# An Overview of Data Warehousing and Design Methodology

**Manisha D. Rakhonde1, Namrata R. Borkar2**

*Department of Computer Science & Engineering, Dr. Sau. K.G.I.E.T., Darapur,*
*Tal – Daryapur, Dist – Amravati, State – Maharashtra (India).*
[1]rakhondemanisha1668@gmail.com
[2]namrata.borkar22@gmail.com

***Abstract* – A data warehouse is a "subject-oriented, integrated, time varying, non-volatile collection of data that is used primarily in organizational decision making and reporting with computing". Data warehouses are central repositories of integrated data from one or more disparate sources. They store current and historical data and are used for creating analytical reports for knowledge workers throughout the enterprise. Examples of reports could range from annual and quarterly comparisons and trends to detailed daily sales analysis. The data stored in the warehouse is uploaded from the operational systems (such as marketing, sales, etc.). The data may pass through an operational data store for additional operations before it is used in the Data Warehouse for reporting. This paper presents a roadmap of data warehousing technologies, focusing on the special requirements that data warehouses place on database management systems (DBMSs) also type design methodologies.**

***Keywords* – Analysis, Data warehouse, OLAP, OLTP.**

## I. INTRODUCTION

Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (Executive, manager, analyst) to make better and faster decisions [9]. The past three years have seen explosive growth, both in the number of products and services offered and in the adoption of these technologies by industry. Data warehousing technologies have been successfully deployed in many industries: manufacturing (for order shipment and customer support), retail (for user profiling and inventory management), financial services (for claims analysis, risk analysis, credit card analysis, and fraud detection), transportation (for fleet management), telecommunications (for call analysis and fraud detection), utilities (for power usage analysis), and healthcare (for outcomes analysis). A data warehouse is a "subject-oriented, integrated, time varying, non-volatile collection of data that is used primarily in organizational decision making."[1] The data warehouse supports on-line analytical processing (OLAP), the functional and performance requirements of which are quite different from those of the on-line transaction processing (OLTP) applications traditionally supported by the operational databases.

### *A. History*

The concept of data warehousing dates back to the late 1980s [2] when IBM researchers Barry Devlin and Paul Murphy developed the "business data warehouse". In essence, the data warehousing concept was intended to provide an architectural model for the flow of data from operational systems to decision support environments. The concept attempted to address the various problems associated with this flow, mainly the high costs associated with it. In the absence of a data warehousing architecture, an enormous amount of redundancy was required to support multiple decision support environments. In larger corporations it was typical for multiple decision support environments to operate independently. Though each environment served different users, they often required much of the same stored data. The process of gathering, cleaning and integrating data from various sources, usually from long-term existing operational systems (usually referred to as legacy systems), was typically in part replicated for each environment. Moreover, the operational systems were frequently reexamined as new decision support requirements emerged. Often new requirements necessitated gathering, cleaning and integrating new data from "data marts" that were tailored for ready access by users.

Data Warehousing is open to an almost limitless range of definitions. Data Warehouses are an important asset for organizations to maintain efficiency, profitability and competitive advantages. Organizations collect data through many sources Online, Call Center, Sales Leads, and Inventory Management. The data collected have degrees of value and business relevance. As data is collected, it is passed through a 'conveyor belt', call the Data Life Cycle Management. An organization's data life cycle management's policy will dictate the data warehousing design and methodology. The goal of Data Warehousing is to generate frontend analytics that will support business executives and operational managers. This paper presents a roadmap of data warehousing technologies, focusing on the special requirements that data warehouses place on database management systems (DBMSs) also type design methodologies.

### *B. Pre-Data Warehouse*

The pre-Data Warehouse zone provides the data for data warehousing. Data Warehouse designers determine which data contains business value for insertion. OLTP databases are where operational data are stored. OLTP databases can reside in transactional software applications such as Enterprise Resource Management (ERP), Supply Chain, Point of Sale, and Customer Serving Software. OLTPs are design for transaction speed and accuracy.

Meta-data ensures the sanctity and accuracy of data entering into the data lifecycle process. Meta-data ensures that data has the right format and relevancy. Organizations can take preventive action in reducing cost for the ETL stage by having a sound Metadata policy. The commonly used terminology to describe Meta-data is "data about data" [2].

