The main database Schemas are:

2.3.1 Star Schemas

The star schema is perhaps the simplest data warehouse schema. It is called a star schema because the entity-relationship diagram of this schema resembles a star, with points radiating from a central table [Fon Silvers]. The center of the star consists of a large fact table and the points of the star are the dimension tables. A star query is a join between a fact table and a number of dimension tables. Each dimension table is joined to the fact table using a primary key to foreign key join, but the dimension tables are not joined to each other. The optimizer recognizes star queries and generates efficient execution plans for them. It is not mandatory to have any foreign keys on the fact table for star transformation to take effect. A star join is a primary key to foreign key join of the dimension tables to a fact table. The main advantages of star schemas are that they:

- Provide a direct and intuitive mapping between the business entities being analyzed by end users and the schema design.
- Provide highly optimized performance for typical star queries.
- Are widely supported by a large number of business intelligence tools, which may anticipate or even require that the data warehouse schema contain dimension tables.

Star schemas are used for both simple data marts and very large data warehouses.

2.3.2 Snowflake Schemas

The snowflake schema is a more complex data warehouse model than a star schema, and is a type of star schema [Fon Silvers]. It is called a snowflake schema because the diagram of the schema resembles a snowflake. Snowflake schemas normalize dimensions to eliminate redundancy i. e., the dimension data has been grouped into multiple tables instead of one large table. While this saves space, it increases the number of dimension tables and requires more foreign key joins. The result is more complex queries and reduced query performance. The main advantages of Snowflake schemas are that they:

- save memory space for data.
- increases the number of dimension tables and requires more foreign key joins.
- the result is more complex queries.

2.3.3 Third Normal Form (3NF)

Third normal form modeling is a classical relational-database modeling technique that minimizes data redundancy through normalization [Fon Silvers]. When compared to a star schema, a 3NF schema typically has a larger number of tables due to this normalization process. 3NF schemas are typically chosen for large data warehouses, especially environments with significant data-loading requirements that are used to feed data marts and execute long-running queries. The main advantages of 3NF schemas are that they

- Provide a neutral schema design, independent of any application or data-usage considerations
- May require less data-transformation than more normalized schemas such as star schemas

2.4 Data Warehouse based Data Integration: University of Kashmir Case Study

To remain competitive in today's academic climate, any University needs a foundation of quality data that too in examination system where the data need to be more precise and consistent. Organizations of higher education need this capability as much as Fortune 500 companies. Organizations, both governmental and business, have to manage large amount of information stored in some form of databases or files [Jorge Bernardino, 2002][F. Bancilhon, 1981]. One of the main problems to deal with information managing is the weak interoperability between various databases and information systems. Especially this problem is serious when we want to organize collaboration between the information systems of various departments within the organization. Solution is data-warehousing, the idea of drawing data from several different (heterogeneous) data sources on different platforms and computers. Building a data warehouse is extremely complex and takes commitment from both the information technology department and the business analysts of the organization. It takes planning hard work, dedication, and time to create a relational database that delivers the right data to the right user. University of Kashmir's Examinations data warehouse (UOKEDW) is not a panacea for every data problem, but it is a very good start toward a permanent solution.

University of Kashmir, like any academic organization in North India started realizing need and importance of computerization in late 1980's and early 1990s; however it was not those days that we were provided with RAD tools. To just meet with the dynamic changing scenario of Information Technology an in house computerization unit was established which mostly focused on designing and developing software solutions for University Automation System [Marotta, 1999]. It mostly involved development of those software products which were used to meet day to day requirements of various sections and departments of the University like examination Wing, Registration System, Library etc. in an efficient manner. At beginning of the development process most of the development was carried out in FoxBASE, dbase & FORTRAN as programming platforms, Unix and Windows 9x as operating systems, three teams were set up initially to computerize the Examination related processes following systems.

As year passed on, information explosion within the organization was at its peak, it was not much long before university administration realized that in order to manage data in an efficient manner and provide friendly access to information, more new teams needed to be setup with enhanced budget in order to store/manage data and information in an efficient way to meet the changing demands.

Newly constituted teams were given decent budget along with freedom to choose the tools required for development, within no time university was riding on with the success of Information Technology and almost all the areas of the university administration were computerized and autonomous solutions within single organization each working fine but with no possibility of integration.

As the University and other academic organizations, both governmental and business, have to manage large amount of vital information stored in some form of databases or files. One of the main problems to deal with information managing is the weak interoperability between various databases and information systems. Especially this problem is serious when we want organize collaboration between the information systems of various departments within the organization [J Gray, 1993], therefore the Development which was carried out in the University lagged integration of data from various heterogeneous data sources which were present at various operational levels of the University.

Solution is integration of such heterogeneous data sources. The proposed solution address most of the issues pertaining to efficient storing and retrieving of this vital information using a data-warehouse in which the idea is of drawing data from several different (heterogeneous) data sources from different platforms and sources into a single data warehouse having many data marts.

2.4.1 Towards Computerization

University of Kashmir Examination Department declares 450 evaluation and 450 re-evaluation results of around 3.5 Lakh enrolled students appearing in both under and Post Graduate Examinations of around 200 affiliated colleges. University of Kashmir, like any academic organization in India started realizing need and importance of computerization in late 1990 and early 2002, however it was not those days that we were provided with RAD tools. In house computerization was done in FoxBASE, dbase and SCO Unix as operating system, three teams were set up initially to computerize following systems,

- Tabulation
- Transit
- Secrecy
- Certificates
- Accounts
- And Conduct

As year passed on, information explosion within the department was at its peak, it was not much long before university administration realized that in order to manage such a crucial data of the Examination Department and provide friendly access to information, new teams needs to be setup with enhanced budget be provided in order to store/manage data to meet the changing demands.

Newly constituted teams were given decent budget along with freedom to choose the tools required for development, within no time university was riding on with the success of Information Technology and almost all the areas of the university administration were computerized-n autonomous solutions within single organization each working fine but with no possibility of integration.

2.4.2 UOK Examinations Data Warehouse Development

Development of UOK Examination data warehouse started in the in around 2004 as a client/server project. The tools which were mainly used in the first phase of development were Microsoft SQL Server 2000 and Windows Server 2005 workstation. Every software development carried out on these platforms was done in home. While getting the warehouse server in place, the software development team provided data access software for efficient usage of the warehouse in implementation of both read and write operations [Krishna, 2005]. Although many of the access tools were in their adolescence at the time, accessing data was much easier with these graphical user interface (GUI) tools than with the fourth-generation tools then in use.

University of Kashmir formed a development team of two Project leaders with ten Master of Computer Applications final year students and Computing Assistants from the data administration and Examination Automation Centre to build the data warehouse. The team selected a representative group of business analysts to serve as pilot users to test the warehouse and access software. During the next few months, the team built a “student” warehouse model based on over 150 questions, which the pilot users considered difficult or critical to answer using current information resources [A. Bauer, 2001][J. Lechtenb orger, 2002].

During 2006, many of the original data warehouse team members shifted back to their regular duties, leaving a core of six fulltime equivalent employees working on the project. That core has remained intact, receiving additional help from UOK's institutional research office and many of the business analysts who are regular users of the warehouse. Also, the data administration department initiated a formal program to train users on the warehouse. To date, there are over 250 trained warehouse users which have been trained while carrying out training programs for various employees. The major goal is to train 400 odd employees, approximately 10 percent of UOK's employees.

2.4.3 Design

2.4.3.1 Data Mart Design

Let us take an example of a Registration System and Examination Automation System. Existing Registration Data Mart has Registrar Dimension, Course Dimension, Subject Dimension, Transaction Dimension, Log Dimension, REG10 Dimension, REG11 Dimension, REG12 Dimension etc. Table Reg10 & Reg11 are list of students having submitted registration fee for batch 10 & 11, every year this table is created which stores information of list of students having admitted for current year. The table 2.1 below shows the various data marts dimension that are used in designing the data warehouse of the University.

Registration Data Mart	**Examination Data Mart**	.
Registrar	Enrolment	
Course	Theory Marks	
Faculty	Practical Marks	
College	Result	
University	Subject	
Log	Course	
Transaction	Faculty	
REG11	College	
REG12	…	
…	…	

Table 2.1: Data Marts for Registration & Examination

Existing Examination database Enrolment Dimension, Theory Marks Dimension contains theory marks, Practical Marks Dimension contains practical marks, Subject Dimension, Course Dimension, Faculty Dimension and the final Result Dimension which contains the final result. On the same line other marts are created, warehouse administrator has capability to pick and choose data from various data sources. These will be followed by performing ETL on these newly created Data Marts. The Examination and Registration Data Marts is shown in the figure below

2.4.3.2 Data Warehouse Design

In our solution the data integration into a single data warehouse with data marts is the prime focus. The solution thus developed is using Linux-Operating System, Oracle Database Management System, Apache Tom Cat Web Server, Java/JSP server with above mentioned software configuration is set up. This server is part of CAN where in it has been given a controlled access to all existing solutions in university e.g examination, registration , accounts etc in other words this server is made to perform all the ETL functions on all the heterogeneous data sources and place extracted data into data marts. The basic idea of this solution is to create Data Mart for every autonomous information source and then integrate these data marts to have single warehouse which could be named as Examination Mart, Registration Mart, Academic Mart, HRM Mart, Budge Data Mart, Accounts Data Mart etc.

University of Kashmir's is interconnected by a Campus Area Network (CAN) in which each operational section/Wing has a separate working software Application and Data Source. While as the examination solution has been developed on LAMP (Linux, Apache, My SQl, PHP), registration system is developed on Microsoft technologies MSSQL,.Net framework etc, at the same time budget is still very much legacy system developed in dBase 4.0. The newly created marts are connected to create a single warehouse as shown below in (Fig 2.1).

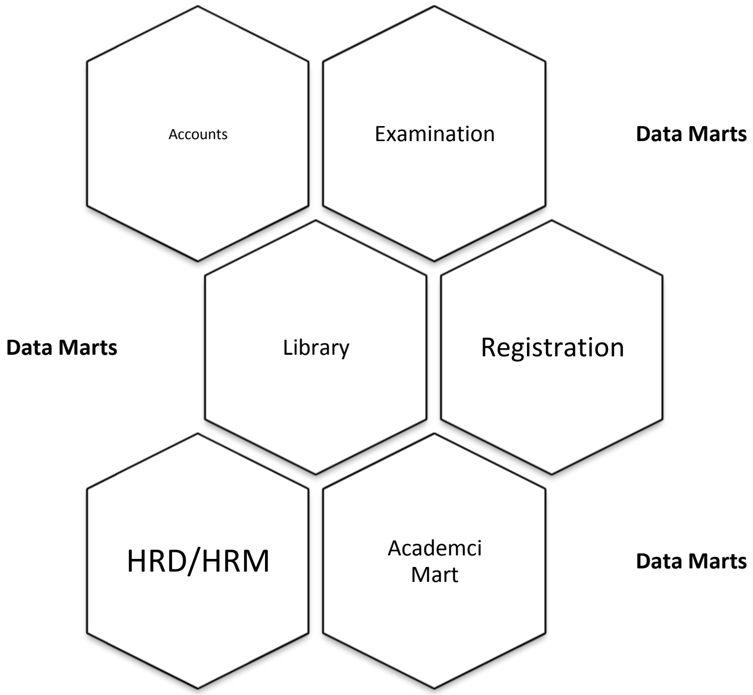


Figure 2.1: University of Kashmir Data Warehouse

2.5 Examination Data Warehouse implementation Phases

The following steps were under taken for designing of UOKEDW

I. **Requirement Gathering:** The first thing that the constituted project team was engaged in gathering requirements from the various employees working in the examination system. Because end users were typically not familiar with the data warehousing process or concept, requirement gathering was implemented using one-to-one meetings or as Joint Application Development (JAD) sessions, where multiple stake holders in the examination wing were interacted with so that the requirement analysis done in a proper manner.

II. **Physical Environment Setup:** Once the requirements gathering were somewhat clear, it became necessary to set up the physical servers and databases. At a minimum, it was necessary to set up a development environment and a production environment. It was not enough to simply have different physical environments set up [D.L. Moody, 1999]. The different processes many data warehousing projects where there were three environments: Development, Testing, and Production (such as ETL, OLAP Cube, and reporting) also need to be set up properly for each environment. The primary goal of this phase was to identify what constitutes as a success for this particular phase of the data warehouse project.

III. **Data Modeling:** This was a very important step in the data warehousing project. Indeed, it was fair to say that the foundation of the data warehousing system is the data model [Marotta, 1999].

A good data model will allow the data warehousing system to grow easily, as well as allowing for good performance. In UOKEDW project, the logical data model was built based on user requirements, and then it is translated into the physical data model.

IV. **ETL** (Extraction, Transformation, Loading) process typically took the longest to develop the UOKE's data warehouse implementation cycle as there were many heterogeneous data bases involved in extraction transformation and loading process. The reason for this was that it took time to get the source data, understand the necessary columns, understand the business rules, and understand the logical and physical data models before ETL would have been successfully carried out.

V. **OLAP Tube Design:** The OLAP cube was derived from the Requirement Gathering phase [Y. Zhuge, 1997]. The users working in the Examination Wing had some idea on what they want, but it was difficult for them to specify the exact report / analysis they wanted to see and anlyse. When this was the case, it is usually a good idea to include enough information so that they feel like they have gained something through the data warehouse, but not so much that it stretched the data warehouse scope by a mile. Hence front end development became an important part of a data warehousing initiative of UOK.

VI. **Front End Options:** The front-end options ranged from an internal front-end development using scripting languages such as VB, VB .NET, ASP, PHP, to off-the-shelf products such as Crystal Reports, to the more high-level products such as Actuate. When choosing vendor tools, it was made sure that it could be easily customized to suit the business of examination, especially the possible changes to the reporting requirements of the Examination System. Possible changes included not just the difference in report layout and report content, but also included possible changes in the back-end structure.

VII. **Report Specification:** Report specification typically came directly from the requirements phase [T.Sellis, 1999]. To the end user/employee working in the examination system, the only direct touch point he or she had with the data warehousing system is the reports they see and analyse. So, report development, although not as time consuming as some of the other steps such as ETL and data modeling, nevertheless play a very important role in determining the success of the data warehousing project.

VIII. **Query Processing** – In this the OLAP reports or reports were made to run directly against the RDBMS often exceeded the time limit, and it was hence ideal for the data warehousing team to invest some time to tune the query, especially the most popularly ones.

2.6 Examination Warehouse Infrastructure

UOK's Examination data warehouse resides in a client/ server environment. UOKEDW extracts data from loaded on to the Unix Server Majorly kept for Secure Data Entry Process for marks entry and loads it into a Microsoft Windows server running an MS SQL as RDBMS. UOKEDW server is a IBM Xenon Server with 8 GB of memory and two processors, running the Windows 2008 Server operating system. Users connect through Ethernet to the warehouse over UOK Campus Area Network backbone via Transmission Control Protocol/Internet Protocol (TCP/IP). The suggested GUI data access has was first implemented in Visual Basic

6.0 now transformed to Visual Basic .NET 2008., which runs identically on the Windows. Microsoft Access® is another tool used mostly for data migration from one database Server to other. The process of using GUI tools to build structured query language (SQL) requests and bring the results back to a client machine. With client/server architecture, once the data are in the workstation, users "own" the data, cutting and pasting at will into their favorite software (e.g., spreadsheet, word processor, graphic tools) [Larry, 1997].

