Data warehouse & Data Mining logical design Implementation

Mohammad S. Qaseem, Dr. A. Govardhan, S. Nasira Tabassum, Syed.Asadullah Hussaini

Abstract— Based on the data warehouse, data mining techniques, using the method of system, The paper analyzed the use of data mining strategy in enterprise financing decisions, Studied the financing decision system design goal, the function and the logic structure model, explored the data warehouse model and data mining model structure and realization method. It is found that heterogeneous data integration is the basis for financing decision system design. Key lies in the multi-dimensional data warehouse model building, data mining algorithm design and expression. These results are importan t significance on promoting the construction of enterprise financing decision system, realizing enterprise financing decision automation and intelligent.

Index Terms— Data warehouse, data mining techniques, data mining strategy, multi-dimensional data warehouse.

1 INTRODUCTION

Presently almost all businesses have operational systems and these systems are not giving them any competitive advantage. These systems have gathered a vast amount of "data" over the years. The companies are now realizing the importance of this "hidden treasure" of information. Efforts are now on to tap into this information that will improve the quality of their decision-making.

A "data warehouse" is nothing but a repository of data collected from the various operational systems of an organization. This data is then comprehensively analyzed to gain competitive advantage. The analysis is basically used in decision making at the top level.

Data Warehousing technology has grown much in scale and reputation in the past few years, as evidenced by the increasing number of products, vendors, organizations, and yes books, even books, devoted to the subject. Enterprises that have successfully implemented data warehouses find it strategic and often wonder how they ever managed to survive without it in the past.

Data mining is a field at the intersection of computer science and statistics. It is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

• *Mohammad S. Qaseem an Associate Professor and Head of the CSIT department in Nizam Institute of Enggineering and Technology, AP. India. E-mail: ms_qaseem@yahoo.com*
• *Dr. A. Govardhan is Professor of CSE department,& DE of Jawaharlal Nehru Technical University, Hyderabad, Andhra Pradesh, India. (Email: govardhan_cse@yahoo.co.in)*
• *S. Nasira Tabassum is currently pursuing M.Tech in Soft-ware Engineering from Nizam Institute of Engineering and Technology, AP India. (Email: nasira_tabassum@gmail.com).*
• *Syed.Asadullah Hussaini Master of Technology From Shadan College of Engineering & Technology, Affiliated to JNTU Hyderabad, AP India. His areas of interest include data mining, web technologies. (Email: asad_mtech@yahoo.com)*

2 DATA WAREHOUSING

To successfully engage data mining in our processes, the first step is to know who our customers are. We are able to list them by name, job title, function, and business unit, and communicate with them regularly. Next we must be able to identify the appropriate business opportunities. In MIS, our priorities are based on business needs as articulated to us by our clients through ad hoc requests and project management meetings and processes. Constant communication, integration and feedback are required to ensure we are investing our resources in proper ways. Once having identified our customer base and business cases, we must be able to transform data into useful information. Transforming and presenting data as information is our primary function in the corporation. We are constantly looking for new and improved ways to accomplish this directive. The latest evolution in efficiently transforming and presenting data is formal data warehousing practices with browser based front ends. Source data is crucial to data quality and mining efforts. As each new on-line transactional system and data base plat form is introduced the complexity of our tasks increases. "Using operational data presents many challenges to integrators and analysts such as bad data formats, confusing data fields, lack of functionality, legal ramifications, organizational factors, reluctance to change, and conflicting timelines Also, the more disparate the input data sources, the more complicated the integration. A clear definition of the business need is also required to ensure the accuracy of the end results. Defining a logical view of the data needed to supply the correct information, independent of source data restraints, is necessary. Here clients and analysts get the opportunity to discuss their business needs and solutions proactively.Next, a mapping from the physical source data to the logical view is required and usually involves some compromises from the logical view due to physical data constraints. Then questions about presentation can begin to be answered. Who needs it? How often? In what format? What technology is available?

The first iteration of our SAS Data Warehousing solution accesses five operational systems existing on six data platforms. In addition to printed reports the users expect, the data warehouse is also accessible through MDDB OLAP technology over the intranet. Users can now ask and answer their own questions, enabling the creativity needed for successful data mining. With help from the SAS System, we are busily integrating additional data, accessing more data platforms and streamlining our processes.