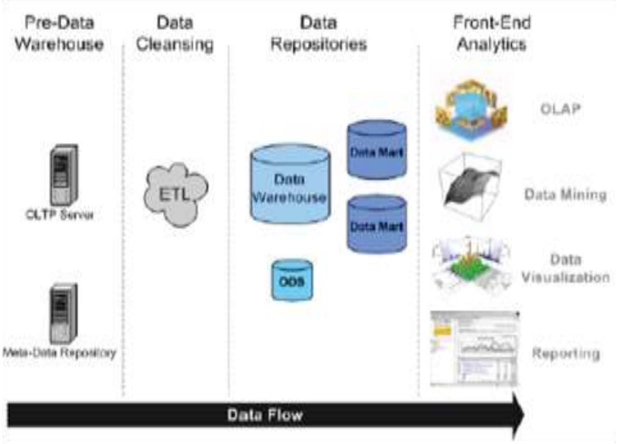

**Fig. 1. Overview of data warehousing infrastructure.**

### *C. Data Cleansing*

Before data enters the data warehouse, the extraction, transformation and cleaning (ETL) process ensures that the data passes the data quality threshold. ETLs are also responsible for running scheduled tasks that extract data from OLTPs.

### *D. Data Repositories*

The Data Warehouse repository is the database that stores active data of business value for an organization. The Data Warehouse modeling design is optimized for data analysis. There are variants of Data Warehouses Data Marts and ODS. Data Marts are not physically any different from Data Warehouses. Data Marts can be thought of as smaller Data Warehouses built on a departmental rather than on a companywide level.

Data Warehouses collects data and is the repository for historical data. Hence it is not always efficient for providing up-to-date analysis. This is where ODS (Operational Data Stores), come in. ODS are used to hold recent data before migration to the Data Warehouse. ODS are used to hold data that have a deeper history that OLTPs. Keep large amounts of data in OLTPs can tie down computer resources and slow down processing imagine waiting at the ATM for 10 minutes between the prompts for inputs.

### *E. Front End Analysis*

The last and most critical portion of the Data Warehouse overview are the frontend applications that business users will use to interact with data stored in the repositories. Data Mining is the discovery of useful patterns in data. Data mining are used for prediction analysis and classification e.g. what is the likelihood that a customer will migrate to a competitor. OLAP, Online Analytical Processing, is used to analyze historical data and slice the business information required. Reporting tools are used to provide reports on the data. Data are displayed to show relevancy to the business and keep track of key performance indicators (KPI). Data Visualization tools is used to display data from the data repository. Often data visualization is combined with Data Mining and OLAP tools. Data visualization can allow the user to manipulate data to show relevancy and patterns.

### *F. Data Warehousing Architectures*

Following are the factors that affect the architectures selections decision [10]:

- Nature of end-user tasks
- Information interdependence between organizational units
- Social/political factors
- Constraints on possessions
- Professed ability of the in-house IT staff
- Upper management’s information needs
- Necessity of need for a data warehouse
- Strategic view of the data warehouse former to implementation
- Compatibility with existing system
- Technical Issues as:
  - What tools will be used to sustain data recovery and analysis?
  - Which database management system should be used?
  - Will data migration tools be used to load the data warehouse?
  - Will parallel processing or partitioning be us?

### *G. Data Mining Capabilities*

Regardless of how smartly and productively the information management system is planned, built and operated; the information management system is basically a repository, or a storage facility. The value is completely dependent on the analytic applications that access, process, and present the data, information and knowledge to sustain research and problem solving necessities. This is the process of data mining [10].

## II. TYPES OF SYSTEMS

With improvements in technology, as well as innovations in using data warehousing techniques, data warehouses have changed from Offline Operational Databases to include an Online Integrated data warehouse [10].

- ➢ Offline Operational Data Warehouses are data warehouses where data is usually copied and pasted from real time data networks into an offline system where it can be used. It is usually the simplest and less technical type of data warehouse.
- ➢ Offline Data Warehouses are data warehouses that are updated frequently, daily, weekly or monthly and that data is then stored in an integrated structure, where others can access it and perform reporting.
- ➢ Real Time Data Warehouses are data warehouses where it is updated each moment with the influx of new data. For instance, a Real Time Data Warehouse might incorporate data from a Point of Sales system and is updated with each sale that is made.

