DR B.R. AMBEDKAR NATIONAL INSTITUTE OF TECHNOLOGY JALANDHAR

Assignment – 1

Data Mining and Data Warehousing

SUBMITTED TO-
Dr. Geeta Sikka
CSE Department

SUBMITTED BY-
Ankit Goyal (17103011)
Group - G-1
Branch - CSE

The Study on Data Warehouse Design and Usage

I. INTRODUCTION

Before we see the design process, let's see what a data warehouse is. Think of a data warehouse as a central storage facility which collects information from many sources, manages it for efficient storage and retrieval, and delivers it to many audiences, usually to meet decision support and business intelligence requirements. "What is the need of a data warehouse? What goes into a data warehouse design? How are data warehouses used? How do data warehousing and OLAP relate to data mining?" In this research paper we are discussing a business analysis framework for data warehouse design, data warehouse design process, data warehouse usage for information processing and from OLAP to multidimensional data mining. The concept of data warehousing is deceptively simple. Data is extracted periodically from the applications that support business processes and copied onto special dedicated computers. There it can be validated, reformatted, reorganized, summarized, restructured, and supplemented with data from other sources. The resulting data warehouse becomes the main source of information for report generation, analysis, and presentation through ad hoc reports, portals, and dashboards. Building data warehouses used to be difficult. Many early adopters found it to be costly, time consuming, and resource intensive. Over the years, it has earned a reputation for being risky. This is especially true for those who have tried to build data warehouses themselves without the help of real experts.

II. RESEARCH ELABORATIONS

[A] A BUSINESS ANALYSIS FRAMEWORK FOR DATA WAREHOUSE DESIGN:

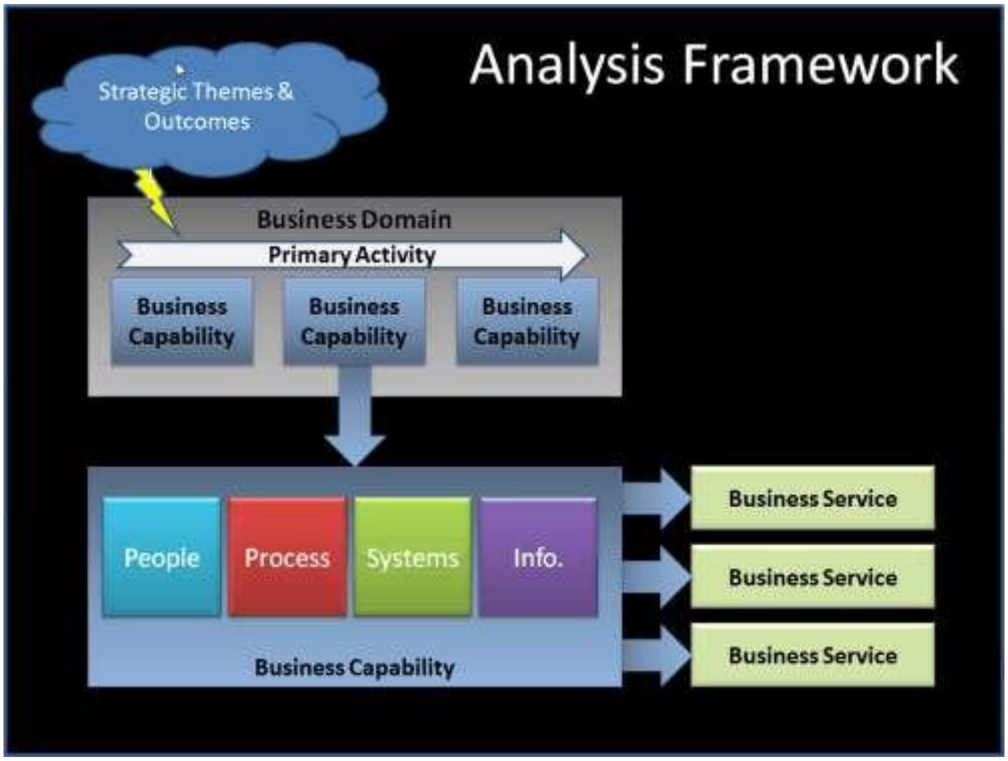
This view is combined to form a complex framework that represents the top-down, business-driven, or owner's perspective, as well as the bottom-up, builder-driven, or implementor view of the information system. Four, different views regarding a data warehouse design must be considered: the top-down view, the data source view, the data warehouse view, and the information system.