3 METADATA

Formal recording of metadata is also crucial to data warehousing and data mining exercises. Metadata describes data in terms of entities, attributes and relationships that are meaningful on the business level. It describes product hierarchy, the customer attributes, the relationships between the business and various partners, and other data attributes such as when and where the data is vailable and what applications use it. Metadata must be flexible to change since a data warehouse is not a static environment and must respond repeatedly to changes in the business and systems environment it was built to support. "Metadata provides the key link between the business users and the data. It describes the data in business terms. A good metadata system gives users the ability to browse through the metadata on their desktops… making users more comfortable with the data warehouse. Metadata what is available and where Data Marts fast, specialized access for end users and analysts Operational Feedback integrates decision support back to operating systems End-users the reason for the Data Warehouse in the first place.

4 DATA WAREHOUSE TO SUPPORT DATA MINING DESIGN

A data warehouse designed for data mining needs 1) a central repository that contains detailed data, 2) a hardware investment for the central repository that supports a variety of tools, and 3) regular use to measure the effectiveness of campaigns, especially those based on results from data mining.
MIS is building an Operational Data Store in our HP-UNIX SAS environment that contains detail data from every operational data source needed to meet our users' business needs. We are implementing a second server to act as our deployment server' for all web based end-user applications. Consistent and timely information and upgrades help ensure our users will continue to use our data warehouse solutions.

5 WHAT IS A DATA WAREHOUSE?

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 can include data from other sources. Data warehouses separate analysis workload from transaction workload and enable an organization to consolidate data from several sources.In addition to a relational database, a data warehouse environment includes an extraction, transportation, transformation, and loading (ETL) solution, online analytical processing (OLAP) and data mining capabilities, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users. A common way of introducing data warehousing is to refer to the characteristics of a data warehouse as set forth by William Inmon:

- Subject Oriented
- Integrated
- Nonvolatile
- Time Variant

Subject Oriented:

Data warehouses are designed to help you analyze data. For example, to learn more about your company's sales data, you can build a data warehouse that concentrates on sales. Using this data warehouse, you can answer questions such as "Who was our best customer for this item last year?" This ability to define a data warehouse by subject matter, sales in this case, makes the data warehouse subject oriented.

Integrated:

Integration is closely related to subject orientation. Data warehouses must put data from disparate sources into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this, they are said to be integrated.

Nonvolatile:

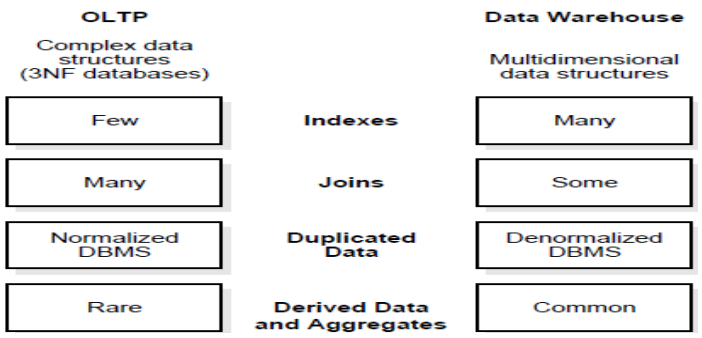
Nonvolatile means that, once entered into the data warehouse, data should not change. This is logical because the purpose of a data warehouse is to enable you to analyze what has occurred.

Time Variant:

In order to discover trends in business, analysts need large amounts of data. This is very much in contrast to online transaction processing (OLTP) systems, where performance requirements demand that historical data be moved to an archive. A data warehouse's focus on change over time is what is meant by the term time variant.

6 CONTRASTING OLTP AND DATA WAREHOUSING ENVIRONMENTS:

Below are illustrated key differences between an OLTP system and a data warehouse.



One major difference between the types of system is that data warehouses are not usually in third normal form (3NF), a type of data normalization common in OLTP environments.Data warehouses and OLTP systems have very different requirements. Here are some examples of differences between typical data warehouses and OLTP systems:

Workload:

Data warehouses are designed to accommodate ad hoc queries. You might not know the workload of your data warehouse in advance, so a data warehouse should be optimized to perform well for a wide variety of possible query operations.