➢ Integrated Data Warehouses are data warehouses that can be used for other systems to access them for operational systems. Some Integrated Data Warehouses are used by other data warehouses, allowing them to access them to process reports, as well as look up current data. There are Three Types of Data Warehouses [10]

- Enterprise Data Warehouse – An enterprise data warehouse provides a central database for decision support throughout the enterprise.
- ODS (Operational Data Store) – This has a broad enterprise wide scope, but unlike the real enterprise data warehouse, data is refreshed in near real time and used for routine business activity. One of the typical applications of the ODS (Operational Data Store) is to hold the recent data before migration to the Data Warehouse. Typically, the ODS are not conceptually equivalent to the Data Warehouse albeit do store the data that have a deeper level of the history than that of the OLTP data.
- Data Mart – Data mart is a subset of data warehouse and it supports a particular region, business unit or business function.

### *A. Data mart*

A data mart is a simple form of a data warehouse that is focused on a single subject (or functional area) hence, they draw data from a limited number of sources such as sales, finance or marketing. Data marts are often built and controlled by a single department within an organization. The sources could be internal operational systems, a central data warehouse, or external data [3]. Denormalization is the norm for data modeling techniques in this system. Given that data marts generally cover only a subset of the data contained in a data warehouse, they are often easier and faster to implement.

### *B. Online analytical processing (OLAP)*

OLAP is characterized by a relatively low volume of transactions. To facilitate complex analyses and visualization, the data in a warehouse is typically modeled multidimensional. For OLAP systems, response time is an effectiveness measure. OLAP applications are widely used by Data Mining techniques. OLAP databases store aggregated historical data in multi-dimensional schemas (usually star schemas). OLAP systems typically have data latency of a few hours, as opposed to data marts, where latency is expected to be closer to one day. The OLAP approach is used to analyze multidimensional data from multiple sources and perspectives. The three basic operations in OLAP are: rollup (increasing the level of aggregation) and drill-down (decreasing the level of aggregation or increasing detail) along one or more dimension hierarchies, slice and dice (selection and projection), and pivot (re-orienting the multidimensional view of data) [4].

For example, in a sales data warehouse, time of sale, sales district, salesperson, and product might be some of the dimensions of interest. Often, these dimensions are hierarchical; time of sale may be organized as a day-month-quarter-year hierarchy, product as a product-category-industry hierarchy.

Data warehouses, in contrast, are targeted for decision support. Historical, summarized and consolidated data is more important than detailed, individual records. The different sources might contain data of varying quality, or use inconsistent representations, codes and formats, which have to be reconciled. Finally, supporting the multidimensional data models and operations typical of OLAP requires special data organization, access methods, and implementation methods, not generally provided by commercial DBMSs targeted for OLTP. It is for all these reasons that data warehouses are implemented separately from operational databases.

Data warehouses might be implemented on standard or extended relational DBMSs, called Relational OLAP (ROLAP) servers. These servers assume that data is stored in relational databases, and they support extensions to SQL and special access and implementation methods to efficiently implement the multidimensional data model and operations.

In contrast, multidimensional OLAP (MOLAP) servers are servers that directly store multidimensional data in special data structures (e.g., arrays) and implement the OLAP operations over these special data structures.
There is more to building and maintaining a data warehouse than selecting an OLAP server and defining a schema and some complex queries for the warehouse.

### *C. Online transaction processing (OLTP)*

OLTP is characterized by a large number of short on-line transactions (INSERT, UPDATE, DELETE). OLTP systems emphasize very fast query processing and maintaining data in multi-access environments. For OLTP systems, effectiveness is measured by the number of transactions per second. OLTP databases contain detailed and current data. The schema used to store transactional databases is the entity model (usually 3NF). Normalization is the norm for data modeling techniques in this system [5].

OLTP applications typically automate clerical data processing tasks such as order entry and banking transactions that are the bread-and-butter for day-to-day operations of an organization. These tasks are structured and repetitive, and consist of short, atomic, isolated transactions. The transactions require detailed, up-to-date data, and read or update a few (tens of) records accessed typically on their primary keys. Operational databases tend to be hundreds of megabytes to gigabytes in size. Consistency and recoverability of the database are critical, and maximizing transaction throughput is the key performance metric. Consequently, the database is designed to reflect the operational semantics of known applications, and, in particular, to minimize concurrency conflicts.

### *D. Predictive analysis*

Predictive analysis is about finding and quantifying hidden patterns in the data using complex mathematical models that can be used to predict future outcomes. Predictive analysis is different from OLAP in that OLAP focuses on historical data analysis and is reactive in nature, while predictive analysis focuses on the future. These systems are also used for CRM (customer relationship management).