- The top - Down view allows the selection of the relevant information necessary for the data warehouse. This information matches current and future business needs.
- The Data source view exposes the information being captured, stored, and managed by the operating system. This information may be documented at various levels of detail and accuracy, from individual data source tables to integrate at various levels of detail and accuracy, form individual data source tables to integrated data source tables. Data sources are often modeled by traditional data modeling techniques, such as the E-R model or DASE tools.
- The Data warehouse view includes fact tables and dimension tables. It represents the information that is stored inside the data warehouse, including precalculated totals and

counts, as well as information regarding the source, date and time of origin added to provide historical context.

- The Business Query View is the data perspective in the data warehouse form the end-user's view point

So, building and using a data warehouse is a complex task because it requires business skill technology skills, and program management skills. Regarding business skills, building a data warehouse involves understanding how systems store and manage their data, how to build extractors that transfer data from the operational system to the data warehouse, and how to build warehouse refresh software that keeps the data warehouse reasonably up-to-date with the operational system's data.



A framework provides the structure for the upsurge of using analysts of all sorts: business analysts, business process analysts, risk analysts, system analysts, and provides a standardized way to gather, communicate and develop the desired information required by:

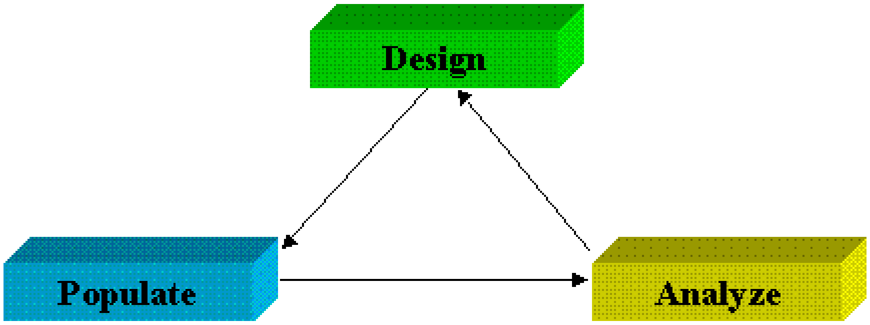
- The Program Management Office;
- Business users;
- key stakeholders and
- Technology developers.

Based on our experience, even for projects that are completed on time and on budget, there may be significant inefficiencies in performing business analysis functions. These inefficiencies include:

- lost opportunities

- Rework
- No realization of benefits

[B] DATA WAREHOUSE DESIGN PROCESS



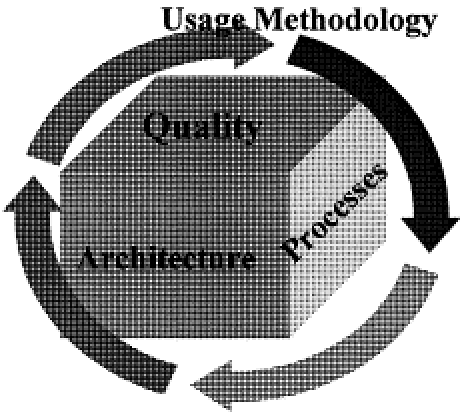
This is always considered as a good choice for data warehouse development, especially for data marts, because the turnaround time is short, modifications can be done quickly, and new designs for the technologies and that can be adapted in a timely manner. So, here we are discussing the warehouse design process. This includes various steps as follows:

- Choose a Business Process to Model if the business process is organizational and involves multiple complex object collections, a data warehouse model should be followed. However, if the process is departmental and focuses on the analysis of one kind of business process, a data mart model should be chosen.
- Choose the business process gain, which is the fundamental, atomic level of data to be represented in the fact table for this process.
- Choose the dimension that will apply to each and every fact table record. Typical dimensions are time, item, customer, supplier, warehouse, transactions type, and status.
- Choose the measures that will populate each fact table record. Typical measures are numeric additive quantities like dollars_sold and units_sold.

