Comparative Study of Data Warehouse Architectural Design Approaches
Tanzeela Khanam¹ Pravin S. Metkewar²

¹²SICSR, affiliated to Symbiosis International University (SIU), Pune, Maharashtra, India

Abstract—The development of Data Warehouse starts with requirements gathering, designing the dimensional model which is further followed by testing and maintenance. It is the central repository where all the historical data is stored and maintained which is further used for the analysis. The most important activity in building of a data warehouse successfully is design phase. This work presents a brief description of different architectural approaches and techniques that address the DW Design problem. Different data models have been proposed for data warehouse design but these approaches are based on their own visual modeling languages UML or ER model, and there is no standard method or model that allows us to model all aspects of a DW. There are four stages of data migration in the proposed model: Data extraction, cleansing and refining data, data transforming, data indexing and loading which is further explained in the paper. Also this work presents the brief description of different approaches and techniques that addresses the DW Architecture. This paper shows the typical data warehouse architecture for different level: one, two, three level classical and novel three level architecture. The proposed architecture exhibits some drawbacks at the point when connected to work over extensive number of heterogeneous data sources.

Key words: Data warehouse design, Data warehouse Architecture, Multi-Dimensional Modeling, Unified Modeling Language

I. INTRODUCTION

The definition of Data warehouse starts with its characteristics: “subject oriented, time variant, integrated and non-volatile collection of large amount of data in support of decision making process within an organization”. Data warehouses store huge amount of data from disparate data sources which can be further used for query analysis and reporting. Therefore, the data is stored in the multi-dimensional (M D) structure which stores the information into facts and dimensions. A fact contains the metrics or measures (fact attributes) of business process (sales, deliveries, etc) where as dimension represents descriptive structure for analyzing a fact (product, customer, time etc). Database consists of following five phases first phase is Analysis of operational system whose aim is to collect the information with reference to existing operational system. It produces conceptual or logical schemes. Next phase consists of collecting and filtering the user requirements. It involves the users and designers of DW, and produces the specifications concerning choices of facts (metrics) and dimensions (descriptive attributes or fields). Conceptional design is the next phase where multidimensional schema of data warehouse is characterized and independent conceptual schema for the DW, according to the chosen conceptual model is produced. Logical design takes the input as conceptual schema and creates a corresponding logical schema. And finally, a phase of physical design, that checks the account issues particularly related to the tools chosen for implementation—such as indexing and allocation.

Despite the fact that few theoretical models have been proposed, none of them has been recognized as a standard one.

A. Database Design Methods:

<table>
<thead>
<tr>
<th>Step</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operational System Analysis</td>
<td>Information regarding the operational systems</td>
<td>Database Schemas</td>
</tr>
<tr>
<td>Conceptual design</td>
<td>Database Scheme</td>
<td>Specifications of Data warehouse</td>
</tr>
<tr>
<td>Logical design</td>
<td>Conceptual Schema</td>
<td>Logical Schema</td>
</tr>
<tr>
<td>Physical design</td>
<td>Logical schema</td>
<td>Physical Schema</td>
</tr>
</tbody>
</table>

Table 1:

II. PROBLEM DEFINITION

Different data models both conceptual and logical have been proposed for data warehouse design. These methodologies are based on their own visual modeling languages or well known graphical notation like ER model or UML, however to the best of our insight, there is no standard technique or model that permits us to model all aspects of DW. During our survey, the focus was on the development of MD data models and conceptual design of DW, the interest on the physical design of DWs has been very poor and most of the research was in designing and modeling of DWs. Moreover the databases are non-updateable and data keeps on changing so it is very difficult to design a model that is based on real time data.

The data models for outlining conventional OLTP systems are not appropriate for modeling complex queries in data warehousing environment. The OLTP transactions systems comprised of basic, pre-defined queries. The queries in data warehousing environment which tend to use joins on more tables, have a larger calculation time and are ad-hoc in nature. This sort of processing environment warrants a new perspective to data modeling.