2.7 Text Based Information Integration for Heterogeneous Data Sources

The basic principle of data integration is to combine (integrate) selected information sources from a specific domain, in a way that a whole new data Source is generated. The end user, when querying for data, has the illusion of interacting with one single system, which presents him a unified logical view of the data available. The first attempts to address information integration issues in enterprises where based primarily on data warehousing techniques, however proposed architecture, is described graphically below in (Fig 2.3)[ETH Group, 2010].

Traditional solutions prescribe creation of new data source on Information integration from heterogeneous data sources, which is not cost effective as shown in (Fig 2.2) below.

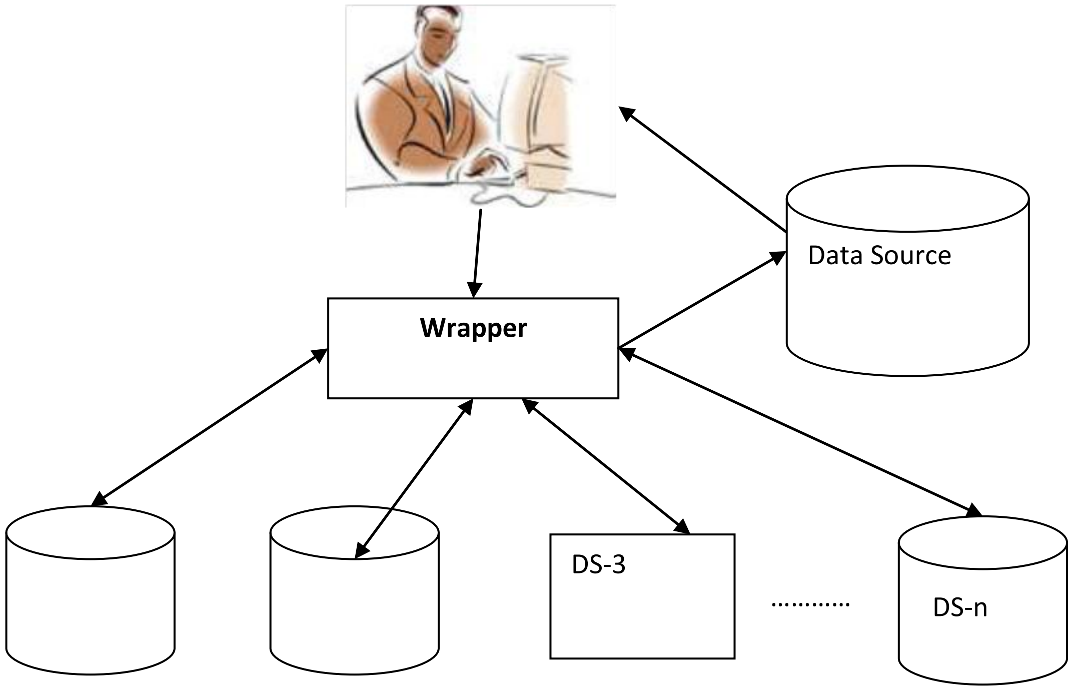


Figure 2.2: Data Integration Traditional View

Proposed Change

The schema provides user with an interface, user is not aware of number and/or structure of data source on which queries are to be executed. In given situation we can have n sources like

1. Database Oracle on Linux Operating System.
2. Database Mysql on Fedora Operating System.
3. Database MSSQL on Windows Operating System.

User inputs set of Keywords which are formulated into query and executed locally on autonomous data sources as shown in (fig 2.3), this particular schema is not designed to store any data; it is purely a logical schema. Hence, the formulation of keywords results in a set of source-specific queries Qi, whose combination will yield the answer to user input.

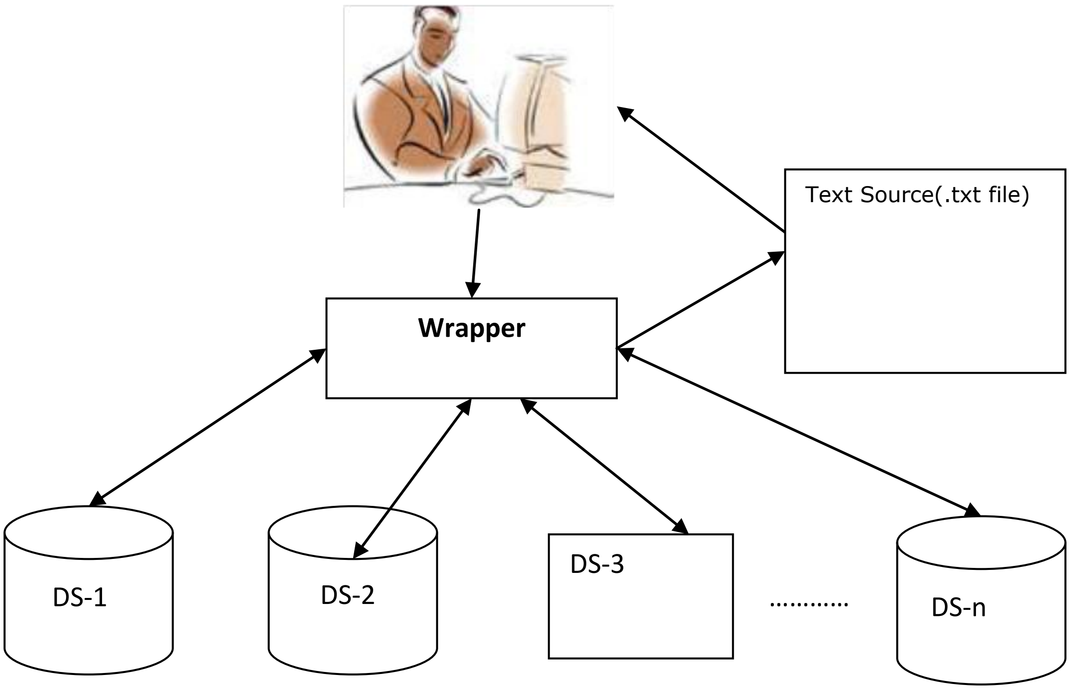


Figure 2.3: Data Integration Proposed View

2.7.1 Wrapper

Performs following tasks

2.7.1.1 Query Reformulation

Query formulation is main problem when it comes to retrieving data from multiple sources while query optimization being secondary problem. Generating queries from user input that are to be executed on multiple sources requires wrapper to have understanding of different database systems and operating system. The problem is that of data integration, where there is set of autonomous heterogeneous data sources. A user inputs key words describing his/her search criteria and data integration system needs to formulate queries to refer to the data sources. In a subsequent phase, the queries over the sources are optimized and executed. Query formulation algorithm has to be designed in such a manner that single user input results in generation of m queries each to be executed on different data sources at the same time optimizing query to best possible extent.

In some data integration applications, the number of data sources may be quite large – for example, data sources may be a set of web sites, a large set of suppliers and consumers in an electronic marketplace, or a set of peers containing fragments of a larger data set in a peer-to-peer environment. Hence, the challenge in this context is to develop an solution that scales up in the number of views.

As such, user query posed in terms of user input, and the data integration system needs to formulate the query to refer to the data sources. Since there are n heterogeneous data source, but user desired result may be present in m sources where n>m, as such it is the responsibility of Wrapper to identify m sources and prepare resultant m queries. In a subsequent phase, the queries over the sources are optimized and executed.

To minimize data retrieval, Wrapper generates precise queries, returning only the data that is needed. To avoid retrieving rows that are not needed, the conditions in Where clauses and predicates are converted to Where clauses in the generated SQL. To avoid retrieving columns that are not needed, the generated SQL specifies the columns actually needed by the user.

2.7.1.2 Data Transformation

Wrapper- sends queries to a data source, receives answers back, possibly applies basic transformations and creates new text source, this newly created text source is defined in accordance to the result generated as a result of executing query on heterogeneous data sources, and transformed data is stored in this text source, finally user is provided result from this text source and this text file is deleted and new text file will be created when user inputs new search criteria, as such system is not burdened because text files are small in size and at the same time are portable.

It is common that applications need to deal with information which is not obtainable in a single format; and that’s the environment where dealing with a sole query language, data model and interface which covers heterogeneous data sources becomes important. Think about a scenario, for example, where a list of auctioned ITEMs is accessible in an XML document, as but particulars about the person who’s proposing the ITEM are existing in a USERS table hosted on a relational database, including information about the user id, name, address and email. Now think about the need of making an application that given a user's email address retrieves all the items that are being auctioned by that user.

The wrapper consuming the result is aware of the physical origin of the data returned as a result of execution of queries even if the result mixes information stored in a relational database and in an XML document. Since data received by the wrapper are in different formats is transformed into generic format, extracted data is transformed and saved into text source, before saving in text format-text source is created in accordance with data retrieved as a result of execution of n queries on n heterogeneous data sources, definition includes column definition. Extracted, refined, cleaned, transformed, saved data in temp text source is passed onto user[ETH Group, 2010].

2.7 Conclusion

Although Data Integration was considered to be "an area of intellectual curiosity" [M. Vincini, 1999] at its early years, the advent of information sharing nowadays is calling for effective integration approaches realized in practice. Users are not compromising with low standards of information accuracy and are willing to find the right information at the right time. The research community, thus far, has shown excellent progress in dealing with the most crucial problems presented on the way of integrating data, however, further challenges arise constantly: The expansion of (semi -) & unstructured data (XML) for example implies that data sources are even more complex and difficult to handle. Coping with semantic heterogeneity in such scenarios seems almost impossible. However, research is getting even more intense and promising ideas are expected to develop[ETH Group, 2010].

Organizations across the globe while focusing on computerization of departments do not pay much stress upon uniformity and consistency of data [E. Bertino, 2001], and almost all the organizations across the globe ended up with numerous heterogeneous data sources. While data in warehouse must be credible, it must be carefully assembled from a variety of sources around the organization. Data Warehouse not only makes organization information easily accessible but has become tool for data integration.

The future of UOKE's data warehouse is becoming more clear. Initially, the warehouse served as a resource for accessing information from legacy systems. Eventually, the warehouse will serve as a telescope into UOKDW's distributed data stores. Some of these data will reside in the data warehouse, while other elements will be "viewed" from the RDBMSs where the data reside. UOKEDW foresees a time when the telescope extends beyond UOKEDW to other organizations with common goals, such as the neighboring Maricopa County Community College District. The real power of the warehouse will be actualized in years to come. The data warehouse fills an important data administration role in a client/server environment. As distributed application developers move further away from the central computing core, the data elements in the warehouse ensure the integrity of the organization's enterprise data.

The bottom line is that data warehousing is here to stay. Warehousing gives organizations the opportunity to "get their feet wet" in client/ server technology, distributed solutions, and RDBMS. This is essential for any future mission critical application, making the data warehouse a low-risk, high-return investment.

The data warehousing technology is gaining wide attention, and many organizations are building data warehouses (or, data marts) to help them in data analysis in decision for decision support. Data Warehousing is a newly emerged field of study in Computing Sciences. Due to its viz. multidisciplinary nature, it has overlapping area of studies in three different computing disciplines. This overlapping sometimes may cause contradictory definitions for a specific concept. To overcome this problem of data warehousing for Examination Automation System, it was considered for Star Schema Design. In this regard various functional dimensions of the Examination System were designed and connected to a Fact Transaction Dimension. Furthermore the general issues like the Client Statistics and Query Design were taken up and various Decision Support Databases were designed and implemented using the same star Schema.

3.1 Introduction

20th century resulted in accumulation of two things-wires and data, while both brought enormous success to organization in specific and information technology in general, 21st century is all about management. Industry realized need to get rid of wires and integrate/manage data present everywhere around us. Fiber & Wi-Fi is replacement to wires; however data integration/management is still challenge at large because of varying underlying structure, format, operating system etc. In this paper we propose/introduce various methods of data transformation at application level without having to modifying underlying structure of data storage.

Due to the growth of the number of employees and customers associated with an enterprise, many new business rules have to be implemented which are well served for the use of the computerized database. However flaws in the database have shown up from time to time caused by users who are inexperienced in managing Information Technology Infrastructure in a proper way or are doing it intentionally. Consider the case when two processes executing interfere with one another, thereby producing incorrect results. Sensitive data might be exposed-or worse, changed by unauthorized users. The fact is there are indeed many risks that the data might be exposed to. The other case concerns updates which might change data illegally[B.Thuraisingham, 1996]. Therefore the system has to provide an extensive support to protect the database against such threats which may arise while recovery, concurrency, security and integrity measures are being implemented on a database. For the stability of the database, using one central data warehouse for enterprise data source is perilous since some situations may adversely happen, such as system crash. Since only one central data warehouse is used, the situation may leave the database in an incorrect state in which recovery may not be possible. Consequently, one way to ensure a recoverable database is to certain that every piece of information it contains can be reconstructed from some other information stored in other place redundancy. The solution we provide is that we create another parallel Data Warehouse (TDW) which is used only during the transaction period of the enterprise. We will refer to this new Central Data Warehouse (CDW) as the central database as the main database throughout the paper.

With the advent of computerization primary goal of organization across the globe was automation of their working system, this resulted in massive collection of data in respect of organization business logic and process, not much was thought about integration of application and data. Once a blessing became huge problem in organizations, data all over the organization was becoming difficult to manage and inconsistency of data resulted in creation of team not meant for development but data management.

With the introduction of Data Warehouse which is majorly used to integrate data from many heterogeneous data sources which includes the working and archive data pertaining to the organization. It also includes multiple subject areas and is typically implemented and controlled by a central organizational unit such as the corporate Information Technology (IT) group often called as central or enterprise data warehouse. Data Warehouse integrates data from heterogeneous/homogeneous data sources however data translation is still remains a challenge at large.

On internet there is a huge data explosion going, According to Eric Schmidt, Google CEO "Every two days now we create as much information as we did from the dawn of civilization up until 2003, something like five

Exabyte of data" he says [http://techcrunch.com, 2012]. In 2011 300 million website were added making total number of websites to 555 million(December 2011)[http://royal.pingdom.com, 2012], thus resulting numerous data sources each having its own structure and schema, user desired data presentation still remains issue at large and needs to understood and covered at the earliest.

3.2 Data & Information

Data refers to the lowest abstract or a raw input which when processed or arranged makes meaningful information. It is the group or chunks which represent quantitative and qualitative attributes pertaining to variables. Information is usually the processed outcome of data. More specifically speaking, it is derived from data. Information is a concept and can be used in many domains.

Data can be in the form of numbers, characters, symbols, or even pictures. A collection of these data which conveys some meaningful idea is information. It may provide answers to questions like who, which, when, why, what, and how.

The raw input is data and it has no significance when it exists in that form. When data is collated or organized into something meaningful, it gains significance. This meaningful organization is information [http://differencebetween.net, 2012].

3.3 File Formats & Database

Some file formats are designed for very particular types of data: PNG files, for example, store bit mapped images using loss less data compression. Other file formats, however, are designed for storage of several different types of data: the Ogg format can act as a container for many different types of multimedia, including any combination of audio and/or video, with or without text (such as subtitles), and metadata. A text file can contain any stream of characters, encoded in one of many kinds of character encoding schemes, including possible control characters. Some file formats, such as HTML, Scalable Vector Graphics, and the source code of computer software are also text files with defined syntaxes that allow them to be used for specific purposes[http://en.wikipedia.org, 2012][Md. Sumon Shahriar, 2010].