## III. BENEFITS

A data warehouse maintains a copy of information from the source transaction systems. This architectural complexity provides the opportunity to:

- Congregate data from multiple sources into a single database so a single query engine can be used to present data.
- Mitigate the problem of database isolation level lock contention in transaction processing systems caused by attempts to run large, long running, analysis queries in transaction processing databases.
- Maintain data history, even if the source transaction systems do not.
- Integrate data from multiple source systems, enabling a central view across the enterprise. This benefit is always valuable, but particularly so when the organization has grown by merger.
- Improve data quality, by providing consistent codes and descriptions, flagging or even fixing bad data.
- Present the organization's information consistently.
- Provide a single common data model for all data of interest regardless of the data's source.
- Restructure the data so that it makes sense to the business users.
- Restructure the data so that it delivers excellent query performance, even for complex analytic queries, without impacting the operational systems.
- Add value to operational business applications, notably customer relationship management (CRM) systems.
- Make decision–support queries easier to write.

## V. DESIGN METHODOLOGIES

### A. *Bottom-up design*

In the *bottom-up* approach, data marts are first created to provide reporting and analytical capabilities for specific business processes. These data marts can then be integrated to create a comprehensive data warehouse. The data warehouse bus architecture is primarily an implementation of "the bus", a collection of conformed dimensions and conformed facts, which are dimensions that are shared (in a specific way) between facts in two or more data marts [6].

### B. *Top-down design*

The *top-down* approach is designed using a normalized enterprise data model. "Atomic" data, that is, data at the greatest level of detail, are stored in the data warehouse. Dimensional data marts containing data needed for specific business processes or specific departments are created from the data warehouse [7]

### C. *Hybrid design*

Data warehouses often resemble the hub and spokes architecture. Legacy systems feeding the warehouse often include customer relationship management and enterprise resource planning, generating large amounts of data. To consolidate these various data models, and facilitate the extract transform load process, data warehouses often make use of an operational data store, the information from which is parsed into the actual DW. To reduce data redundancy, larger systems often store the data in a normalized way. Data marts for specific reports can then be built on top of the DW.

The DW database in a hybrid solution is kept on third normal form to eliminate data redundancy. A normal relational database, however, is not efficient for business intelligence reports where dimensional modelling is prevalent. Small data marts can shop for data from the consolidated warehouse and use the filtered, specific data for the fact tables and dimensions required. The DW provides a single source of information from which the data marts can read, providing a wide range of business information. The hybrid architecture allows a DW to be replaced with a master data management solution where operational, not static information could reside.

The Data Vault Modeling components follow hub and spokes architecture. This modeling style is a hybrid design, consisting of the best practices from both third normal form and star schema. The Data Vault model is not a true third normal form, and breaks some of its rules, but it is a top-down architecture with a bottom up design. The Data Vault model is geared to be strictly a data warehouse. It is not geared to be end-user accessible, which when built, still requires the use of a data mart or star schema based release area for business purposes.

In this section, we describe the design of relational database schemas that reflect the multidimensional views of data. Entity Relationship diagrams and normalization techniques are popularly used for database design in OLTP environments. However, the database designs recommended by ER diagrams are inappropriate for decision support systems where efficiency in querying and in loading data is important [9].

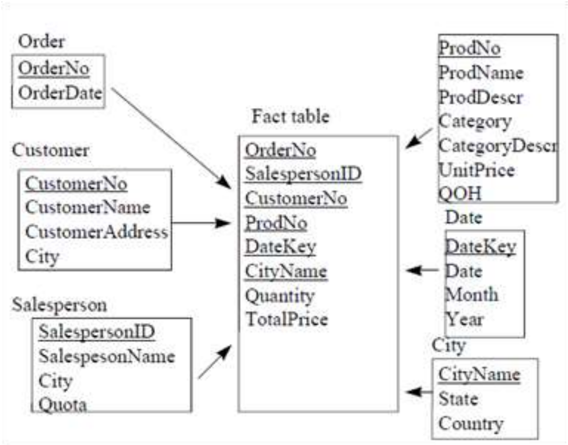

**Fig. 2. A Star Schema**

Most data warehouses use a *star schema* to represent the multidimensional data model. The database consists of a single fact table and a single table for each dimension. Each tuple in the fact table consists of a pointer (foreign key – often uses a generated key for efficiency) to each of the dimensions that provide its multidimensional coordinates, and stores the numeric measures for those coordinates. Each dimension table consists of columns that correspond to attributes of the dimension. Figure 2 shows an example of a star schema.