OLTP systems support only predefined operations. Your applications might be specifically tuned or designed to support only these operations.

Data modifications:

A data warehouse is updated on a regular basis by the ETL process (run nightly or weekly) using bulk data modification techniques. The end users of a data warehouse do not directly update the data warehouse.

In OLTP systems, end users routinely issue individual data modification statements to the database. The OLTP database is always up to date, and reflects the current state of each business transaction.

Schema design:

Data warehouses often use denormalized or partially denormalized schemas (such as a star schema) to optimize query performance.OLTP systems often use fully normalized schemas to optimize update/insert/delete performance, and to guarantee data consistency.

Typical operations:

A typical data warehouse query scans thousands or millions of rows. For example, "Find the total sales for all customers last month."

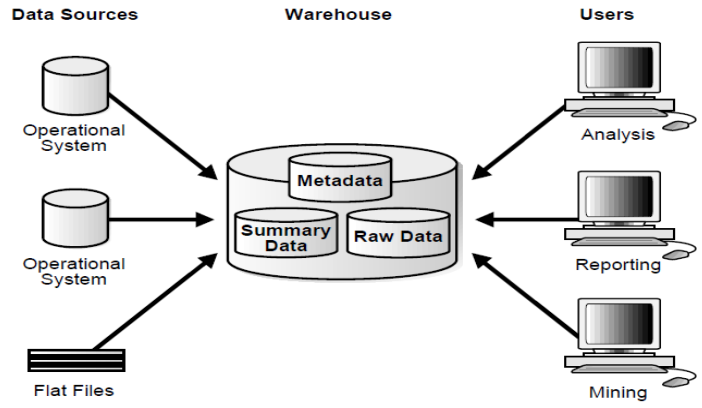
A typical OLTP operation accesses only a handful of records. For example, "Retrieve the current order for this customer."

Historical data:

Data warehouses usually store many months or years of data. This is to support historical analysis.OLTP systems usually store data from only a few weeks or months. The OLTP system stores only historical data as needed to successfully meet the requirements of the current transaction.

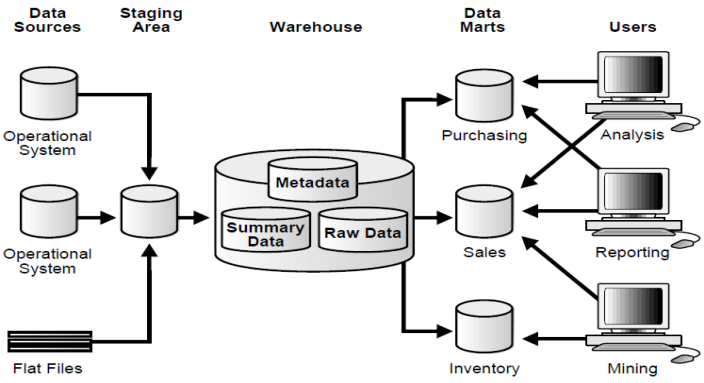
7 Architecture of a Data Warehouse with a Staging Area

In Figure, the metadata and raw data of a traditional OLTP system is present, as is an additional type of data, summary data. Summaries are very valuable in data warehouses because they pre-compute long operations in advance. For example, a typical data warehouse query is to retrieve something such as August sales. A summary in an Oracle database



is called a materialized view.

Note: Data marts are an important part of many data warehouses, but they are not the focus of this book.



8 Logical Design in Data Warehouses

This chapter explains how to create a logical design for a data warehousing environment and includes the following topics:

- Logical Versus Physical Design in Data Warehouses
- Creating a Logical Design
- Data Warehousing Schemas
- Data Warehousing Objects

Creating a Logical Design:

A logical design is conceptual and abstract. You do not deal with the physical implementation details yet. You deal only with defining the types of information that you need. One technique you can use to model your organization's logical information requirements is entity-relationship modeling. Entity-relationship modeling involves identifying the things of importance (entities), the properties of these things (attributes), and how they are related to one another (relationships).