On the other hand a database is a collection of data that is organized so that it can easily be accessed, managed, and updated. In computing, databases are sometimes classified according to their organizational approach. The most prevalent approach is the relational database, a tabular database in which data is defined so that it can be reorganized and accessed in a number of different ways. A distributed database is one that can be dispersed or replicated among different points in a network. An object-oriented programming database is one that is congruent with the data defined in object classes and subclasses [http://searchsqlserver.com, 2012][Marc Van Cappellen, 2008].

3.4 Problems Pertaining to New Systems after Migrating from legacy Sources

The various problems pertaining to new systems after migration from legacy sources are discussed below:

3.4.1 Problem Pertaining to Recovery

Before we go into the details of why we need recovery controls, we would first like to clarify the meaning of recovery. In Date's words [Date, 1995], recovery is depicted as "Recovery in database system means, primarily, recovering the database itself-that is, restoring the database to a state that is known to be correct after some failure has rendered the current state incorrect, or at least suspicious." There are several possible reasons for a transaction to fail in the middle of the execution. For example, system crash may cause error in the computer system during transaction execution. Some transactions might violate the concurrency control enforcement and the control may decide to abort the transaction because it violates serializability or because several transactions are in a state of deadlock.

Physical problems (media failure) may happen such as head crash on the disk, fire or sabotage or physical theft. For a system crash, the content in the buffer memory is a critical point. The state of any transaction in a progress is not known; such a transaction did not successfully complete, and so must be undone (rolled back) when the system restarts. For example, while some users are updating the record a power failure occurs. Hence those unfinished transactions must be rolled back to its previous state. For data protection when media failure occurs, the backup copy of registration database is needed for restoration; there is no need for a roll back.

3.4.2 Problems Pertaining Concurrency

Enterprise Information System's Data Warehouse is a shared resource. It is assumed at a particular instance of time many concurrent users will be accessing the database for doing read and write operations on data. With concurrent processing involving updating the data in parallel, a database without concurrency control will be compromised due to interference between users. Concurrency control allows many users to access and update the database simultaneously while preventing partially completed updates from happening [Date, 1995][E.Bertino, 2001]. This technique is essential to our Enterprise Information System, such as when more than one user is registering a customer in the same groups or service with a limited number of available spaces for customers. When the user decides to register a customer for the service, the database integrity parameters will check first whether or not the Service Group is full by having one variable used to store the number of Customers per group who have registered for that service. While in a multiuser structure it can lead to a violation of data Integrity. If the problem of violation takes place then it is not only last updates that concurrency control mechanism has to address, uncommitted dependency and inconsistent analysis problems are also possible. These problems can cause the database to be in inconsistency state [J. Gray, 1993].

3.4.3 Problems Pertaining Security

The unauthorized access can modify the sensitive data pertaining to both employees and customers so emphasis has to be given to how proper roles and privileges are given to the users so that the security of the data is not violated [K.P. Eswaran, 1976]. It can be concluded that Security concerns are implemented so as to ensure that users can do only what they are allowed to do and nothing more. In the EIS data, all information concerned with every customer and employee is kept in it. If no security system is implemented into the system, tremendous problems will arise since a user might alter the sensitive data of the enterprise. Hence a security system is needed for maintaining a usable database for EIS[R. Bayer, 1977]. Some constraints must be enforced to ensure that authorized users are doing correct operations, satisfying all constraints.

3.5 Proposed Solutions

By using a central database for Enterprise Information Systems, many problems are likely to occur due to unaware or deliberate action. All the errors will be adversely directed to the central database Log. With the great importance that the central database/Warehouse is used for most enterprises for performing various tasks, a small Parallel Warehouse (TDW) is implemented in order to be used as a substitute during the transaction period of the Enterprise. So any adverse effect will only be confined to the small enterprise CDW [T. Johnson, 1993][J. Gray, 1978]. Consequently, advantages concerning security are also provided. Moreover, by separating warehouse, the enterprise can modify or change the structure of the database easily, without too much concern about its side effect on the CDW.

The TDW can be generated easily from the CDW by transferring only the necessary information and Structures. Some information is transformed into more appropriate form just like emp, dept, finance, log schemas and relationships, etc. This information will be transformed into the form of which can be temporarily used in the system. Any unauthorized users who try to change these details will fail since no as the changes are not permanent for the CDW. It is also more convenient for our enterprise database since the transformed transaction is more relevant to the objective of performing transaction. The transactions are consolidated in this TDW and when the transaction operations are completed the data is reflected in the CDW. During every export a log database is maintained keeping track of new records added and the tables which are being modified. So every time the export takes place only those data tables are modified which have an update associated in the transaction and the other data table are kept un-altered.

3.5.1 Proposed Solution for Recovery

Since we are using separate databases any failure that might cause the database to corrupt is now limited only to the TDW, leaving the CDW untouched. For the case that TDW is collapsed, it can easily be reconstructed within three minutes, since it contains only information Transaction details, it is then ready to be used again to provide uninterrupted service during the transaction period. Furthermore, for any terminated transaction. The transaction manager is used to provide the atomicity of important transactions. In other words, it guarantees that if the transaction executes some updates and a failure occurs (whatever the cause) before the transaction finishes (reach its plan), then those updates will be undone. Thus, the transaction either executes in its entirety or is totally canceled. In this way a sequence of operations that is fundamentally non-atomic can be viewed as if it were atomic from this point of view. The commit transaction and rollback transaction are the key to the way recovery works. For commit transaction, it tells the Database Management System (DBMS) that the atomicity of the process has been thoroughly finished. The database is in a consistent state and the updates made by the process can now be made that is fundamentally non-atomic can be viewed as if it were atomic from this point of view.

The commit transaction and rollback transaction are the key to the way recovery works [H.V. Jagadish, 1990][J. Hellerstein, 1995]. For commit transaction, it tells the Database Management System (DBMS) that the atomicity of the process has been thoroughly finished. The database is in a consistent state and the updates made by the process can now be made permanent or committed. In contrast, the signal of failure to end the transaction is indicated by the rollback transaction. The database might be in an inconsistent state and the updates by that transaction must be undone or rollback. A log will be maintained by the system about the details of all update operations. So, if it is necessary to undo any specific update, a log file will be used to update value to its previous value. For EIS, the technique of commit and rollback transaction is implemented in Microsoft Visual Basic .NET for SQL Server 2005 by using the reserved words BeginTrans, CommitTrans and Rollback. These three functions are used for important transactions that might be able to compromise the consistency of the database. For example, by using these functions, when user decides to register a customer, the program adds the customer to the EIS record and updates the number of Customers in Service Group with a fallback recovery against any failure. A clearer idea of the importance of recovery might be depicted when the transaction is concerned with payment. Therefore transaction should be made as atomicity in order to avoid the inconsistency of the database

3.5.2 Proposed Solution for Concurrency

As mentioned before, without concurrency control the problems of lost updates, uncommitted dependency, and inconsistent analysis are expected to occur. Since the CDW is a shared and classified resource, careful database management must be incorporated. Microsoft SQL Server 2005 provides three levels of locking. These are record locking, table and recordset locking, and opening with exclusive access. For record locking, only the record currently being edited is locked. For table and recordset locking, an entire table or all tables underlying a form are locked while any user is editing any record in the form. Finally, for opening with exclusive access, the entire database is locked by a single user-change into single-user environment. Microsoft SQL Server 2005 automatically locks the record currently being edited even though the programmer did not predefine the lock mechanism [R. Hagmann, 1987][R.W. Hamming, 1950]. In TDW, since we always operate under multiuser environment, the opening with exclusive access is irrelevant and will not be mentioned twice. The most used mechanism in our CDW is table and recordset locking. Since this database is a relational database and has a quite complicated relation and query, recordset will be used for most of the time. As mentioned before about concurrency, the inconsistency about counting the number of customers class can be solved by using table and recordset locking which will be referenced as lock. As the Customers decide to register any Group, the lock is made active so that counting the number of Customers in each Group is correct. The reason why the original system locks the entire table is that it uses one table to store all records of who had registered for which courses. Consequently, the counting method also uses this table to count the number of customers for each class. In other words, from the CDW uses the table "REGEMP" to store and count the number of customers in each course by using Customer ID and Group ID as criteria for counting the number of Customers registered in each group. In our new system, the TDW stores records in table "TCusTRegradable" which is then used to count the number of customers in each course, DCount is used in Visual Basic .NET code. Hence, if these tables are locked while any user is currently updating then, the lost updates problem can be solved.

3.5.3 Proposed Solution for Security

In the past, the original system used table TUSERS to store login name and password for the super users. As the user log on to the database, the application only verifies login name and password using ordinary Visual Basic .NET code to check the information in TUSERS. UserLookUp is used to check the data in the table. Hence, if anyone is able to look into this table, he can find the login names and passwords easily. If the Transaction database is not encrypted, anyone with a disk editor can view the contents of the file. Although the data within the file will not appear in an easy-to-read format, the data is there and available for unauthorized individuals to see. Therefore, the encryption is used for the Transaction database even though the performance of the application will drop but it is necessary to keep it encrypted. In other words, another level of security is made from encrypting the database. Typically, the DBMS supports either or both of two broad approaches to data security [W. Hsu, 2004][A. Kashyap]. The approaches are known as discretionary and mandatory control. In the case of discretionary control, a given user will get different authority or privilege. Discretionary schemes are very flexible. In contrast to discretionary, mandatory control defines each data object to be labeled with a certain classification level, and each user is given a specific level of clearance. A given data object can then be accessed only by users with the appropriate clearance. Consequently, mandatory control is rigid but appropriate to use with a EIS database [A. Kashyap]. It is easier and clearer to maintain a class of users than to concentrate on individual users, since every customer must have the same right. Login name and password table will not be used; TUSER and new approach should be used here. Microsoft SQL Server 2005 provides a very powerful and comprehensive feature to maintain user account. The information on each clearance level of user and password, and access right for each object will be stored in the system file. This file is distinct from the database file which Microsoft SQL Server uses to store information database security. The privileges of each level will be discussed in the next section.

3.5.4 Solution for Integrity

As mentioned before, integrity concerns protection against authorized users. For customers they must be assigned the rights to read all that data and update only the table TREGISTER since this table is used to store who is registering which services. Customer must not be able to view database application as in the design view to avoid any adverse alteration by them. The other thing of concern is that customer should not be allowed to use the toolbar since it provides features beyond the necessity of customer registration. For the super class of users, it depends on the policy of the Enterprise to what the registrar is allowed to do as does the higher Admin of users [LISA, 2004]. The other policies such as how many credits customers must enroll in each group, the maximum number of Services a Customer can register each Group, prerequisite, co-requisite, etc, are deliberately ignored from user privileges levels since it varies from organization to organization.

3.6 Information Translation Problem Definition

Most of the internet and intranet users are not well versed with technology. It has been observed that even top level managers are dependent on technical support of the organization for carrying out there day to day tasks.

Data in the organization may be present in database however user wants the same data as hardcopy, or as in most cases written text in the website is copied and pasted on Microsoft word. User wants part of the image but does not understand if it is possible to edit the picture or not.

The problem is that there is not a single generic data format available, information is present in different formats requiring different tools for its mining.

In prevailing circumstances user is required to have comprehensive knowledge of system/database/file formats in order to use information the way he/she wants to. The target system needs to be built which hides the technology from users and provides him with the information in desired format.

3.6.1 System Assumption & Solution

Heterogeneous data spread across multiple sources having varying underlying structure and data format which are extracted, transformed and loaded into single Data Warehouse. We assume Warehouses/Marts depending on enterprise architecture are created as such data is centralized [Mohammad Ghulam Ali, 2009][Stefan Biffl, 2010].

Solution is conversion of result in user desired format i.e user query is executed on warehouses and generated result is converted into user desired format, (Excel/odt/pdf etc.).User can also describe feature of his/her file format i.e. he/she wants Vardana 12 as font size in word 2007 format.

3.6.2 Proposed Information Translation Algorithm

- **ISL- INTELLIGENT SOFTWARE LAYER** is placed between user and warehouses, user input is received by and converted into query by ISL and same is executed on warehouse, auxiliary information is saved for later use e.g font type size format, thus user input is received by ISL.

- User is provided with GUI so that he/she can input his/her query(google sought) along with desired format in which user wants his/her result e.g(.Microsoft word. determine extension of the said format[S. Agarwal,2002][F. Song,1999][Bhalotia, 2002].

- Result generated as a result of execution of query is not passed on to user but is received by ISL for converting it into user desired format.

- ISL receives result from warehouse, creates new text file and saves result in this newly created text file(.txt), file name is based on time stamping principle e.g 1545220412.txt where 15 is hours, 45 is mins, 22 is day 04 is month and 12 is year.

- ISL converts it into user desired file format, translation requires

  a. User desired format is already know.

  b. Create new file, with the same name as that of text file but with user desired extention i.e if use wants output in word format then 1545220412.docx is created.

  c. User requirement such as font, size,margins etc is taken care of at the time of file creation, e.g word file is created in which font size, margins etc become integral part of this newly created file.

  d. Data saved in text files is read char by char and appended into the newly created file.

- File created is passed on to the user, and both text file and application file are deleted, as such disk storage is not a issue.

- ISL does not need to buy application license such as Microsoft office, pdf, etc.

- ISL can be initially tested for few formats e.g word, pdf etc before support for all formats can be extended.
- Diagrammatic representation of algorithm is shown below(fig 3.1)

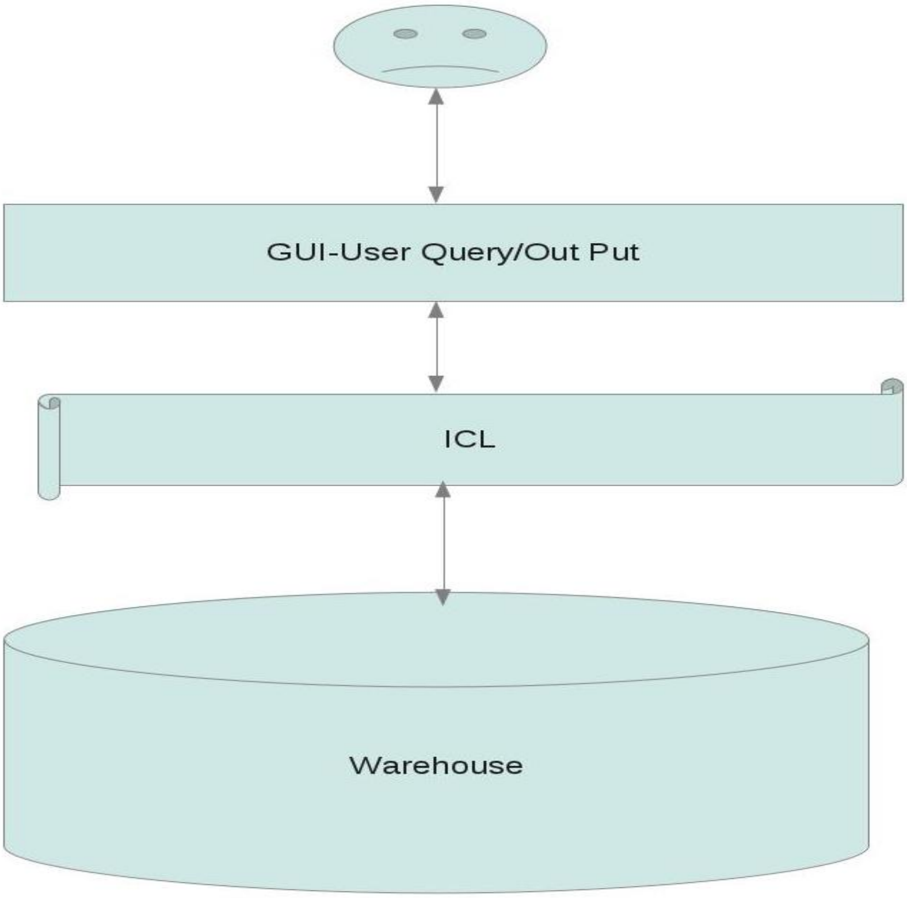


Fig 3.1: Diagrammatic Representation of Algorithm

3.7 Conclusion

User over the years has become more demanding; he/she does not only need information but wants it in specific format which should be complete and correct. Globally centralization was prioritized because of collection of massive data in heterogeneous data sources, however much was not thought for naive user and information system was still at the mercy of technocrats time has now come to stress more upon user demands so as to meet user demands and make user dependency on technocrats minimal.