The main issues in designing dimensional approach are:

- Loading data in the data warehouse from different operational systems is complicated in order to maintain the integrity of facts and dimensions.
- It is difficult to alter the data warehouse structure if the organization in which it does business changes the dimensional approach implementation.
- The architecture proposed so far exhibits some downsides when connected to work over substantial number of heterogeneous data sources. This paper addresses the issues of traditional architecture of data.
store and also it introduces the architecture for future generation typical data warehouse.

- The one-level architecture impels the planning activities like data source identification, data integration, transformation and so forth to be completed for each and every query which prompts to an unpredictable access time for the end customer.
- In two-level architecture each decision support system has its own derived data; there is no probability of storing a single copy of useful data for all decision support system, which illustrates data duplication.
- In three-level architecture the reconciled activity removes irregularities, anomalies and inconsistencies of operational data however again data replication found in this level.

III. LITERATURE REVIEW

UML design is preferred for designing the data warehouse, which characterized four UML profiles for modeling different aspects of data warehouse:

1) The UML profile for Multidimensional Modeling
2) The Data Mapping Profile
3) The ETL Profile
4) Database Deployment Profile

Authors propose a methodology that provides a hypothetical as well as theoretical foundation for implementing OO databases and Object relational databases in DW. This approach presents, arrangements of minimal constraints and extensions to UML for representing the properties of multidimensional modeling for DW. These all UML and conceptual designs are proposed on the Data warehouse architecture which is mentioned below:

In one-level architecture each piece of data is stored once. A middleware level works as an interface between the operational databases and users. This sort of architecture permits quick development of data warehouse with cost reduced.

In two-level architecture source data has been isolated from the derived data where the first level of this architecture stores source data and second level stores derived data. This architecture is suitable especially when operational sources are homogenous.

In three-level architecture consists of both reconciled and operational, derived data. There are two phases for data determination in decision support: the reconciliation of operational data belonging to disparate sources, derivation of decision support data from reconciled one. Data is stored in different level in this architecture: the first level stores operational data, reconciled data is stored in second level stores and the third level stores support decision data.
IV. TYPICAL DATA WAREHOUSE ARCHITECTURE

The components of typical data warehouse architecture are as follows:

1) The main component is software tools which is used for extracting data from multiple operational data sources and external data sources which is stored outside the organization for cleaning, transforming and integrating this data which is further used for loading data in data warehouse and periodically refreshing the data warehouse to reflect updates at the source.

2) It consists of several data marts.

3) OLAP server present multidimensional views of data to a variety of front end tools like report writer, analysis tools, query tool, and data mining tools.

4) Repository for storing and managing the meta Data.

5) Tools for monitoring and administrating the warehouse system.

V. DESIGN TECHNIQUES

Following are the approaches for architectural design:

<table>
<thead>
<tr>
<th>Sr.No</th>
<th>Proposal</th>
<th>Framework/Architecture</th>
<th>Approach/Technique</th>
<th>Scheme Used</th>
<th>Extended to logical level</th>
<th>Extended to Physical Level</th>
<th>Tool Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>-</td>
<td>Graphical Conceptual model for DW called Dimension Fact model</td>
<td>Methodology to build from pre-existing schema</td>
<td>Star</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
</tr>
<tr>
<td>2.</td>
<td>General framework for DW Design, based on dimensional-fact Model</td>
<td>Three-Level architecture</td>
<td>After collecting the user requirements and analyzing the existing Information system, a conceptual design is prepared from operational database scheme</td>
<td>Star</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>Object oriented approach to achieve the conceptual modeling of DW, MD, Databases and OLAP application</td>
<td>-</td>
<td>Introduce a set of constraints and extensions in UML for viewing MD modeling properties</td>
<td>Star</td>
<td>No</td>
<td>No</td>
<td>UML</td>
</tr>
<tr>
<td>4.</td>
<td>Manipulating and presenting MD models on the web by means of extensible style sheet language transformations(XSLT)</td>
<td>-</td>
<td>Object oriented approach Is used which is based on the UML</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>Modeling the physical design of a DW by using the component diagram and the deployment diagram of UML</td>
<td>Framework of a DW with 5 stages (i.e. source, integration, DW, customization and client) and 3 levels of design (conceptual, logical and physical).</td>
<td>It reduces the overall development time of DW and includes all main design phases of DW from conceptual modeling to final implementation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>Framework for the design of DW back stage including transformation rules at</td>
<td>Approach is based on the usage of UML packages</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td>Extend UML</td>
<td></td>
</tr>
</tbody>
</table>
VI. PROPOSED OBJECT ORIENTED FRAMEWORK FOR DATA WAREHOUSE CONCEPTUAL DESIGN

