International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 1 | ISSN : 2456-3307

A Survey on Data Warehousing

Sapna Mittal
Assistant Professor, Computer Science, Post Graduate Department of Computer Science and Applications, R.S.D College, Ferozepur, Punjab, India

ABSTRACT

Data warehousing is a phenomenon that grew from the huge amount of electronic data stored in recent years and from the urgent need to use that data to accomplish goals that go beyond the routine tasks linked to daily processing. In a typical scenario, a large corporation has many branches, and senior managers need to quantify and evaluate how each branch contributes to the global business performance. The corporate database stores detailed data on the tasks performed by branches. To meet the managers' needs, tailor-made queries can be issued to retrieve the required data. In order for this process to work, database administrators must first formulate the desired query (typically an aggregate SQL query) after closely studying database catalogs. Then the query is processed. This can take a few hours because of the huge amount of data, the query complexity, and the concurrent effects of other regular workload queries on data. Finally, a report is generated and passed to senior managers in the form of a spreadsheet.
Keywords : Data Warehousing, SQL query, Craftsmanship, Warehouse Architectures, Health care service, Telecommunication services

I. INTRODUCTION

Data Warehouse Design

Let's review some fields of application for which data warehouse technologies are successfully used:

- **Trade** Sales and claims analyses, shipment and inventory control, customer care and public relations
- **Craftsmanship** Production cost control, supplier and order support
- **Financial services** Risk analysis and credit cards, fraud detection
- **Transport industry** Vehicle management
- **Telecommunication services** Call flow analysis and customer profile analysis
- **Health care service** Patient admission and discharge analysis and bookkeeping in accounts departments

The field of application of data warehouse systems is not only restricted to enterprises, but it also ranges from epidemiology to demography, from natural science to education. A property that is common to all fields is the need for storage and query tools to retrieve information summaries easily and quickly from the huge amount of data stored in databases or made available by the Internet. This kind of information allows us to study business phenomena, learn about meaningful correlations, and gain useful knowledge to support decision-making processes at a Warehouse.

Data Warehouse Architectures

The following architecture properties are essential for a data warehouse system (Kelly, 1997):

- **Separation :** Analytical and transactional processing should be kept apart as much as possible.
- **Scalability:** Hardware and software architectures should be easy to upgrade as the data volume, which has to be managed and processed, and the number of users' requirements, which have to be met, progressively increase.
- **Extensibility:** The architecture should be able to host new applications and technologies without redesigning the whole system.

• **Security** Monitoring accesses is essential because of the strategic data stored in data warehouses.
• **Administer ability:** Data warehouse management should not be overly difficult.
Two different classifications are commonly adopted for data warehouse architectures.

II. Architecture

Single-Layer Architecture

A single-layer architecture is not frequently used in practice. Its goal is to minimize the amount of data stored; to reach this goal, it removes data redundancies.layer physically available: the source layer. In this case, data warehouses are virtual.

Two-Layer Architecture

The requirement for separation plays a fundamental role in defining the typical architecture for a data warehouse system. Although it is typically called a two-layer architecture to highlight a separation between physically available sources and data warehouses, it actually consists of four subsequent data flow stages:
1. Source layer A data warehouse system uses heterogeneous sources of data. That data is originally stored to corporate relational databases or legacy1 databases, or it may come from information systems outside the corporate walls.
2. Data staging The data stored to sources should be extracted, cleansed to remove inconsistencies and fill gaps, and integrated to merge heterogeneous sources into one common schema. The so-called Extraction, Transformation, and Loading tools (ETL) can merge heterogeneous schemata, extract, transform, cleanse, validate, filter, and load source data into a data warehouse (Jarke et al., 2000). Technologically speaking, this stage deals with problems that are typical for distributed information systems, such as inconsistent data management and incompatible data structures (Zhuge et al., 1996). Section 1.4 deals with a few points that are relevant to data staging.

3. Data warehouse layer Information is stored to one logically centralized single repository: a data warehouse. The data warehouse can be directly accessed, but it can also be used as a source for creating data marts, which partially replicate data warehouse contents and are designed for specific enterprise departments. Meta-data repositories (section 1.6) store information on sources, access procedures, data staging, users, data mart schemata, and so on.

4. Analysis In this layer, integrated data is efficiently and flexibly accessed to issue reports, dynamically analyze information, and simulate hypothetical business scenarios. Technologically speaking, it should feature aggregate data navigators, complex query optimizers, and user-friendly GUIs. deals with different types of decision-making support analyses.
The architectural difference between data warehouses and data marts needs to be studied closer. The component marked as a data warehouse in Figure 1-3 is also often called the primary data warehouse or corporate data warehouse. It acts as a centralized storage system for
1The term

Three-Layer Architecture

In this architecture, the third layer is the reconciled data layer or operational data store. This layer materializes operational data obtained after integrating and cleansing source data. As a result, those data are integrated, consistent, correct, current, and data warehouse that is not populated from its sources directly, but from reconciled data.

The main advantage of the reconciled data layer is that it creates a common reference data model for a whole enterprise. At the same time, it sharply separates the problems of source data extraction and integration from those of data warehouse population. Remarkably, in some cases, the reconciled layer is also directly used to better accomplish some operational tasks, such as producing daily reports

that cannot be satisfactorily prepared using the corporate applications, or generating data flows to feed external processes periodically so as to benefit from cleaning and integration. However, reconciled data leads to more redundancy of operational source data. Note that we may assume that even two-layer architectures can have a reconciled layer that is not specifically materialized, but only virtual, because it is defined as a consistent integrated view of operational source data. Finally, let's consider a supplementary architectural approach, which provides a comprehensive picture. This approach can be described as a hybrid solution between the single-layer architecture and the two/three-layer architecture. This approach assumes that although a data warehouse is available, it is unable to solve all the queries formulated. This means that users may be interested in directly accessing source data from aggregate data (drill-through). To reach this goal, some queries have to be rewritten on the basis of source data (or reconciled data if it is available). This type of architecture is implemented in a prototype and it needs to be able to go dynamically back to the source data required for queries to be solved (lineage).

III. REFERENCES

[1]. Dedic, N. and Stanier C., 2016., "An Evaluation of the Challenges of Multilingualism in Data Warehouse Development" in 18th International Conference on Enterprise Information Systems - ICEIS 2016, p. 196.

[2]. Jump up to:a b "9 Reasons Data Warehouse Projects Fail". blog.rjmetrics.com. Retrieved 2017-04-30.

[3]. Jump "Exploring Data Warehouses and Data Quality". spotlessdata.com. Retrieved 2017-04-30.

[4]. Jump "What is Big Data". spotlessdata.com. Retrieved 2017-04-30.

[5]. Jump Patil, Preeti S.; Srikantha Rao; Suryakant B. Patil (2011). "Optimization of Data Warehousing System: Simplification in Reporting and Analysis"