So far, all fundamental problems have been cleared up. Recovery, Concurrency and Security of the Enterprise Information System Data Warehouse have been solved. However, a lot of delicate improvements can be implemented because the further use of this CDW in the future will tell what is appropriate and what should be improved. User-interface is another area for improvement since it can reduce the error that users might make but requires further work to produce an appropriate and elegant design that suits the user’s needs.

4.1 Introduction

Once the data has become machine readable in an Information System many issues related to its security arise. Although the information systems provide users greater access than ever to vast information resources, however they are equally subject to threats that jeopardize the privacy and confidentiality of sensitive information, the integrity of data, and the availability of critical information system resources [Battacharya et al, 2007]. Protecting information and the resources that process and maintain information is critical to the continuity of operations [Ingrid et al, 2007]. Security of information resources must include controls and safeguards to offset possible threats as well as controls to ensure timeliness, availability, integrity, confidentiality, etc. Information technology (IT) security encompasses the total infrastructure for maintenance and delivery of information, including physical computer hardware, supporting equipment, communication systems, and logical processes defined by software, procedures etc.

Many of the basic requirements for security are well-known, and apply equally to a document as to any other system: The system must prevent unauthorized users from accessing the system and similarly authorized users from modifying data un-authorizedly. The applications and underlying data must not be susceptible to data-theft by hackers, the data must be available to the right users at the right time, and the system must keep a log of activities performed by its authorized users [MEHARI, 2004].

When a paper document is scanned it will usually end up in one of the following formats:

1. Image file format

   Paper document is scanned and is stored in the picture format, it is in read only format however picture in windows platform can be edited in paint application. The security of such document lies in the hands of System Administrator who has to ensure access to the system and document. Many image file formats are available with each having its own merits as well as de-merits.

2. Text File format

   Paper document is scanned and data from it is extracted and stored in the text format (text files). Such files can easily be manipulated and therefore the security of such document again lies in the hands of System Administrator who has to ensure access to the system and document. The word processors as well as several other utilities mostly store the data in the text format and here also several file formats are available. .

3. Binary file format

Paper document is scanned and data from it is retrieved and stored in the database file. The databases have built in security features however the Database Administrator again has to ensure the security issues. The database administrator has to ensure secured access to the database in the form of Database privileges and roles that ensures that a user can only perform an operation on a database object if he has been authorized to perform that operation. A very large number of databases are available each having its own file format.

Whatever may be the format of the document, the following are the possible states which it can have during its life cycle (fig 6.1):

a) Processing

b) Storage

c) Transmission

At every state the threats as well as the security procedures vary. The threats can be both intentional and un-intentional. They can also be internal as well as external or both.

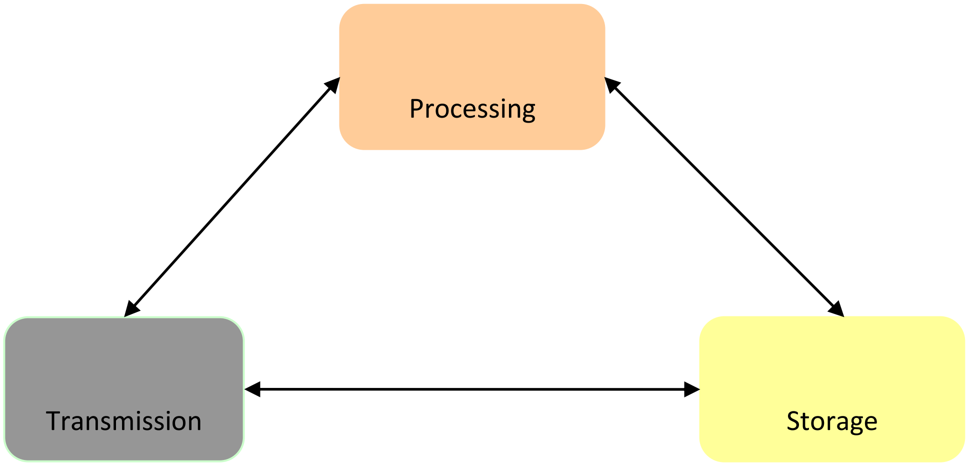


Figure 4.1: Document Stages

In this chapter we start with the introduction to the Database Access System based on privileges and roles, we then elaborate on the existing system and understand the need of encryption. Document security is proposed using encryption and University Registration System is taken as case study in understanding the system. For the access control of the system, the built in Database Control features are used however the user operations are monitored and controlled by introducing the process of encryption. In this chapter we propose a model for system security involving the available database security features and the encryption technique. The model is designed specifically for preventing un-authorized modification of the data by its users which enjoy different levels of authorizations. In the rest of the paper document and data would be used inter-changingly. In our research since the data extracted from the pre-printed forms is stored in the databases therefore we begin with the security systems of the databases.

4.2 Access to Database Systems

Any security solution must meet the following criteria: secrecy or confidentiality, integrity, and availability (Bertino, E) which are together called as CIA triad (fig 4.2). Secrecy and confidentiality may be the most recognized of the criteria. They are both items that end users are well aware of. For instance, a customer stores their credit card information within their account at an online retailer. If the database is compromised, either because of poor database design or poor application design, the customer information is made available. This example compromises the secrecy and confidentiality criterion. The next criterion is integrity. Keishi Tajima of Kyoto University states that "User access to a database is either an action to get some information from the database, or an action to give some information to the database in order to make it reflected by the database state."(Tajima, Keishi) With this in mind we see that database users share a great deal of responsibility. Data integrity ensures that data is being modified by an authorized user and that it is being modified properly. It is possible for a user to be given improper rights to database objects. This means that a user could modify a table that they should not have access to. It is very important for Database Administrators (DBA's) to constantly monitor user security. Imagine an employee changing their salary information.

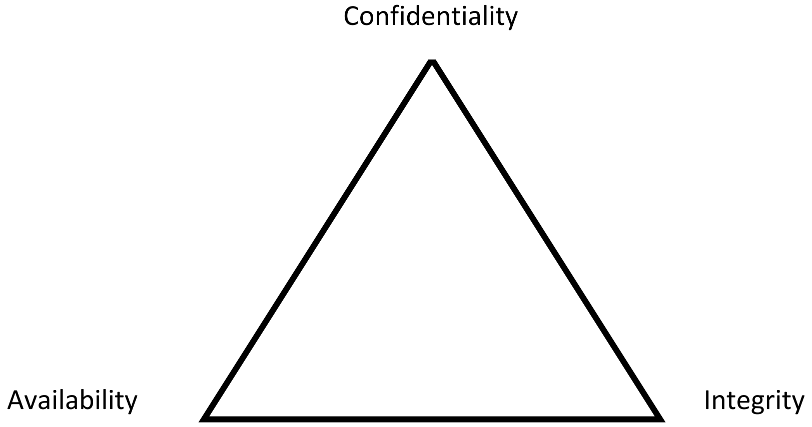


Figure 4.2: Objectives of Security

The last criterion that a database security solution should meet is availability. Availability is ensuring that the database is available even through hardware and software problems, and malicious attacks. Although databases can handle 1000's of concurrent connections simultaneously, it is possible to cause a denial of service attack against the database. For example, an Oracle database uses a listener that runs on port 1521 by default. The listener listens for connections to the database and hands the connections off to the database. By flooding the listener with requests, it is possible to fill the listener's log file and cause the listener to stop accepting connections. As you can see from these examples it is important to have database security solutions that meet the three criteria.

We know that a database is of no value if users cannot access the data contained in it. There are many tools and clients on the market that allow a user to access data within a database. The client software includes a tool called SQL*Plus. SQL*Plus is a command line tool that allows users to execute Structured Query Language (SQL). SQL*Plus connects to a database listener running on the database server. Once the client connects, the listener starts a dedicated process on the server and passes the connection to the new process.

Most of the databases consider security as system security and data security. System security asks the questions, "Has the user provided a valid username and password?" and "What actions can the user perform on the database?" Data security asks the questions: "What objects does the user have access to?" and "Are the user's actions being audited?" Oracle uses discretionary access control, meaning access to information is based on privileges. For example, you have full access to the objects that you create in your schema. You also have the ability to grant permissions to objects in your schema to other users.

As mentioned earlier, Database permissions are based on privileges. Users access the database using their username and password [Whitman & Mattord, 2007]. Once connected, users are able to create objects in their own schema only if they have a quota on their default table space. It is important to realize that users can only create objects in table spaces that they have quotas. An example of a user misusing quotas would be if he created a table in the table space user data and then filled the table space by adding too many rows to the table. This may cause other users who had objects in the table space user data to not be able to insert or update their data. It is important to examine user quotas periodically. There are many types of privileges associated with an Oracle database. This chapter will only examine very basic privileges like select, insert, update, and delete.

Auditing allows DBA's to monitor misuse and abuse of privileges. There are several types of auditing. The first is statement Auditing.

Statement Auditing audits the statements that specified users are executing against any schema. Fine-grained auditing audits access to objects based on their content. Schema object auditing allows DBA's to audit statements that are being executed against a specific object. Schema object auditing applies to all database users. Privilege auditing audits system privileges. Any number of users can be specified to be monitored with Privilege auditing. Auditing can be configured to store the audit information within the database itself or at the Operating System level in files. It is important to know that with auditing enabled, tremendous amounts of information will be logged. This information will need to be cleaned up periodically. As mentioned previously, Oracle stores data unencrypted by default. Depending on the version of the Oracle database it may be possible to use the DBMS_CRYPTO, DBMS_OBFUSCATION_TOOLKIT, or Transparent Data Encryption to encrypt data. Regardless of the encryption method, the data cannot be unencrypted without a key.

The databases as said earlier have a well defined system of security. Upon creating a database user and granting him rights to access the database, the administrator of the Database must ensure secured and control access to data. The access of a user to the database records can be limited to the level of individual records based on the identity and privilege of that user. However, maintaining this type of complex access control code is not only costly, but also risk-prone. The access control is built into an application however a user has access to the database itself, which is common in database environments. For these principal reasons, we understand this hard-to-solve problem to build security as a whole, and access control in particular, onto the data itself, inside the database.

Database privileges and roles ensure that a user can only perform an operation on a database object if he has been authorized to perform that operation. A privilege is an authorization to perform a particular operation and without explicitly granted privileges, a user cannot access any information in the database. System privileges authorize a user to perform a specific operation, such as the CREATE TABLE privilege, which allows a user to create a database table. Object Privileges authorize a user to perform a specific operation on a particular object. An example of object privileges is SELECT ON STUDENT, to allow a particular user to query to the table, but not query other database objects, nor modify any of them. Granting a user SELECT, UPDATE, INSERT allow a user to read and write to this view. By providing these types of privileges, the database system facilitates to ensure that the database users are only authorized to perform those specific operations required by their job functions. In addition, other features, such as roles and stored procedures, not only allow you to control which privileges a user has, but under what conditions he can use those privileges.

While privileges let you restrict the types of operations a user can perform, managing these privileges may be complex. To address the complexity of privilege management, database roles encapsulate one or more privileges that can be granted to and revoked from users.

4.3 Encryption & Decryption

The process of disguising a message in such a ways as to hide its substance is encryption. An encrypted data is cipher-text. The process of turning cipher-text back into plaintext/data is decryption (figure 4.3).

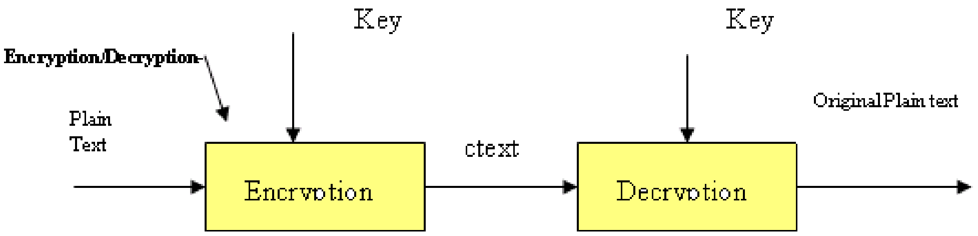


Figure 4.3: Encryption Procedure

Encryption is used to secure communications and data storage, particularly authentication credentials and the transmission of sensitive information. It can be used throughout a technological environment, including the operating systems, middleware, applications, file systems, and communications protocols. Encryption can be used as a preventive control, a detective control, or both. As a prevention control, encryption acts to protect data from disclosure to unauthorized parties. As a detective control, encryption is used to allow discovery of unauthorized changes to data and to assign responsibility for data among authorized parties. When prevention and detection are joined, encryption is a key control in ensuring confidentiality, data integrity, and accountability.

Properly used, encryption can strengthen the security of an institution’s systems. Encryption also has the potential, however, to weaken other security aspects. For instance, encrypted data drastically lessens the effectiveness of any security mechanism that relies on inspections of the data, such as anti-virus scanning and intrusion detection systems. When encrypted communications are used, networks may have to be reconfigured to allow for adequate detection of malicious code and system intrusions.

Although necessary, encryption carries the risk of making data unavailable should anything go wrong with data handling, key management, or the actual encryption. For example, a loss of encryption keys or other failures in the encryption process can deny the institution access to the encrypted data. The products used and administrative controls should contain robust and effective controls to ensure reliability.

Financial institutions employ encryption strength sufficient to protect information from disclosure until such time as the information's disclosure poses no material threat. For instance, authenticators should be encrypted at strength sufficient to allow the institution time to detect and react to an authenticator theft before the attacker can decrypt the stolen authenticators. Decisions regarding what data to encrypt and at what points to encrypt the data are typically based on the risk of disclosure and the costs and risks of encryption. The costs include potentially significant overhead costs on hosts and networks. Sensitive information is also encrypted when passing over a public network and also may be encrypted within the institution.

Encryption cannot guarantee data security. Even if encryption is properly implemented, for example, a security breach at one of the endpoints of the communication can be used to steal the data or allow an intruder to masquerade as a legitimate system user.