Star schemas do not explicitly provide support for attribute hierarchies. *Snowflake schemas* provide a refinement of star schemas where the dimensional hierarchy is explicitly represented by normalizing the dimension tables, as shown in Figure 3. This leads to advantages in maintaining the dimension tables. However, the denormalized structure of the dimensional tables in star schemas may be more appropriate for browsing the dimensions [9, 11].

*Fact constellations* are examples of more complex structures in which multiple fact tables share dimensional tables. For example, projected expense and the actual expense may form a fact constellation since they share many dimensions. In addition to the fact and dimension tables, data warehouses store selected summary tables containing pre-aggregated data. In the simplest cases, the pre-aggregated data corresponds to aggregating the fact table on one or more selected dimensions. Such pre-aggregated summary data can be represented in the database in at least two ways.

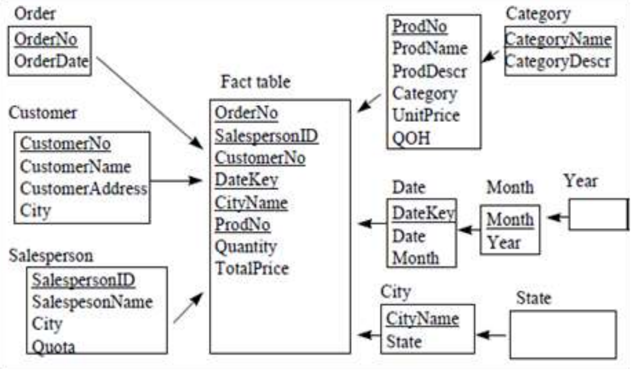

**Fig. 4. A Snowflake Schema.**

Let us consider the example of a summary table that has total sales by product by year in the context of the star schema of Figure 2. We can represent such a summary table by a separate fact table which shares the dimension Product and also a separate shrunken dimension table for time, which consists of only the attributes of the dimension that makes sense for the summary table (i.e., year). Alternatively, we can represent the summary table by encoding the aggregated tuples in the same fact table and the same dimension tables without adding new tables.

This may be accomplished by adding a new level field to each dimension and using nulls: We can encode a day, a month or a year in the Date dimension table as follows: (id0, 0, 22, 01, 1960) represents a record for Jan 22, 1960, (id1, 1, NULL, 01, 1960) represents the month Jan 1960 and (id2, 2, NULL, NULL, 1960) represents the year 1960. The second attribute represents the new attribute level: 0 for days, 1 for months, 2 for years. In the fact table, a record containing the foreign key id2 represents the aggregated sales for a Product in the year 1960. The latter method, while reducing the number of tables, is often a source of operational errors since the level field needs be carefully interpreted [11].

## VII. EVOLUTION IN ORGANIZATION USE

These terms refer to the level of sophistication of a data warehouse:

### *A. Offline operational data warehouse*

Data warehouses in this stage of evolution are updated on a regular time cycle (usually daily, weekly or monthly) from the operational systems and the data is stored in an integrated reporting-oriented data

### *B. Offline data warehouse*

Data warehouses at this stage are updated from data in the operational systems on a regular basis and the data warehouse data are stored in a data structure designed to facilitate reporting.

### *C. On time data warehouse*

Online Integrated Data Warehousing represent the real time Data warehouses stage data in the warehouse is updated for every transaction performed on the source data.

### *D. Integrated data warehouse*

These data warehouses assemble data from different areas of business, so users can look up the information they need across other systems [8].

## VIII. APPLICATION AREAS

Following section briefly describes the different application areas for which data warehouses are built.

### *A. Retail Sales*

Data is collected at several interesting places in a grocery store. Some of the most useful data is collected at the cash registers as customers purchase products. Modern grocery store scans the bar codes directly into the point of sale system. The POS system is at the front door of the grocery store where consumer takeaway is measured. The back door, where vendors make deliveries, is another interesting data collection point [8]. At the grocery store, management is concerned with logistics of ordering, stocking, and selling products while maximizing profit. Some of the most significant management decisions are on pricing and promotions. Both store management and marketing spend a great deal of time tinkering with pricing and promotions. In such scenarios, data warehouses come to rescue.

### *B. Telecommunications*

A telecommunications company generates hundreds of millions of call-detail transactions in a year. For promoting proper products and services, the company needs to analyze these detailed transactions. The data warehouse for the company has to store data at the lowest level of detail.

