ISSN XXXX XXXX © 2019 IJESC

Research Article Volume 9 Issue No.3

Data Warehousing

Anshu
Department of Computer Science
G.V.M. Girls College, Haryana, India

Abstract:
Data ware housing is a booming industry with many interesting research problem. A data warehouse is a global repository that stores pre-processed queries on data which resides in multiple, possibly heterogeneous, operational or legacy sources. The information stored in the data warehouse can be easily and efficiently accessed for making effective decisions. The On-Line Analytical Processing (OLAP) tools access data from the data warehouse for complex data analysis, such as multidimensional data analysis, and decision support activities. Current research has lead to new developments in all aspects of data warehousing, however, there are still a number of problems that need to be solved for making data warehousing effective. The data warehouse is concentrated on only few aspects. Let's look at various principles of good data warehouse design and the steps involved. Data warehouse can be built using a top-down approach, bottom – down approach or a combination of both.

Keywords: Data Warehouse, ETL, OLAP

1. INTRODUCTION

Data warehouse is a subject oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process. It is a technology that aggregates structured data from one or more sources so that it can be compared and analyzed for greater business intelligence. It is used to provide greater insight in to the performance of a company by comparing data consolidated from multiple heterogeneous sources. A data warehouse is designed to run query and analysis on historical data derived from transactional sources. It is a useful tool ,gives benefit from the ability to store and analyze data and this can allow in making sound decisions. It is also important to make sure that the correct information is published and it should be easy to access by the people who are responsible for making decisions. There are two elements that make up the data warehouse environment, and these are presentation and staging. The staging could also be known as the acquisition area. It is composed of ETL operations, and once the data has been prepared, it will be sent to the presentation area. Once the data has been incorporated in to warehouse, it does not change and cannot be altered since a data warehouse runs analytics on events that have already occurred by focusing on the changes in data over time. So, warehoused data must be stored in a manner that is secure, reliable, easy to retrieve and easy to manage. To build an effective data warehouse, it is important to understand the data warehouse principles. If the data warehouse is not built correctly, it runs in to a number of different problems.

2. BENEFITS OF DATA WAREHOUSING

1. The Enablement of Better Decision-Making

As companies are now able to get closer to their consumers than ever before, the corporate decision-makers no longer have to make important business decisions based on partial or limited data. They're now backed up by facts and statistics housed within data warehouses that can be recalled ad hoc.

2. Quick and Easy Data Access

Users can access an array of information, stored across multiple sources, almost instantly. It means you won't be wasting time attempting to manually pull information from various sources.

3. Consistent Quality Data

As Data warehouses gather information from countless sources, but they convert it into a unified format to be used throughout the organization. As a result, each department will be producing results which are consistent with each other, which in turn ensures organization wide accuracy.

3. PRINCIPLES OF DESIGNING GOOD DATA WAREHOUSE

As there is a need for data warehouses to be designed, implemented as well as properly maintained. So, the following principles of effective data warehouse design should be followed. They consist of both business as well as IT principles.

Business principles

Organization-wide consensus

At the very onset of deploying data warehousing, there is the need for a consensus-building process, which assists in guiding the planning process as well as the design plus implementation process. So if our workers as well as managers see data warehousing as unnecessary they will not use it as they ought to or see it as a threat to their jobs and so will not use it at all. Therefore, make early effort to include all stakeholders and gain their acceptance regarding data warehousing before it is implemented.

Data integrity

Data integrity is crucial to data warehousing. Consequently, any design should start by reducing the possibility of data replication as well as inconsistency. Also, data integration plus standardization should equally be promoted.

Implementation efficiency
Data Warehouse design ought to be straightforward plus efficient to carry out. In other words, designing a technically robust data warehouse without giving any consideration to the difficulty or implementation of such design is counter-productive. Instead, we should go for simplicity.

Operational efficiency
Data Warehouses ought to be easy to support. They should equally ensure rapid responses as far as business change requests are concerned. In addition to this, it should also be easy to correct errors and even exceptions.

User friendliness
Warehouse design should be easy to use because if the people who make use of data warehouse find it difficult to use, then there is a business end-user problem.