4.4 How Encryption Works

In general, encryption functions by taking data and a variable, called a "key" and processing those items through a fixed algorithm to create the encrypted text. The strength of the encrypted text is determined by the entropy, or degree of uncertainty, in the key and the algorithm. Key length and key selection criteria are important determinants of entropy. Greater key lengths generally indicate more possible keys. More important than key length, however, is the potential limitation of possible keys posed by the key selection criteria. For instance, a 128-bit key has much less than 128 bits of entropy if it is selected from only certain letters or numbers. The full 128 bits of entropy will only be realized if the key is randomly selected across the entire 128-bit range.

The encryption algorithm is also important. Creating a mathematical algorithm that does not limit the entropy of the key and testing the algorithm to ensure its integrity are difficult. Since the strength of an algorithm is related to its ability to maximize entropy instead of its secrecy, algorithms are generally made public and subject to peer review. The more the algorithm is tested by knowledgeable worldwide experts, the more the algorithm can be trusted to perform as expected. Examples of public algorithms are AES, DES and Triple DES, HSA-1, and RSA.

4.4.1 Encryption Key Management

Since security is primarily based on the encryption keys, effective key management is crucial. Effective key management systems are based on an agreed set of standards, procedures, and secure methods that address.

- Generating keys for different cryptographic systems and different applications;
- Generating and obtaining public keys;
- Distributing keys to intended users, including how keys should be activated when received;
- Storing keys, including how authorized users obtain access to keys;

- Changing or updating keys, including rules on when keys should be changed and how this will be done;
- Dealing with compromised keys;
- Revoking keys and specifying how keys should be withdrawn or deactivated;
- Recovering keys that are lost or corrupted as part of business continuity management;
- Archiving keys & Destroying keys;
- Logging the auditing of key management-related activities; and
- Instituting defined activation and deactivation dates, limiting the usage period of keys.

Secure key management systems are characterized by the following precautions:

- Key management is fully automated (e.g., personnel do not have the opportunity to expose a key or influence the key creation).
- No key ever appears unencrypted.
- Keys are randomly chosen from the entire key space, preferably by hardware.
- Key-encrypting keys are separate from data keys. No data ever appears in clear text that was encrypted using a key-encrypting key. (A key encrypting key is used to encrypt other keys, securing them from disclosure.)
- All patterns in clear text are disguised before encrypting.

- Keys with a long life are sparsely used. The more a key is used, the greater the opportunity for an attacker to discover the key.

- Keys are changed frequently. The cost of changing keys rises linearly while the cost of attacking the keys rises exponentially. Therefore, all other factors being equal, changing keys increases the effective key length of an algorithm.

- Keys that are transmitted are sent securely to well-authenticated parties.

- Key-generating equipment is physically and logically secure from construction through receipt, installation, operation, and removal from service.

4.4.2 Encryption Types

Three types of encryption exist: the cryptographic hash, symmetric encryption, and asymmetric encryption. A cryptographic hash reduces a variable-length input to a fixed-length output. The fixed length output is a unique cryptographic representation of the input. Hashes are used to verify file and message integrity. For instance, if hashes are obtained from key operating system binaries when the system is first installed, the hashes can be compared to subsequently obtained hashes to determine if any binaries were changed. Hashes are also used to protect passwords from disclosure. A hash, by definition, is a one-way encryption. An attacker who obtains the password cannot run the hash through an algorithm to decrypt the password. However, the attacker can perform a dictionary attack, feeding all possible password combinations through the algorithm and look for matching hashes, thereby deducing the password. To protect against that attack, "salt," or additional bits, are added to the password before encryption. The addition of the bits means the attacker must increase the dictionary to include all possible additional bits, thereby increasing the difficulty of the attack.

Symmetric encryption is the use of the same key and algorithm by the creator and reader of a file or message. The creator uses the key and algorithm to encrypt, and the reader uses both to decrypt. Symmetric encryption relies on the secrecy of the key. If the key is captured by an attacker, either when it is exchanged between the communicating parties, or while one of the parties uses or stores the key, the attacker can use the key and the algorithm to decrypt messages or to masquerade as a message creator.

Asymmetric encryption lessens the risk of key exposure by using two mathematically related keys, the private key and the public key. When one key is used to encrypt, only the other key can decrypt. Therefore, only one key (the private key) must be kept secret. The key that is exchanged (the public key) poses no risk if it becomes known. For instance, if individual A has a private key and publishes the public key, individual B can obtain the public key, encrypt a message to individual A, and send it. As long as individual A keeps his private key secure from discovery, only individual A will be able to decrypt the message.

4.5 Protection with Encryption

Encryption can be used as a substitute for effective access control. An additional measure of security can be introduced by selectively encrypting sensitive information, such as credit card numbers, marks before it is stored in the database. It’s the application package that encrypts and decrypts stored data and encrypting stored data can provide the assurance of securing data regardless of access method. However, encrypting data inside a data base can be complex, and it usually adds overhead to the system. Additionally, the encryption keys must be stored somewhere in an application, in a file, or in a table—and managing these keys is widely recognized as a very difficult security issue.

4.6 Risk Calculation

On their own, vulnerabilities and threats don't pose a security risk. However, when a threat combines with a vulnerability, you have a situation known as a risk— something that should be corrected. Security experts often describe risk using this equation:

Risk = Threat × vulnerability

There's a good reason that the multiplication operator (×) is used to describe the relationship between threats and vulnerabilities. A risk is high only when both the threat and vulnerability are high. This concept is illustrated in the risk matrix shown in Fig 4.4.

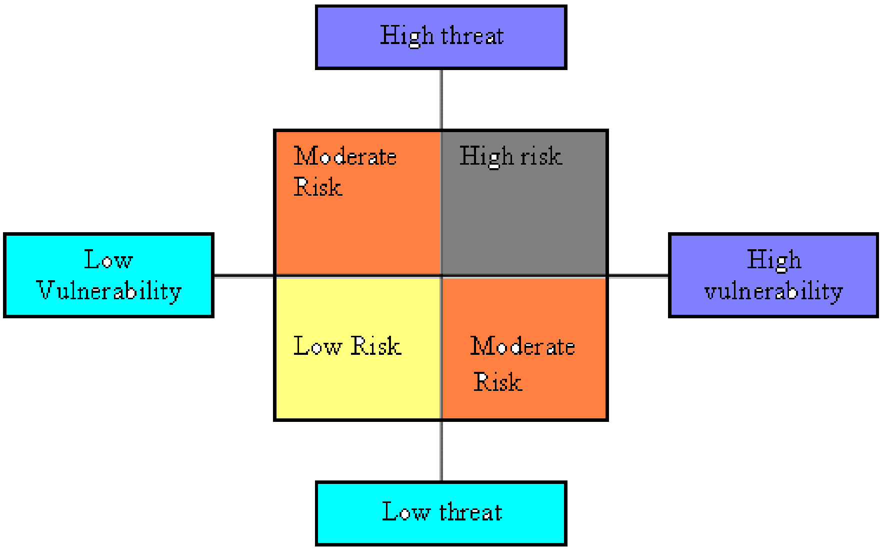


Figure 4.4: Risk Matrix

4.7 Proposed System

In the existing system shown in (fig 4.5), the user logs on to the system with the access to the Data (as per the privileges) and the existing database security system can not protect data any more. So what is required from Hacker/Intruder point of view is to hack into the system, so that data is at disposal and manipulation without restriction is possible because the main feature of the access based system is that of privileges and roles associated with the user. Therefore within the given privileges one can manipulate/hack or simply steal the information with out fear of being identified.

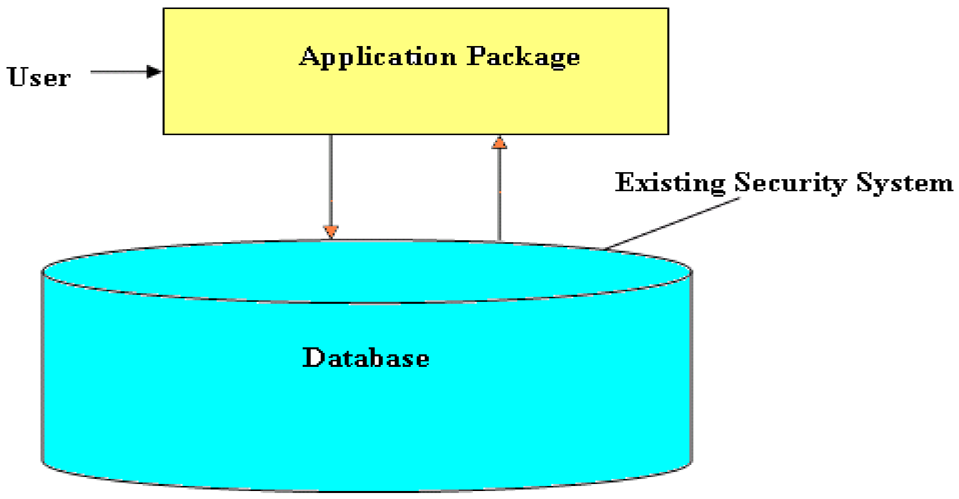


Figure 4.5: Existing System Security

What is required is to impart security to sensitive data even when the Intruder/Hacker has successfully intruded into the system, we extend the existing system by inserting layer of software between application package and Database which will be responsible for encryption of data at the time of data being uploaded into database and decryption of data at the time of data retrieval (fig 4.6). If the decryption keys are not provided at the time of retrieval an intrusion into the system would be reported.

The Encryption key as shown in (fig 4.3), need not to be provided by the user as it will be pre-installed/updated by the Database administrator without the knowledge of database users. However at the time of data retrieval of decrypted data user has to provide Decryption key/otherwise the system can send red alert.

The additional security features can be implemented using Encryption protocols like

1. DES
2. Blowfish
3. Triple DES

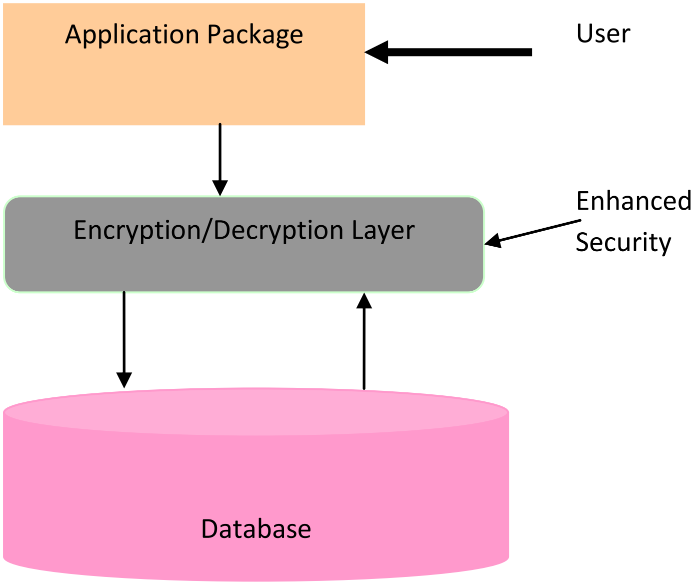


Figure 4.6: Security of the Proposed System

4.8 Proposed Implementation Algorithm for Encryption/Decryption

Implementation part of the system shown in (fig 4.7) pertains to the layer of software installed between the Application Package and Database System. The interfaces between the Application Package and Encryption/Decryption module and that between the Encryption/Decryption module and Database System is invisible to all the three layers namely Application Package, Encryption/Decryption Module and the Database System.

As shown in the figure 4.7, the overall Security Package has following three layers

1. Interface and tools for Application Package(**IT-AP**)
2. Encryption and Decryption software(**EDS)**
3. Interface and tools for Database(**IT-DW**)

4.9 Interface and tools for the Application Package

The primary role of this layer is to interact with the Application Package as if it is interacting with the Database, analyse queries and subsequently route the query/data via either Encryption/Decryption or to Interface and tools for Data Warehouse Manipulation.

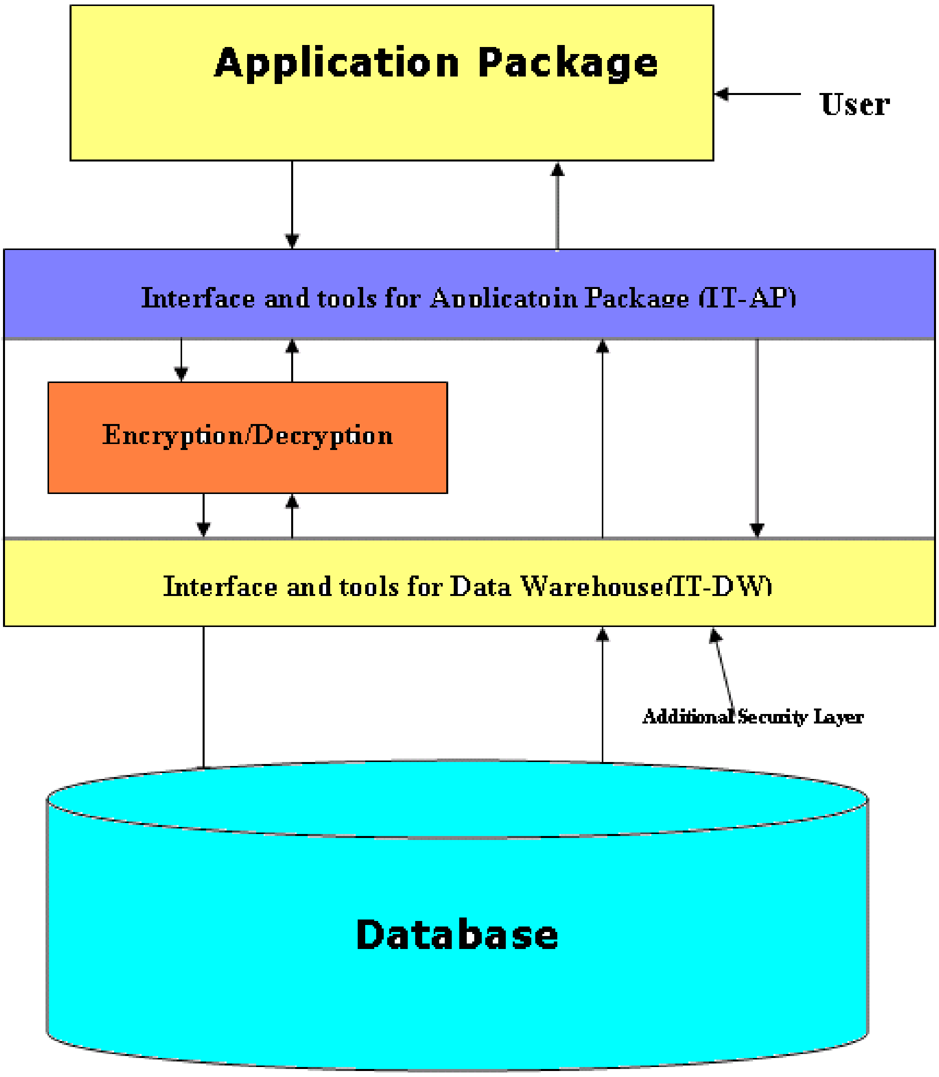

Application Package
User
Interface and tools for Applicatoin Package (IT-AP)
Encryption/Decryption
Interface and tools for Data Warehouse(IT-DW)
Additional Security Layer
Database

Figure 4.7: Security Implementation Procedure

IT-AP sends the query to the IT-DW while the data in the query is either send to IT-DW directly or via Encryption/Decryption. As is shown in the figure 6.7, the data pertaining to the encrypted fields are send to IT-DW via Encryption/Decryption and the data pertaining to unencrypted fields is sent to IT-DW directly.