### *C. Transportation*

In this case, the airline's marketing department wants to analyze the flight activity of each member of its frequent flyer program. The department is interested in seeing which flights the company's frequent flyers take, which planes they fly, what fare basis they pay, how often they upgrade, how they earn. These requirements can be fulfilled by data warehouse.

### *D. Education*

There are some efforts in the area of data warehouse for building data warehouse for education domain. The paper by Carlo DELL'AQUILA [13, 14] summarizes the experience in designing and modeling an academic data warehouse.

Existing facilities and databases affect the chosen data warehouse that brings them together to support decisional activities leading the whole university environment, including administrators, faculties and students. The choice to develop a dedicated system is mainly forced by the peculiar information type that defines the basic information in data warehouse widely different from institution to institution .In the article titled 'What academia can gain from building a data warehouse' by David Wierschem, et.al [15].The authors have identified the opportunities associated with developing a data warehouse in an academic environment. They begin by explaining what a data warehouse is and what its informational contents may include, relative to the academic environment. Next they addressed the current environment drivers that provide the opportunities for taking advantage of a data warehouse and some of the obstacles inhibiting the development of an academic data warehouse.

## XI. CONCLUSION

This paper presents a roadmap of data warehousing technologies, focusing on the special requirements that data warehouses place on database management systems (DBMSs) also type design methodologies. We have described data warehouse operational processes and techniques to design, administrate and facilitate the evolution of the data warehouse.

Creating and managing a warehousing system is hard. Many different classes of tools are available to facilitate different aspects of the process. Development tools are used to design and edit schemas, views, scripts, rules, queries, and reports. Planning and analysis tools are used for what-if scenarios such as understanding the impact of schema changes or refresh rates, and for doing capacity. The star and snowflake schemas are more efficient for data warehouse design as they are easy to learn and need fewer joins [12].

## REFERENCES

[1]. Inmon, W.H., Hackathorn, and R.D (1994) Using the data warehouse. Wiley-QED Publishing, Somerset, NJ, USA.

[2]. Rainer, R. Kelly (20120501). Introduction to Information Systems: Enabling and Transforming Business, 4th Edition (Kindle Edition). Wiley. pp. 127-133.

[3]. http://docs.oracle.com/html/E10312_01/dm_concepts.htm Data Mart Concepts

[4]. https://intellipaat.com/tutorial/datawarehousetutorial "Data Warehouse in Nutshell"

[5]. http://datawarehouse4u.info/OLTPvsOLAP. html OLTP vs OLAP

[6]. "The Bottom Up Misnomer Decision Works Consulting". DecisionWorks Consulting. Retrieved 2016-03-06.

[7]. Gartner, Of Data Warehouses, Operational Data Stores, Data Marts and Data Outhouses, Dec 2005.

[8]. http://www.tech-faq.com/data-warehouse.html

[9]. Chaudhuri S., Dayal U.," An Overview of Data Warehousing and OLAP Technology", *Proc. of EDBT*, 1998.

[10]. Ofori Boateng, Jagir Singh, Greeshma, P Singh, "data warehousing" Business Intelligence Journal - July, 2012 Vol.5 No.2, PP – 224 to 234.

[11]. Gray J. et.al. "Data Cube: A Relational Aggregation Operator Generalizing Group-by, Cross-Tab and Sub Totals" Data Mining and Knowledge Discovery Journal, Vol 1, No 1, 1997.

[12]. Kamal Alaskar and Akhtar Shaikh. (2009)" Object Oriented Data Modeling for Data Warehousing (An Extension of UML approach to study Hajj pilgrim's private tour as a Case Study). International Arab Journal of e-Technology, Vol. 1, No. 2.

[13]. Carlo DELL'AQUILA,'An Academic Data Warehouse' World Scientific and Engineering Academy and Society (WSEAS) Stevens Point, Wisconsin, USA ©2007.

[14]. Manjunath T.N, Ravindra S Hegadi, Ravikumar G K."Analysis of Data Quality Aspects in Data Warehouse Systems",(IJCSIT) International Journal of Computer Science and Information Technologies, vol.2 (1) , 2010, 477-485.

[15]. Jaideep Srivastava, Ping-Yao Chen,"Warehouse Creation-A Potential Roadblock to Data Warehousing", IEEE Transactions on Knowledge and Data Engineering January/February 1999 (vol. 11, no.1) pp.118-126.