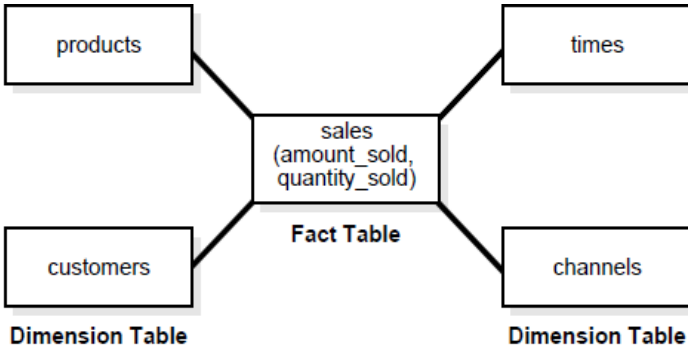
The process of logical design involves arranging data into a series of logical relationships called entities and attributes. An entity represents a chunk of information. In relational databases, an entity often maps to a table. An attribute is a component of an entity that helps define the uniqueness of the entity.

In relational databases, an attribute maps to a column. To be sure that your data is consistent, you need to use unique identifiers. A unique identifier is something you add to tables so that you can differentiate between the same item when it appears in different places. In a physical design, this is usually a primary key. While entity-relationship diagramming has traditionally been associated with highly normalized models such as OLTP applications, the technique is still useful for data warehouse design in the form of dimensional modeling. In dimensional modeling, instead of seeking to discover atomic units of information (such as entities and attributes) and all of the relationships between them, you identify which information belongs to a central fact table and which information belongs to its associated dimension tables. You identify business subjects or fields of data, define relationships between business subjects, and name the attributes for each subject. Your logical design should result in a set of entities and attributes corresponding to fact tables and dimension tables and a model of operational data from your source into subject-oriented information in your target data warehouse schema. You can create the logical design using a pen and paper, or you can use a design tool such as Oracle Warehouse Builder (specifically designed to support modeling the ETL process) or Oracle Designer (a general purpose modeling tool).

Data Warehousing Schemas:

A schema is a collection of database objects, including tables, views, indexes, and synonyms. You can arrange schema objects in the schema models designed for data warehousing in a variety of ways. Most data warehouses use a dimensional model.The model of your source data and the requirements of your users help you design the data warehouse schema. You can sometimes get the source model from your company's enterprise data model and reverse-engineer the logical data model for the data warehouse from this. The physical implementation of the logical data parameters are size of machine, number of users, storage capacity, type of network, and software.

StarSchemas: The star schema is the simplest data warehouse schema. It is called a star schema because the diagram resembles a star, with points radiating from a center. The center of the star consists of one or more fact tables and the points of the star are the dimension tables, as shown in Figure.



The most natural way to model a data warehouse is as a star schema, where only one join establishes the relationship between the fact table and any one of the dimension tables. A star schema optimizes performance by keeping queries simple and providing fast response time. All the information about each level is stored in one row.

Note: Oracle Corporation recommends that you choose a star schema unless you have a clear reason not to.

Other Schemas: Some schemas in data warehousing environments use third normal form rather than star schemas. Another schema that is sometimes useful is the snowflake schema, which is a star schema with normalized dimensions in a tree structure.

Data Warehousing Objects:

Fact tables and dimension tables are the two types of objects commonly used in dimensional data warehouse schemas.Fact tables are the large tables in your data warehouse schema that store business measurements. Fact tables typically contain facts and foreign keys to the dimension tables. Fact tables represent data, usually numeric and additive, that can be analyzed and examined. Examples include sales, cost, and profit.Dimension tables, also known as lookup or reference tables, contain the relatively static data in the data warehouse. Dimension tables store the information you normally use to contain queries. Dimension tables are usually textual and descriptive and you can use them as the row headers of the result set. Examples are customers or products.