4.9.1 Encryption/Decryption Layer

The sole purpose of this layer is to encrypt the data received from IT-AP and decrypt the data received from IT-DW. Any algorithm can be used for this purpose however the algorithm used should follow the norms of National Bureau of Standards (NBS); and one such norm which has to be strictly applied here is "the security of algorithm must reside in the key; the security should not depend on the secrecy of the algorithm."

Data Encryption Standard (DES) is Symmetric algorithm, the same algorithm and key are used for encryption and Decryption. The Key length is 56 bits, which can be changed any time.

4.9.2 Interface and tools for Database (IT-DW)

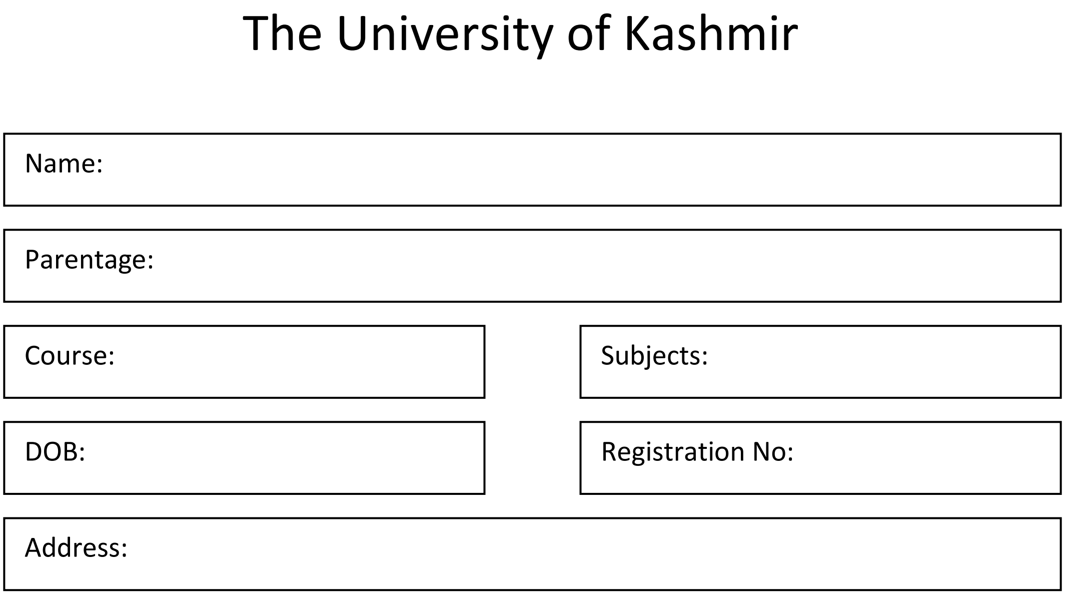
The primary purpose of this Layer is to interact with the database System as if Application Package is interacting with the Database System, send quires for execution and subsequently receive data/result from Database and send the same to either IT-AP directly or via Encryption/Decryption.

4.10 Case Study

Shown below is structure of one of the table used in the Registration database of the University of Kashmir. These data fields are extracted from the scanned pre-printed form which is filled up by the student at the time of admission.

Date of Birth	Registration Number	Name	F name	Course	Subject	Session

The structure of such a form may be like the one shown below



The University of Kashmir

Name:

Parentage:

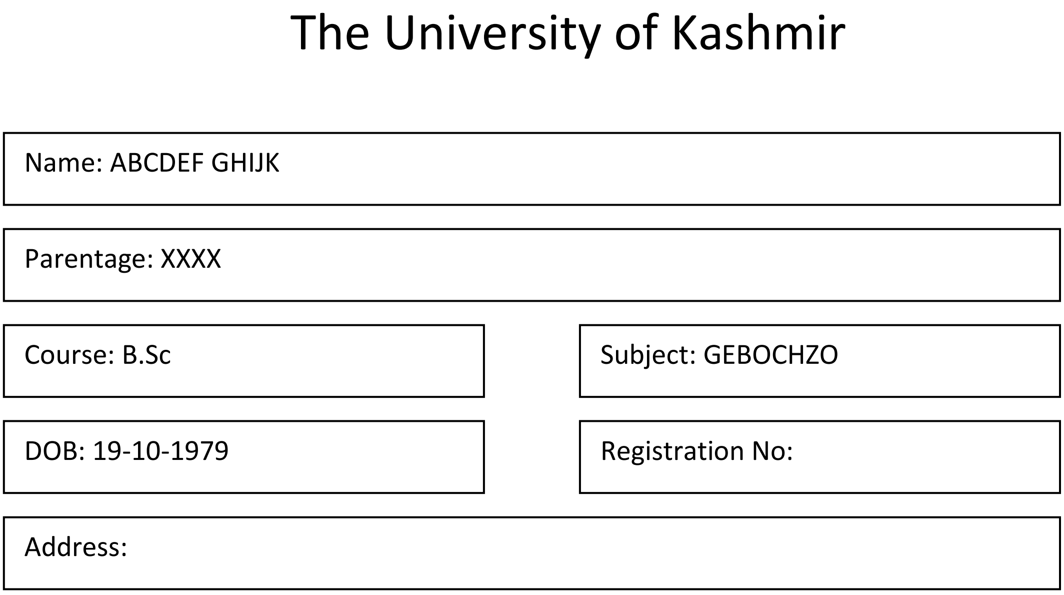
Course:

Subjects:

DOB:

Registration No:

Address:



The University of Kashmir

Name: ABCDEF GHIJK

Parentage: XXXX

Course: B.Sc

Subject: GEBOCHZO

DOB: 19-10-1979

Registration No:

Address:

The data from the above field is extracted and uploaded into the relevant table of the database shown below:

Date of Birth	Registration Number	Name	Parentage	Course	Subject	Session
19-10-1979	5698-S-2006	ABCDEF GHIJK	XXXX	B.Sc	GEBOCHZO	2006

It is obvious that all the data in the table is not only extracted from the form, some data can be obtained from other source as well.

4.11 Security of the Extracted Data

Extracted data is mostly sensitive and its security is paramount to maintain its integrity. The additional measure of security is introduced by selectively encrypting sensitive information, such as Registration No, Course, subject and session before it is stored in the database. The encryption as well as decryption of the data is carried out by the application package thereby securing data regardless of access method. This way even if the database enforced security fails to protect the data still the additional encryption security level prevents it from being used un-authorizedly.

It is found that not whole of the data elements require to be encrypted and depending upon the level of sensitivity only few elements are encrypted before storing into the database. The information classification structure includes these five classification levels, listed in ascending order of sensitivity:

- Unclassified: This information might be freely shared with the public without risk of damage to security. This information is subject to public disclosure under the Freedom of Information Act (FOIA).
- Sensitive But Unclassified (SBU): This level is also known as For Official Use Only (FOUO); its information might not adversely affect the security.
- Confidential: This information would cause damage to security if released to unauthorized parties. This is the lowest level at which information is considered “classified” by the Organization and requires all recipients to have formal security clearance.
- Secret: This information would cause *serious* damage to security if released to unauthorized parties.
- Top Secret: This information would cause *exceptionally grave* damage to security if released to unauthorized parties.

The snap shot of the table when displayed on the screen is shown below:

Date of Birth	Registration Number	Name	Fname	Course	Subject	Session
19-10-1979	********	ABCDEF GHIJK	XXXX	*****	******	*****

4.12 Example

User logs on to the Database System with SELECT privileges, the data in the database is encrypted at the time data is uploaded and decrypted only at the time of data retrieval provided user provides decrypt key. Snap shot of the table would look something like this

Date of Birth	**Registration Number**	**Name**	**Fname**	**Course**	**Subject**	**Session**
19-10-1979	********	ABCDEF GHIJK	XXXX	*****	******	*****
24-10-1977	********	SAMNO	AAAA	*****	******	*****
11-06-1970	********	AIENF	YYYY	********	********	******

- Select Registration Number, Course from Enrolment where Date of Birth= "19-10-1979" ;

As a result of execution of this query the output will be

```
Date of Birth     Registration Number     Course

 19-10-1979          ******               *****
```

This is because the key has not been provided, however if the same query is executed as well as key is provided the result will be as shown below

- Select Registration Number, Course from Enrolment where Date of Birth= "19-10-1979", key;

```
Date of Birth     Registration Number        Course

 19-10-1979          5698-S-2006              B.Sc
```

4.13 Where & When to insert the Decrypt Key

The most suitable is to provide the key with each sql statement i.e. of the form

(Select Registration from Enrolment where Date of Birth = 1000; key)

Key is provided only when encrypted data has to be viewed; it has to be provided at the front end while the user wants to access encrypted data. The user does not need to provide the key if the access to the sensitive data is not required.

4.14 Conclusion

In this chapter the information security was discussed. It is observed that as the businesses are getting more dependent upon the use of information systems the need for better IS security is also increasing. The main goal of defining an IS security policy which is the .protection of information systems against unauthorized access to or modification of information whether in storage, processing or transit, and against the denial of service to authorized users, including those measures necessary to detect, document, and counter such threats was discussed. It was observed that a general purpose solution is not available and in order to ensure a secured environment for the data extraction a need based solution of adding an additional layer of protection is required. This was achieved by using the cryptography on the key data elements before their insertion in the relevant database.

5.1 Introduction

Data Retrieval is still a persistent challenge confronted in applications that need to query across multiple autonomous and heterogeneous data sources. Data integration & extraction is critical in every organization that own a host of data sources, where data sets are being created independently by various divisions of the organization, for better collaboration among various divisions within the organization, data retrieval method that cuts across heterogeneous data sources and is independent of size and access techniques required to access information sources as well as does not require user to have any sort of knowledge that includes location, query language etc, is need of hour, example being piece of data to be searched can be available in m data sources and in k different formats, thus obtaining information from n sources can yield inconsistent/contradictory values/results

There is good amount of regularization as far as World-Wide Web is concerned, while google is worldwide access tool to search and determine source of the information user requires there is still no such tool that can be implemented at enterprise level where there are host of data sources and organization users are still facing difficulty in retrieving data accessible on the intranet of the organization and not on the WWW, in order to access such information users within the organizations need to know a lot including location, access techniques etc while still data consistency & redundancy is beyond the scope of common organization user/s.

GENERIC SEARCH PRINCPLE: Solution making use of Knowledge base where in users of the organization irrespective of their technical ability, data source knowledge and location can search heterogeneous data sources including legacy data sources of organization and retrieve information, also taking into consideration user attributes like his/her location, work profile, designation etc so as to make search more relevant and results more precise.

5.2 Data Mining& Marts

Data Warehouse was need of the hour in enterprises in order to analyze business process for better decision making as per founder of Data Warehouse Inmon: "A Data Warehouse is a subject oriented, integrated, non-volatile and time-variant collection of data in support of management's decisions".

Data Mining is the process of extracting/exploring data from n data sources and re-organizing it for purposes other than what the databases were originally intended for. What data is to be mined varies from company to company, user to user in other words depends on the nature and organization of the data, so there can be no such thing as a generic "data mining tool".

Of course the data must be continuously refreshed, so the scrubbing and reconciliation process must be a permanent feature of the Warehouse, and will have to be modified every time the databases are modified or new databases become available.

Creating and maintaining a Data Warehouse is a huge job even for the largest companies. It can take a long time and cost a lot of money. In fact, it is such a major project companies are turning to Data Mart solutions instead.

A Data Mart is an index and extraction system[Tari, L, 2010]. Rather than bring all the company's data into a single warehouse, the data mart knows what data each database contains and how to extract information from multiple databases when asked[http://www.aaxnet.com, 2012].

Creating a Data Mart can be considered the "quick and dirty" solution, because the data from different databases is not scrubbed and reconciled, but it may be the difference between having information available and not having it available[http://www.aaxnet.com, 2012].

OLAP can be used for data mining or the discovery of previously undiscerned relationships between data items. An OLAP database does not need to be as large as a data warehouse, since not all transactional data is needed for trend analysis. Using Open Database Connectivity (ODBC), data can be imported from existing relational databases to create a multidimensional database for OLAP [http://searchdatamanagement.techtarget.com].

5.4 Artificial Intelligence

Artificial Intelligence, or AI for short, is a combination of computer science, physiology, and philosophy. AI is a broad topic, consisting of different fields, from machine vision to expert systems. The element that the fields of AI have in common is the creation of machines that can "think".

In order to classify machines as "thinking", it is necessary to define intelligence. To what degree does intelligence consist of, for example, solving complex problems, or making generalizations and relationships? And what about perception and comprehension? Research into the areas of learning, of language, and of sensory perception have aided scientists in building intelligent machines. One of the most challenging approaches facing experts is building systems that mimic the behaviour of the human brain, made up of billions of neurons, and arguably the most complex matter in the universe. Perhaps the best way to gauge the intelligence of a machine is British computer scientist Alan Turing's test. He stated that a computer would deserve to be called intelligent if it could deceive a human into believing that it was human.

Artificial Intelligence has come a long way from its early roots, driven by dedicated researchers. The beginnings of AI reach back before electronics, to philosophers and mathematicians such as Boole and others theorizing on principles that were used as the foundation of AI Logic. AI really began to intrigue researchers with the invention of the computer in 1943. The technology was finally available, or so it seemed, to simulate intelligent behavior. Over the next four decades, despite many stumbling blocks, AI has grown from a dozen researchers, to thousands of engineers and specialists; and from programs capable of playing checkers, to systems designed to diagnose disease.

AI has always been on the pioneering end of computer science. Advanced-level computer languages, as well as computer interfaces and word-processors owe their existence to the research into artificial intelligence. The theory and insights brought about by AI research will set the trend in the future of computing. The products available today are only bits and pieces of what are soon to follow, but they are a movement towards the future of artificial intelligence. The advancements in the quest for artificial intelligence have, and will continue to affect our jobs, our education, and our lives [http://library.thinkquest.org/2705/,July 2012].

5.5 Expert Systems

Expert Systems are computer programs that are derived from a branch of computer science research called Artificial Intelligence (AI). AI's scientific goal is to understand intelligence by building computer programs that exhibit intelligent behaviour. It is concerned with the concepts and methods of symbolic inference, or reasoning, by a computer, and how the knowledge used to make those inferences will be represented inside the machine.

Of course, the term intelligence covers many cognitive skills, including the ability to solve problems, learn, and understand language; AI addresses all of those. But most progress to date in AI has been made in the area of problem solving -- concepts and methods for building programs that reason about problems rather than calculate a solution.

AI programs that achieve expert-level competence in solving problems in task areas by bringing to bear a body of knowledge about specific tasks are called knowledge-based or expert systems. Often, the term expert systems is reserved for programs whose knowledge base contains the knowledge used by human experts, in contrast to knowledge gathered from textbooks or non-experts. More often than not, the two terms, expert systems (ES) and knowledge-based systems (KBS), are used synonymously. Taken together, they represent the most widespread type of AI application. The area of human intellectual endeavor to be captured in an expert system is called the task domain. Task refers to some goal-oriented, problem-solving activity. Domain refers to the area within which the task is being performed. Typical tasks are diagnosis, planning, scheduling, configuration and design.