This paper gives the comparative study of different architectural design approaches used for data warehouse. Our framework considers the requirements of the users. There are two levels in the proposed framework: first is Requirements level and second is design level. At the requirement level, the requirements are gathered from the disparate users and a thorough analysis is made. The gathered requirements are integrated in the integrator component. There is quite a few numbers of components in each level to control particular tasks along with detailed meta data repository to speed up the whole process.

In next level, that is design level; UML designer helps in extracting major objects and classes from data gathered from multiple data sources and develop UML class diagrams. The UML class diagrams are then converted to multi dimensional model. This model is further represented in the form of star or snow flake schema. For effective conversion or transformations, certain mapping rules are applied that helps in mapping the classes to facts and dimension.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>7.</td>
<td>UML profile to represent MD and security aspects of conceptual modeling</td>
<td>Uses UML packages to group classes into higher level units</td>
<td>Star</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Conceptual UML model is designed and then it is translated into XML logical model which is later converted into XML document as physical model</td>
<td></td>
<td></td>
<td>UML</td>
</tr>
<tr>
<td>8.</td>
<td>UML profile for modeling DW usage on a conceptual level</td>
<td>An overview on UML based techniques and tools</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Conceptual level integration framework based on UML sources .First convert UML schema to UML class diagrams and then build multidimensional model from it</td>
<td>Object oriented approach for DW design. Mapping rules to convert/change UML class diagram to multidimensional model</td>
<td>Snowflake</td>
<td>UML</td>
</tr>
<tr>
<td>9.</td>
<td>Uses UML to build a DW model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Table 2:</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 4:**

All rights reserved by www.ijsrd.com 535
VII. RESULTS
Different Object oriented frameworks are proposed by different authors for data warehouse conceptual design which has many benefits. Firstly, the best approach for architectural designing is the object oriented multi dimensional approach because it fulfills all the criteria required for the data warehouse architectural design and its more versatile (adaptable) as the user requirements keeps on changing constantly. Secondly, UML is used which is easy to learn and access. Further it can model all real world objects. Thirdly, the more efficient data warehouse design is star and snowflake schemas are as they are easy to learn and need fewer joins. Different Architecture are proposed for example one level, two level and three level to facilitate different aspects of data warehouse design.

VIII. CONCLUSION
In this paper, we make a comparative study of different design and architectural approaches used for data warehouse. In the literature survey, we have seen different authors have proposed different techniques at different levels of designing i.e. conceptual, logical and physical level. Our study is based on following criteria: Proposal, Framework/Architecture, Approach or technique or methodology proposed, Schema used and whether the design can be extended to Logical and physical design also.

We have seen the limitations in two level and typical three-level architecture with heterogeneous operational source systems. The Meta data are derived and updated separately from operational and derived data; therefore they are not well related to corresponding schemes. Therefore novel three-tier architecture is proposed in this paper which eliminates all the drawbacks of the classical or traditional one which handles the structured data.

IX. FUTURE SCOPE
In future, three-tier architecture can be extended to deal with big data and perform analysis on it which can further process structured as well as unstructured data. In such case a distributed file system like Hadoop acts as link between source data systems and data warehouse. It will collect the data, perform aggregation, and processes huge volumes of unstructured data and stages it for loading into data warehouse. This architecture will produce a semi automatic data warehouse. Further we are in process of testing the projected framework and implementing the same using JAVA at the front and Oracle 10g at the back end.

REFERENCES
[6] M.Breslin, Data warehousing battle of giants: comparing the basics of the Kimball and Inmon models, Business Intelligence Journal9,2004