Fact Tables: A fact table typically has two types of columns: those that contain numeric facts (often called measurements), and those that are foreign keys to dimension tables. A fact table contains either detail-level facts or facts that have been aggregated. Fact tables that contain aggregated facts are often called summary tables. A fact table usually contains facts with the same level of aggregation. Though most facts are additive, they can also be semi-additive or non-additive. Additive facts can be aggregated by simple arithmetical addition. A common example of this is sales. Non-additive facts cannot be added at all. An example of this is averages. Semi-additive facts can be aggregated along some of the dimensions and not along others. An example of this is inventory levels, where you cannot tell what a level means simply by looking at it.

Dimension Tables: A dimension is a structure, often composed of one or more hierarchies, that categorizes data. Dimensional attributes help to describe the dimensional value. They are normally descriptive, textual values. Several distinct dimensions, combined with facts, enable you to answer business questions. Commonly used dimensions are customers, products, and time. Dimension data is typically collected at the lowest level of detail and then aggregated into higher level totals that are more useful for analysis. These natural rollups or aggregations within a dimension table are called hierarchies.

9 LEVELS

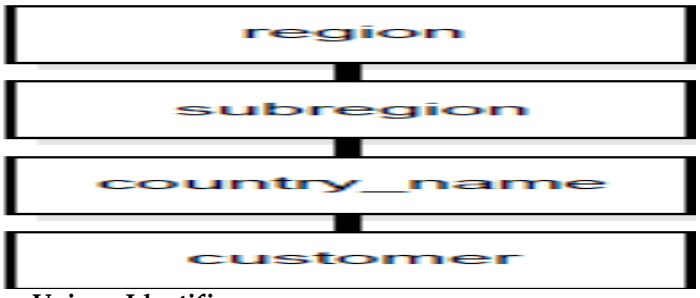
A level represents a position in a hierarchy. For example, a time dimension might have a hierarchy that represents data at the month, quarter, and year levels. Levels range from general to specific, with the root level as the highest or most general level. The levels in a dimension are organized into one or more hierarchies.

Level Relationships: Level relationships specify top-to-bottom ordering of levels from most general (the root) to most specific information. They define the parent-child relationship between the levels in a hierarchy.

Hierarchies are also essential components in enabling more complex rewrites. For example, the database can aggregate existing sales revenue on a quarterly base to a yearly aggregation when the dimensional dependencies between quarter and year are known.

Typical Dimension Hierarchy:

Figure illustrates a dimension hierarchy.



Unique Identifiers:

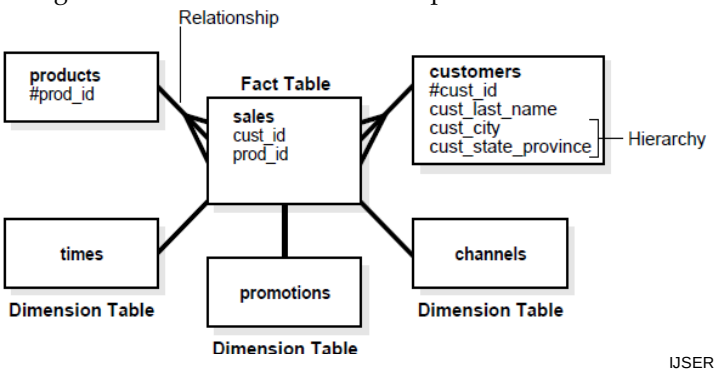
Unique identifiers are specified for one distinct record in a dimension table. Artificial unique identifiers are often used to avoid the potential problem of unique identifiers changing. Unique identifiers are represented with the # character. For example, #customer_id.

Relationships:

Relationships guarantee business integrity. An example is that if a business sells something, there is obviously a customer and a product. Designing a relationship between the sales information in the fact table and the dimension tables products and customers enforces the business rules in databases.

Example of Data Warehousing Objects and their Relationships:

Figure illustrates a common example of a sales fact table and dimension tables customers, products, promotions, times, and channels.



10 THE FUTURE OF DATA WAREHOUSING:

We are moving to the stage of a data ware housing applications that can provide information to many decision makers operational, strategic, and tactical and also to the customers as well in an integrated fashion.

Data Mining: Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determinethe impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data. With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments. For example, Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers. American Express can suggest products to its cardholders based on analysis of their monthly expenditures. WalMart is pioneering massive data mining to transform its supplier relationships.