Building an expert system is known as knowledge engineering and its practitioners are called knowledge engineers. The knowledge engineer must make sure that the computer has all the knowledge needed to solve a problem. The knowledge engineer must choose one or more forms in which to represent the required knowledge as symbol patterns in the memory of the computer -- that is, he (or she) must choose acknowledge representation. He must also ensure that the computer can use the knowledge efficiently by selecting from a handful of reasoning methods.

Every expert system consists of two principal parts: the knowledge base; and the reasoning, or inference, engine.

The knowledge base of expert systems contains both factual and heuristic knowledge. Factual knowledge is that knowledge of the task domain that is widely shared, typically found in textbooks or journals, and commonly agreed upon by those knowledgeable in the particular field.

The most important ingredient in any expert system is knowledge. The power of expert systems resides in the specific, high-quality knowledge they contain about task domains. AI researchers will continue to explore and add to the current repertoire of knowledge representation and reasoning methods. But in knowledge resides the power. Because of the importance of knowledge in expert systems and because the current knowledge acquisition method is slow and tedious, much of the future of expert systems depends on breaking the knowledge acquisition bottleneck and in codifying and representing a large knowledge infrastructure [http://www.wtec.org/loyola/kb/c1_s1.htm, july 2012].

5.6 Proposed Generic Search Algorithm

5.6.1 Development of Keyword preprocessor

Preprocessor is meant to check, correct, rearrange/modify user input as and when required. It not only will check & correct spelling mistakes if any making use of traditional dictionary, but also delete extra space, unnecessary ",.;:" and if required rearrange the words, making the best arrangement of user input without damaging the content and context of the input. Preprocessor will have access to organization main databases like employees thus enabling it to access such database to make user input to more precise queries[E. Alfonseca, 2009][S. Agarwal, 2002][N.L. Sarda].

5.6.2 Construction of knowledge base

Defining how our knowledge base will be set up is a crucial first step before we initially populate our knowledge base. If our knowledge base is not well organized and actively managed, the answers can easily become outdated and the information can become disjointed. As a result, finding answers can become more difficult for our users. In addition to initial planning, we need to develop specific procedures for maintaining the knowledge base over the long term so that we can keep the information organized and updated to maximize the effectiveness of our knowledge base.

Before we initially populate our knowledge base with question/answer pairs, we must plan for the growth of the knowledge base by defining and organizing the information you want to present. By organizing information into distinct and logical categories, the information is more accessible and will avoid having to reorganize our knowledge base as it becomes larger. Once our class, categories, and custom fields are in place, we can then develop the processes for proposing, publishing, reviewing & if required executing queries[E. Alfonseca, 2009].

To design and build an effective knowledge base, and ultimately to manage it effectively, we start by addressing the following areas:

a. Consider the amount of information to be presented

b. Identify our audience and the scope of information to be included

c. Define users and categories

d. Define additional custom fields as necessary

e. Develop writing and style guidelines

f. Designate responsibilities for managing our knowledge base

g. Define a process for proposing new answers

h. Define an approval review process for new answers

i. Determine the display position of new answers on the answers lists

j. Notify users of new answers, if any

k. Evaluate customer feedback

l. Determine a process for reviewing existing answers

5.6.3 Creating Mediator Based Integration

Mediator based integration architectures [Md. Sumon, 2010][Stefan Biffl, 2010][Alon Halevy, 2009] define a framework to deal with such problems. In such systems, the mediator holds a schema (mediator schema) which semantically subsumes the interesting parts of the source schemas. Technical and syntactical heterogeneity in the sources is hidden by wrappers which offer a uniform interface to the mediator. In our approach this interface is comprised of a relational export schema (source schema) and the set of possible queries [E. Alfonseca, 2009] against this schema. The mediator tries to find answers for queries against the mediator schema by combining data from different sources which are accessed through their wrappers. In this process, many types of schematic and semantic discrepancies have to be bridged.

5.6.4 Primary and Secondary Classifiers

Primary and Secondary Classifiers play a major role in the Data Mining process. The algorithm involved is shown as under and is also diagrammatically represented in (fig 5.2):

Step 1: Get Input Query from the User.

Step 2: Perform Preprocessing like checking syntax, size and data type involvement.

Step 3: Feed the query to a Primary Classifier. If already such kind of information is mined in the system then using Knowledge base and Data Mart provide the generic information to the user. If there is a failure go to step 4.

Step 4: Broadcast this query to various Heterogeneous data sources[Mohammad Ghulam Ali, 2009][Marc Van Cappellen, 2008][Srujana Merugu, 2005][Ulf Leser]. Also update the user profile information data source and transfer the output to mediator data source.

Step 5: The large data volume present in the mediator data Source is provided to the secondary classifier .

Step 6: Using a Knowledge base information and Mediator Data Base information proper information is filtered. If there is a success then corresponding entries are made in the Knowledge base and the resultant data is passed to the user via a data mart.

Step 7: If there is a failure go back to step 4 till generic information is not mined.

Step 8: Clear the Mediator Temp Data Source for next Query.

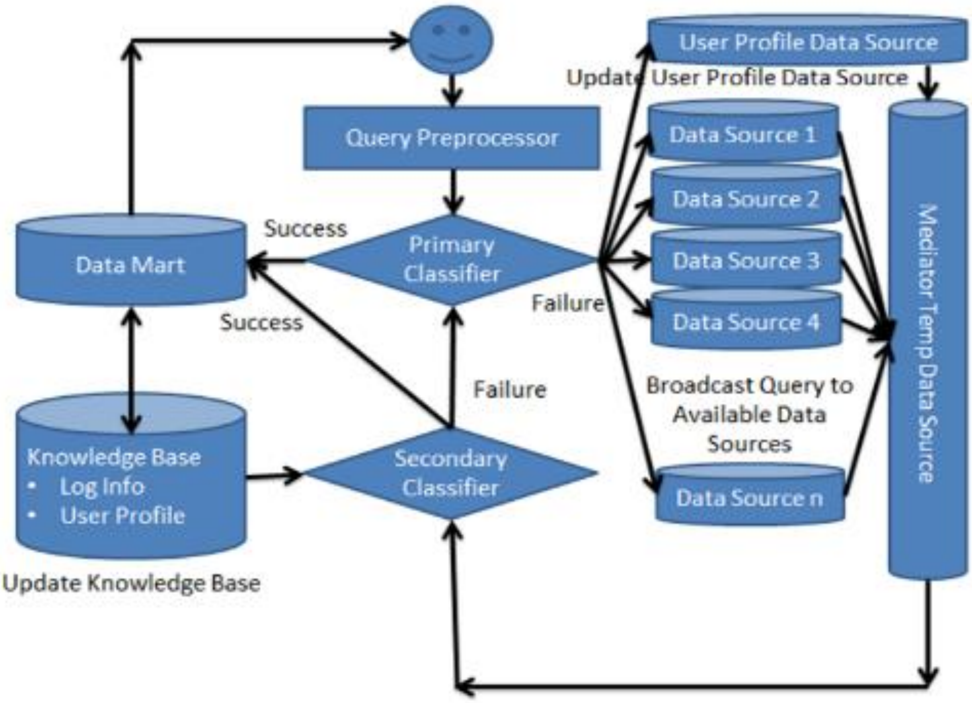


Figure 5.2: Generic Search Optimization

5.7 Performance Details

The performance details for the system under investigation are shown in the graph shown below (fig 5.3). A set of queries were made to run on the system and it was observed that after some training trails the search time was reduced to a large extent and after that remained constant. This was due to direct access from the data mart as the system got trained to such queries as the knowledge and user profile data source base were updated concurrently thus reducing the search time.

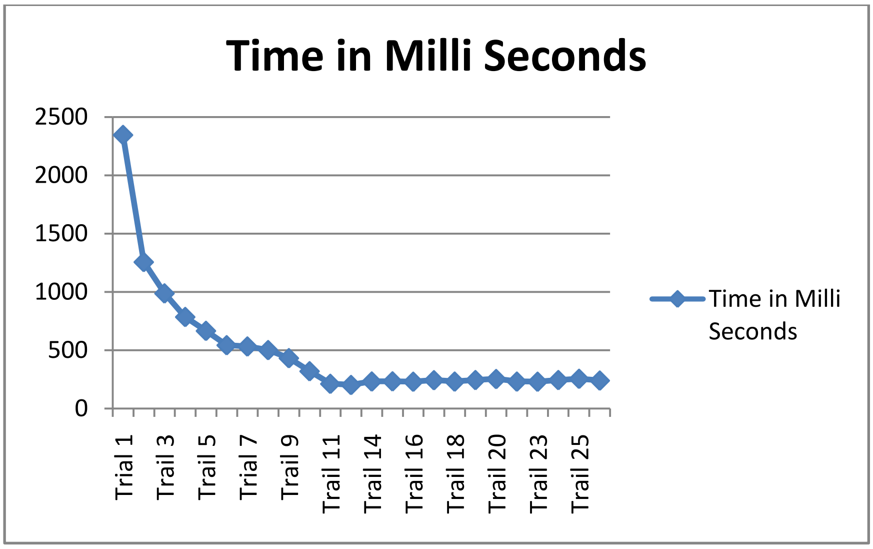


Figure 5.3: Performance Graph for Generic Search

5.8 Conclusion

Data retrieval across multiple heterogeneous[Srujana Merugu, 2005][Ulf Leser] data sources is still a challenge at large because of many reasons including m formats, requires users to be aware of the underlying database schema and also the knowledge of the structured query language specific to the database, data inconsistency across n sources, most of the enterprise database needs lots of reports to be derived/extracted from the underlying data for analysis and decision making purposes & these reports can neither be generic and nor can be determined in advance.

To overcome such problems we proposed GENERIC SEARCH PRINCPLE where in users of the organization irrespective of their technical ability, data source knowledge and location can search n heterogeneous data sources including legacy data sources of organization and retrieve consistent information. This principle also taking into consideration user attributes like his/her location, work profile, designation etc so as to make search more relevant and results more precise.

Data retrieval across multiple heterogeneous [Srujana Merugu, 2005][Ulf Leser] data sources is still a challenge at large because of many reasons including m formats, requires users to be aware of the underlying database schema and also the knowledge of the structured query language specific to the database, data inconsistency across n sources, most of the enterprise database needs lots of reports to be derived/extracted from the underlying data for analysis and decision making purposes & these reports can neither be generic and nor can be determined in advance.

Although Data Integration was considered to be "an area of intellectual curiosity"[M. Vincini, 1999] at its early years, the advent of information sharing nowadays is calling for effective integration approaches realized in practice. Users are not compromising with low standards of information accuracy and are willing to find the right information at the right time. The research community, thus far, has shown excellent progress in dealing with the most crucial problems presented on the way of integrating data, however, further challenges arise constantly: The expansion of (semi -) & unstructured data (XML) for example implies that data sources are even more complex and difficult to handle. Coping with semantic heterogeneity in such scenarios seems almost impossible. However, research is getting even more intense and promising ideas are expected to develop[ETH Group, 2010].

Organizations across the globe while focusing on computerization of departments did not pay much stress upon uniformity and consistency of data [E. Bertino, 2001], and almost all the organizations across the globe ended up with numerous heterogeneous data sources. While data in warehouse must be credible, it must be carefully assembled from a variety of sources around the organization. Data Warehouse not only makes organization information easily accessible but has become tool for data integration.

The future of UOKE's data warehouse is becoming clearer. Initially, the warehouse served as a resource for accessing information from legacy systems. Eventually, the warehouse will serve as a telescope into UOKDW's distributed data stores. Some of these data will reside in the data warehouse, while other elements will be "viewed" from the RDBMSs where the data reside. UOKEDW foresees a time when the telescope extends beyond UOKEDW to other organizations with common goals, such as the neighboring Maricopa County Community College District. The real power of the warehouse will be actualized in years to come. The data warehouse fills an important data administration role in a client/server environment. As distributed application developers move further away from the central computing core, the data elements in the warehouse ensure the integrity of the organization's enterprise data.

The bottom line is that data warehousing is here to stay. Warehousing gives organizations the opportunity to "get their feet wet" in client/ server technology, distributed solutions, and RDBMS. This is essential for any future mission critical application, making the data warehouse a low-risk, high-return investment.

The data warehousing technology is gaining wide attention, and many organizations are building data warehouses (or, data marts) to help them in data analysis in decision for decision support. Data Warehousing is a newly emerged field of study in Computing Sciences. Due to its viz. multidisciplinary nature, it has overlapping area of studies in three different computing disciplines. This overlapping sometimes may cause contradictory definitions for a specific concept. To overcome this problem of data warehousing for Examination Automation System, it was considered for Star Schema Design. In this regard various functional dimensions of the Examination System were designed and connected to a Fact Transaction Dimension. Furthermore the general issues like the Client Statistics and Query Design were taken up and various Decision Support Databases were designed and implemented using the same star Schema.

So far, all fundamental problems have been cleared up. Recovery, Concurrency and Security of the Enterprise Information System Data Warehouse have been solved. However, a lot of delicate improvements can be implemented because the further use of this CDW in the future will tell what is appropriate and what should be improved. User-interface is another area for improvement since it can reduce the error that users might make but requires further work to produce an appropriate and elegant design that suits the user's needs.

User over the years has become more demanding; he/she does not only need information but wants it in specific format which should be complete and correct. Globally centralization was prioritized because of collection of massive data in heterogeneous data sources, however much was not thought for naive user and information system was still at the mercy of technocrats time has now come to stress more upon user demands so as to meet user demands and make user dependency on technocrats minimal, Information Translation Algorithm was proposed where in users can have their piece of information in their desired format. While globally stress is still on data mining not much is thought for data presentation is user desired format, coming years will see drift towards data presentation so as to meet the needs of users.

It is observed that as the businesses are getting more dependent upon the use of information systems the need for better IS security is also increasing. The main goal of defining an IS security policy which is the protection of information systems against unauthorized access to or modification of information whether in storage, processing or transit, and against the denial of service to authorized users, including those measures necessary to detect, document, and counter such threats was discussed. It was observed that a general purpose solution is not available and in order to ensure a secured environment for the data extraction a need based solution of adding an additional layer of protection is required. This was achieved by using the cryptography on the key data elements before their insertion in the relevant database accordingly Algorithm for Encryption/Decryption was proposed.

Finally GENERIC SEARCH PRINCPLE was proposed, where in users of the organization irrespective of their technical ability, data source knowledge and location can search n heterogeneous data sources including legacy data sources of organization and retrieve consistent information. This principle also taking into consideration user attributes like his/her location, work profile, designation etc so as to make search more relevant and results more precise.

2012 **“Information Translation: A Practitioners Approach”, World Congress on Engineering and Computer Science (WCECS), San Francisco, USA. October, 2012.**

(To be presented)

2012 “User Desired Information Translation”, Journal of Global Research in Computer Science, JGRCS, Illinois, USA (ISSN-2229-371X), Volume 3 Issue 6., July, 2012.

2012 “Star Schema Implementation for Automation of Examination Records”, International Conference on Computer Science, Computer Engineering and Applied Computing Las Vegas, USA, July 16-19, 2012 ISBN 1-60132-050-7

2012 “Data Warehouse Implementation of Examination Databases”, International Conference on Computer Science, Computer Engineering and Applied Computing Las Vegas, USA, July 16-19, 2012 ISBN 1-60132-050-7

2012 “Migrated Legacy Data Issues: Recovery, Concurrency & Security”, Journal of Global Research in Computer Science, JGRCS, Illinois, USA (ISSN-2229-371X), Volume 3 Issue 5.