Because the process of construction of a data warehouse is a quite difficult and long-term task, its implementation scope should be clearly defined. The goals of a fundamental data warehouse implementation should be specific, achievable and measurable. This involves determining the time and budget allocations, the subset of the organization that is to be served. So, once a data warehouse is designed and constructed, the fundamental deployment of the warehouse includes the initial installations, roll – out planning, training, and orientations. And platform upgrades and maintenance must also be considered. So, the data warehouse administration includes data refreshment, data source synchronization, planning for disaster recovery, managing access control and security, managing data growth, managing database performances and of course data warehouse enhancement and extension.

Data warehouse development tools provide functions to define and edit metadata repository contents (i.e. schemas, scripts, or rules), answer queries, output reports, and ship metadata to and from relational database system catalogs. Planning and analysis tools study the impact of schema changes and of refresh performance when changing refresh rates or time windows.

[C] DATA WAREHOUSE USAGE FOR INFORMATION PROCESSING



There are a total three kinds of data warehousing applications: Information processing, Analytical processing, and data mining.

- Information Processing supports querying, basic statistical querying, basic statistical analysis, and reporting using cross tabs, tables, charts or graphs. A current trend in data warehouse information processing is to construct low-cost web-based accessing tools that are then integrated with web browsers.
- Analytical Processing supports basic OLAP operations, including slice-and-dice, drill-down, roll-up, and pivoting. It generally operates on historic data in both summarized and detailed forms. The major strength of online analytical processing over information processing is the multidimensional data analysis of data warehouse data.
- Data Mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualizations tools.

[D] FROM ONLINE ANALYTICAL PROCESSING TO MULTIDIMENSIONAL DATA MINING

Among the many different paradigms and architectures of data mining systems, multidimensional data mining is particularly important for the various reasons which are as follows:

- High Quality of data in data warehouse: Most data mining tools need to work on integrated, consistent, and cleansed data, which requires costly data cleaning, data integration and data transformation as preprocessing steps. A data warehouse constructed by such preprocessing steps. While a data warehousing constructed by such preprocessing serves as a valuable source of high-quality data for OLAP as well as for data mining. Now, we notice that data mining may serve as a valuable tool for data cleaning and data integration as well.

- Available information processing infrastructure surrounding data warehouses: Comprehensive information processing and data analysis infrastructures have been or will be systematically constructed surrounding data warehouses, which includes the accessing, integration, consolidation and transformation of multiple heterogeneous databases and OLAP analytical tools. It is prudent to make best use of the available infrastructure rather than constructing everything from scratch.
- OLAP-Based exploration of multidimensional data: Effective data mining needs exploratory data analysis. A user will often want to traverse through a database, select portions of relevant data, analyze them at different granularities, and present knowledge in different forms. Multidimensional data mining provides facilities of pivoting filtering, dicing, and slicing on a data cube and intermediate data mining results.
- Online Selection of data mining functions: Users may not always know the specific kinds of knowledge they want to mine. By integrating OLAP with various data mining functions, multidimensional data mining provides users with the flexibility to select desired data mining functions and swap data mining tasks dynamically.

III. CONCLUSION

Creating and managing a warehousing system is hard. Many different classes of tools are available to facilitate different aspects of the process described in Section 2. 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.

REFERENCES

[1] Inmon, W.H., Building the Data Warehouse. John Wiley, 1992.
[2] http://www.olapcouncil.org
[3] Codd, E.F., S.B. Codd, C.T. Salley, “Providing OLAP (On-Line Analytical Processing) to User Analyst: An IT Mandate.”