The logical Design: The logical design of data warehouse is defined by the dimensional data modeling approach. The dimensioning design process followed in this project adheres to the methodology described by Kimball and Ross (2002).

11 CONCLUSION

In this paper, we have been able to demonstrate the process of designing and developing data-warehouse and data mining applications using a case study in an academic environment. It is to be noted however that this technique can be applied to any organization wishing to implement business intelligence as part of their strategic decision support operations. The power of Data-Warehousing in data analysis is tremendous and data-mining can discover hidden treasures in the data-warehouse. Organizations, particularly in Nigeria can begin to implement this project as part of their strategic decision making process tools. With it they can begin to see trends of things historically as their day to day operational data accumulates over the years. They can forecast the future using neural Networks, regression analysis, and other data mining operations incorporated into datawarehouse is the solution we need at the moment to catapult our business or our organization to the next level.

REFERENCES

[1] Michael J. A., & Linoff, Gordon (1997), Data Mining Techniques For Marketing, Sales and Customer Support. USA: John Wiley & Sons, Inc.SAS Institute

Inc. (1998)

[2] Rapid Warehousing Methodology, Cary,NC: SAS Institute Inc.Taylor, David A. (1995)

[3] Business Engineering With Object Technology. USA: John Wiley & Sons.Efraim T., Jay E., Teng-Peng L., Ramesh S, (2010).

[4] Decision Support and Business Intelligence Systems (8thed) Prentice Hall. Infogold Data Warehouse, Data Mart, Data Mining, and Decision Support Resources, http://infogoal.com/dmc/dmcdwh .htm. BIN (2007)

[5] Business Intelligence Network, http://www.beyenetwork.com/home/ Bill Palace (1996).

[6] Data Mining Technology Note prepared for Management 274A Anderson Graduate School of Management at UCLA

[7] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/index.htm. David Heise: An online publication on Data Warehousing in higher Education, http://dheise.andrews.edu/dw/DWData.htm Fang, R. and Tuladhar, S. 2006.

[8] "Teaching Data Warehousing and Data Mining in a Graduate Program in Informa-tion Technology," Journal of Computing Sciences in Colleges, Vol. 21, Issue 5, pp. 137-144. Kimball R. and Ross M. (2002).

[9] The Data Warehouse Toolkit, second edition. John Wiley and Sons, Inc., USA. Michael A. King (2009).

[10] A Realistic Data Warehouse Project: An Integration of Microsoft Access and Microsoft Excel Advanced Features and Skills Journal of Information Technology Education Vol. 8, 2009 Innovations in Practice. Pierce E.M. (1999).

[11] "Developing and Deli-vering a Data Warehousing and Data Mining Course,"Communications of the AIS, Vol. 2, Article 16, pp. 1-22.Jacobson R. (2000).

[12] Microsoft SQL Server (2000). Analysis Services, Step by Step; Microsoft Press, Redmond, Washington.Stephen Brobst, Joe Rarey (2003).

[13] Five Stages of Datawarehouse Decision Support Evolution http://dssresources.com/papers/features/brob&rarey01062003.html.

[14] Wierschem D., McMillan J. and McBroom R. (2003). "What Academia Can Gain from Building a data Warehouse.

Author Profile:

Mohammad S. Qaseem is an Associate Professor and Head of the CSIT department in Nizam Institute of Enggineering and Technology, AP. India, His areas of interest include data mining, web technologies. (Email: ms_qaseem@yahoo.com)

Dr. A. Govardhan is Professor of CSE department,& DE of Jawaharlal Nehru Technical University, Hyderabad, Andhra Pradesh, India. (Email: govardhan_cse@yahoo.co.in)

S. Nasira Tabassum is currently pursuing M.Tech in Software Engineering from Nizam Institute of Engineering and Technology, Deshmukhi, Nalgonda Dist, Affiliated to JNTU Hyderabad, AP India. (Email: nasira_tabassum@gmail.com).

Syed.Asadullah Hussaini Master of Technology From Shadan College of Engineering & Technology, Affiliated to JNTU Hyderabad, AP India. His areas of interest include data mining, web technologies. (Email: asad_mtech@yahoo.com)