2012 “Integrating Information from Heterogeneous Data Sources: University of Kashmir Case Study”, Journal of Global Research in Computer Science, JGRCS (ISSN-2229-371X), Volume 3 Issue 5 Illinois, USA.

2012 “Information Integration for Heterogeneous Data Sources”, IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 640-643, New York USA.

2012 “Secure Data Entry of Operational Systems for Data Warehouse”, IOSR Journal of Engineering Apr. 2012, Vol. 2(4) , New York USA.

2012 “Generic Search Optimization for Heterogeneous Data Sources” International Journal of Computer Applications, Foundation of Computer Science, New York, USA. April, 2012.

2012 “Data Warehouse Implementation of Examination Databases” International Journal of Computer Applications, Foundation of Computer Science, New York, USA. April, 2012.

2007 “Short Messaging Services Based Examination Information System”, International Conference on Computer Science, Computer Engineering and Applied Computing. Las Vegas, USA, June 25-28, 2007, ISBN 1-60132-050-9

2007 “University Information System Integration Plan: Database Perspective”, International Conference on Computer Science, Computer Engineering and Applied Computing Las Vegas, USA, June 25-28, 2007 ISBN 1-60132-050-7

2007 “Document Security through Encryption”, International Conference on Advances in Computer Vision and Information Technology, Dr. Baba Ambedkar University, Aurngabad. November 15-18, 2007.

2006 “Data Security in Examination System through encryption” J & K Science Congress University of Kashmir. 25-27 July, 2006

- ✓ A. Bauer, H. G¨unzel, eds., “Data Warehouse Systeme — Architektur, Entwicklung, Anwendung, dpunkt.verlag”, 2001.

- ✓ A. Gupta, V. Harinarayan, and D. Quass. Aggregate query processing in data warehousing environments. In Proc. 21thInt. Conf. on Very Large Data Bases,Zurich, Switzerland, 1995.

- ✓ A. Guttman. R-Trees: A Dynamic Index Structure for Spatial Searching. In Proc. ACM SIGMOD Conf., pages 47–57, June 1984.

- ✓ A. Kashyap, J. Dave, M. Zubair, C. P. Wright, and E. Zadok. Using the Berkeley Database in the Linux Kernel.A. Kashyap, S. Patil, G. Sivathanu, and E. Zadok. I3FS: An In-Kernel Integrity Checker and Intrusion Detection File System.

- ✓ A.W. Brown and G. Booch, "Reusing Open-Source Software and Practices: The Impact of Open-Source on Commercial Vendors," Proc. 7th Intl Conf on Software Reuse: Methods, Techniques, and Tools, Springer, 2002.

- ✓ ACM/ IEEE-CS Joint Task Force for Computing Curriculum 2005. “Computing Curriculum 2005”. The Over view report” 30 Sep, 2005

- ✓ Alon Halevy, "Information Integration". In Encyclopedia of Database Systems, 2009.

- ✓ Alon Halevy, Anand Rajaraman and Joann Ordille. “Data Integration: The Teenage Years”, In VLDBConference, pages 9-16, 2006.

- ✓ Alon Y. Levy, Anand Rajaraman, and Joann J. Ordille. Querying Heterogeneous Information Sources Using Source descriptions. In Proceedings of the International Conference on Very Large Databases (VLDB), 1996.

- ✓ B.Thuraisingham "Data Warehousing, Data Mining and Security" IFIP Database Security Conference, July 1996

- ✓ Baral, "Incremental Information Extraction Using Relational Databases". Knowledge and Data Engineering, IEEE Transactions on Issue:99 , pp 25-35, 28 October 2010

- ✓ Battacharya & Saxena, World Comp 2007, June 25-28, 2007, Las Vegas, Nevada, USA, ISBN1-60132-046-9.

- ✓ Bo Yang and Manohar Mareboyana, "Progressive Content-Sensitive Data Retrieval in Sensor Networks". Journal of Computer Science 5 (7):pp 529-535, 2009.

- ✓ C. Fahrner, and G. Vossen. A survey of database transformations based on the Entity-Relationship model. Data & Knowledge Engineering, vol. 15, n. 3, pp. 213-250. 1995.:

- ✓ CAI Yong, HE Guangsheng, "Designing Model of Data Warehouse with OO Method [J]", Computer Engineering and Applications, 2003.6.

- ✓ D. Theodoratos, T. Sellis (DWQ project). "Designing Data Warehouses." DKE '99

- ✓ DataDirect Technologies. DataDirect XQuery Web Service Framework. http://www.xquery.com

- ✓ Date, C. J. (1995), An Introduction to Database Systems, Addison-Wesley Publishing Company, Inc.

- ✓ E. Alfonseca, K. Hall, and S. Hartmann, "Large-scale computation of distributional similarities for queries". In Proceedings of NAACL-HLT, Association for Computational Linguistics, pp 29-32, 2009.

- ✓ E.Bertino, P.Bonatti, E.Ferrari "TRBAC: A Temporal Role Based Access Control", Information and System Security, 2001

- ✓ F. Bancilhon, N. Spyratos, "Update Semantics of Relational Views," ACM TODS 6, 1981, 557–575.

- ✓ Fon Silvers, "Building and Maintaining a Data Warehouse," AN AUERBACH BOOK", CRC Press is an imprint of the Taylor & Francis Group, an informa business

- ✓ H. Lenz, A. Shoshani, "Summarizability in OLAP and statistical databases," Proc. 9th SSDBM 1997, 132–143.

- ✓ H. V. Jagadish. Spatial Search with Polyhedra. In Proc. 6th IEEE Int'l Conf. on Data Engin., 1990.

- ✓ Hot Topics in Data Management System. Data Integration Underlying problems and Research Approaches, ETH Group 2010

- ✓ http://docs.oracle.com/html/E10312_01/dm_concepts.htm

- ✓ http://en.wikipedia.org/wiki/File_format

- ✓ http://pwp.starnetinc.com/larryg/index.html.
- ✓ http://royal.pingdom.com/2012/01/17/internet-2011-in numbers/ 12 May, 2012.
- ✓ http://searchsqlserver.techtarget.com/definition/database
- ✓ http://techcrunch.com/2010/08/04/schmidt-data 15 May,2012.
- ✓ http://www.differencebetween.net/language/differencebetween-data-and-information/
- ✓ I.R. Cruz, X. Huiyong and H. Feihong, "An ontology-based framework for XML semantic integration," Proc. Intl. Database Engineering and Applications Symp., IEEE, 2004, pp. 217-226.
- ✓ Inmon, W. H., "Building the Data Warehouse", Second Edition, John Wiley & Sons, Inc 1996
- ✓ Ingrid M. Olson and Marshall D. Abrams, Essay 7, Information Security Policy- 2007.
- ✓ J. Gray and A. Reuter. Transaction Processing – Concepts and Techniques. Morgan Kaufmann Publishers, 1993.
- ✓ J. Gray. Notes on Database Operating Systems. In Operating Systems – An Advanced Course, volume 60 of Lecture Notes in Computer Science. Springer-Verlag, 1978.
- ✓ J. Hellerstein, J. Naughton, and A. Pfeffer. Generalized Search Trees forDatabase Systems. In Proc. 21st Int'l Conference on Very Large Databases (VLDB), pages 562–573, September 1995.

- ✓ J. Huang and E. Efthimiadis, “Analyzing and evaluating query reformulation strategies in web search logs”. In Proceedings of CIKM, pp 77-86, ACM, 2009.

- ✓ J. Lechtenb¨orger, G. Vossen, “Multidimensional Normal Forms for Data Warehouse Design,” 2002, to appear in Information Systems, Elsevier Science.

- ✓ J. Lechtenb¨orger, G. Vossen, “On the Computation of Relational View Complements,” Proc. 21st PODS 2002, 142–149.

- ✓ Jason Bloomberg and John Goodson. Best Practices for SOA: Building a Data Services Layer. SOA World Magazine, May, 2008.

- ✓ Jeff Lawyer, Shamsul Chowdhury, “ Best Practices in Data Warehousing to Support Business Initiatives and Needs”, Proceedings of the 37th Hawaii International Conference on System Sciences – 2004

- ✓ Jorge Bernardino, Pedro Furtado, Henrique Madeira,” A Cost Effective Approach for Very Large Data Warehouses”, Proceedings of the International Database Engineering and Applications Symposium, 2002

- ✓ K.P. Eswaran, J.N. Gray, R.A. Lorie, and I.L. Traiger. The notion of consistency and predicate locks in database systems. Communications of the ACM, 19(11):624–633,November 1976.

- ✓ Kimball, Ralph, “The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses”, John Wiley & Sons, Inc, 1996.

- ✓ Krishna. “Principles of Curriculum Design and Revision: A Case Study in Implementing Computing Curricula CC2001”. ITiCSE’05, June 27–29, 2005

- ✓ Larry, Greenfield, LGI Systems Inc., "The Data Warehousing Information Center," 1997

- ✓ LIN Yu,etc, “The Principles and Applications of Data Warehouse [M]”, Posts & Telecommunications Press, 2003.1

- ✓ M. A. R. Kortnik, D. L. Moody. “From Entities to Stars, Snowflakes, Clusters, Constellations and Galaxies: A Methodology for Data Warehouse Design.” 18th. International Conference on Conceptual Modelling. Industrial Track Proceedings. ER’99.

- ✓ M. Berry and G. Linoff, Data Mining Techniques For Marketing, Sales, and Customer Support. John Wiley & Sons, 1997.

- ✓ Marc Van Cappellen, Wouter Cordewiner, Carlo Innocenti, “Data Aggregation, Heterogeneous Data Sources and Streaming Processing: How Can XQuery Help? Bulletin of the IEEE Computer Society, Technical Committee on Data Engineering, 2008.

- ✓ Marotta. “A transformations based approach for designing Data Warehouses Internal Report.” InCo. Universidad de la República, Montevideo, Uruguay. 1999.

- ✓ Maurizio Lenzerini. “Data Integration: A Theoretical Perspective”, In Symposium of Principles of Database Systems, 2002.

- ✓ Md. Sumon Shahriar and Jixue Liu, “Constraint-Based Data Transformation for Integration: An Information System Approach”, International Journal of Database Theory and Application Vol. 3, No. 1,pp 85-92, March, 2010.

- ✓ MEHARI, Information risk analysis and management methodology, V3, Concepts and Mechanisms, CLUSIF, October 2004.

- ✓ Mohammad Ghulam Ali, “Object Oriented Approach for integration of heterogeneous databases in a multi database system and local schemas modifications propagation”, international journal of computer sciences and information security, vol 6, No. 2, 2009

- ✓ N.L. Sarda & Ankur Jain. “A System for Keyword-basedSearching in Databases.”

- ✓ Peter Pach, Attila Gyenesei, and Janos Abonyi, “Compact fuzzy association rule based classifier”. Expert Systems with Applications, 2007.

- ✓ Proceedings of the 18th USENIX Large Installation System Administration Conference (LISA 2004), pages 69.79, Atlanta, GA, November 2004. USENIX Association

- ✓ Proceedings of the International Conference on Very Large Databases (VLDB), 1996.

- ✓ R. Ashok Kumar, Dr Y. Rama Devi, "Efficient Approaches for Record level Web Information Extraction Systems". Published in International Journal of Advanced Engineering & Application, pp 161-164, Jan 2011

- ✓ R. Barquin, and S. Edelstein. "Planning andvDesigning, the Data Warehouse",. Prentice Hall, 1996.

- ✓ R. Bayer and M. Schkolnick. Concurrency of Operations on B-Trees. Acta Informatica, 9:1– 21, 1977.

- ✓ R. Hagmann. Reimplementing the cedar _le system using logging and group commit. ACM SIGOPS Operating Systems Review, 21(5):155.162, 1987.

- ✓ R. W. Hamming. Error detecting and error correcting codes. The Bell System Technical Journal, Vol XXVI, April 1950.

- ✓ Ramakrishna Srikant, Sugato Basu, Ni Wang, Daryl Pregibon, "User browsing models: relevance versus examination". In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 223-232, 2010.

- ✓ REN Jinluan, GU Peiliang, ZENG Zhenxiang, "Research on the Methods of Designing Data Structure of Data Warehouse [J]". Computer Engineering and Applications, 2001.22.

- ✓ S. Agarwal, S. Chaudhary, and G. Das. 'Dbxplorer, "Asystem for keyword based search over RelationalDatabases". In proceedings of ICDE 2002.

- ✓ S. Bergamaschi, S. Castano and M. Vincini, "Semantic integration of semi-structured and structured data sources," SIGMOD Rec., vol. 28, 1999, pp. 54-59.

- ✓ S. Vajjhala and J. Fialli. The Java architecture for XML binding (JAXB) 2.0.http://jcp.org/en/jsr/detail?id=222.

- ✓ Srujana Merugu & Joydeep Ghosh "A Distributed Learning Framework for Heterogeneous Data Sources". KDD'05, August 21–24, 2005, Chicago, Illinois, USA.

- ✓ Stefan Biffl, Wikan Danar Sunindyo, Thomas Moser, "Semantic Integration of Heterogeneous Data Sources for Monitoring Frequent-Release Software Projects". International Conference on Complex, Intelligent and Software Intensive Systems, 2010.

- ✓ Svetlozar Nestorov, Nenad Jukić, "Ad-Hoc Association-Rule Mining within the Data Warehouse", Proceedings of the 36th Hawaii International Conference on System Sciences, 2002

- ✓ Syed Najam-ul-Hassan, Maqbool Uddin Shaikh, Uzair Iqbal Janjua," Data Warehousing an Academic Discipline "Curriculum Development Approach, Methodologies and Issues", 2006

- ✓ T. Johnson and D. Shasha. The Performance of Current B-Tree Algorithms. ACM TODS, 18(1), March 1993.

- ✓ Tari, L. Tu, P. Hakenberg, J. Chen, Y. Son, T. Gonzalez, G. Baral, "Incremental Information Extraction Using Relational Databases". Knowledge and Data Engineering, IEEE Transactions on Issue:99 , pp 25-35, 28 October 2010

- ✓ Ulf Leser. “Combining Heterogeneous Data Sources through Query Correspondence Assertions”.

- ✓ http://docs.oracle.com/html/E10312_01/dm_concepts.htm13 May, 2012.

- ✓ W. Hsu and S. Ong. Fossilization: A Process for Establishing Truly Trustworthy Records. IBM Research Report, 2004.

- ✓ W. J. Labio, Y. Zhuge, J. N. Wiener, H. Gupta, H. Garcia-Molina, J. Widom. Stanford University. “The WHIPS Prototype for Data Warehouse Creation and Maintenance”. SIGMOD 1997

- ✓ W. Lehner, J. Albrecht, H. Wedekind, “Normal Forms for Multidimensional Databases,” Proc. 10th SSDBM 1998, 63–72.

- ✓ Whitman & Mattord, Principles of information security, Thompson Course technology, 2nd edition, 2007.

- ✓ Wu Shuning, Cui Deguang, Cheng Peng ,”The Four-stage Standardized Modeling Method in Data Warehouse System Development” Proceedings of the IEEE International Conference on Mechatronics & Automation Niagara Falls, Canada • July 2005

- ✓ YUAN Hong, HE Houcun, “Online Analysis and Data Warehouse Modeling Technologies [J]”, Computer Application Research, 1999.12