IT principles

IT standard compliance
Compliance to information technology standards is probably the most important of the IT principles. Therefore, conforming to existing information technology standards to ensure that tool-sets plus platforms chosen to implement data warehouse agree with this standard is key.

Scalability
This is usually a huge problem with data warehouse design. In order to resolve this, from the very onset scalability should be factored in. Consequently, platforms and tool-sets, which support data volume expansions in the future, should be chosen.

4. DATA WAREHOUSE DESIGN

Data Warehouse design is the process of building a solution to integrate data from multiple sources that support analytical reporting and data analysis. A poorly designed data warehouse can result in acquiring and using inaccurate source data that negatively affect the productivity and growth of your organization. This blog post will take a high-level look at the data warehouse design process from requirements gathering to implementation.

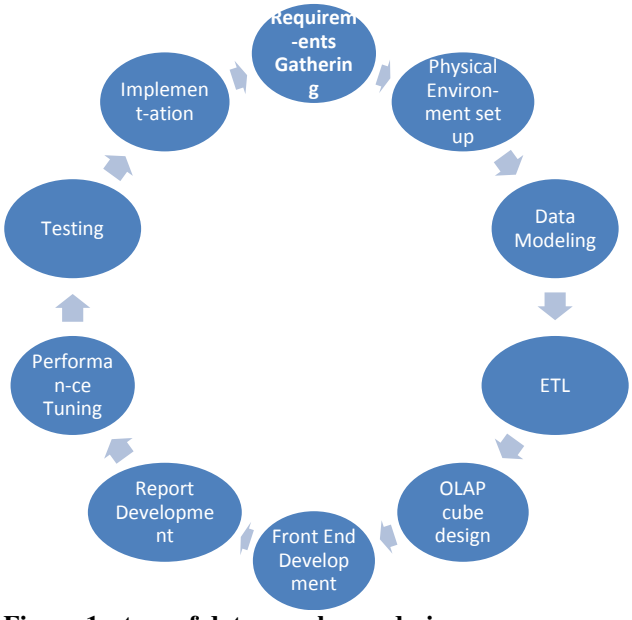

Figure.1. steps of data warehouse design

Requirements Gathering
Gathering requirements is first step of the data warehouse design process. The goal of the requirements gathering phase is to determine the criteria for a successful implementation of the data warehouse. User analysis and reporting requirements must be identified as well as hardware, development, testing, implementation, and user training.

Physical Environment Setup
Once the business requirements are set, the next step is to determine the physical environment for the data warehouse. At a minimum, there should be separate physical application and database servers as well as separate ETL/ELT, OLAP, cube, and reporting processes set up for development, testing, and production. Building separate physical environments ensure that all changes can be tested before moving them to production, development, and testing can occur without halting the production environment, and if data integrity becomes suspect, the IT staff can investigate the issue without negatively impacting the production environment.

Data Modelling
Once requirements gathering and physical environments have been defined, the next step is to define how data structures will be accessed, connected, processed, and stored in the data warehouse. This process is known as data modelling. During this phase of data warehouse design, is where data sources are identified Once the data sources have been identified, the data warehouse team can begin building the logical and physical structures based on established requirements.

ETL
The ETL process takes the most time to develop and eats up the majority of implementation. Identifying data sources during the data modelling phase may help to reduce ETL development time. The goal of ETL is to provide optimized load speeds without sacrificing quality. Failure at this stage of the process can lead to poor performance of the ETL process and the entire data warehouse system.

OLAP Cube Design
On-Line Analytical Processing (OLAP) is the answer engine that provides the infrastructure for ad-hoc user query and multi-dimensional analysis. OLAP design specification should come from those who will query the data. Documentation specifying the OLAP cube dimensions and measures should be obtained during the beginning of data warehouse design process. The three critical elements of OLAP design include:

- **Grouping measures** - numerical values you want to analyze such as revenue, number of customers, how many products customers purchase, or average purchase amount.
- **Dimension** - where measures are stored for analysis such as geographic region, month, or quarter.
- **Granularity** - the lowest level of detail that you want to include in the OLAP dataset.

During development, make sure the OLAP cube process is optimized. A data warehouse is usually not a nightly priority run, and once the data warehouse has been updated, there little time left to update the OLAP cube. Not updating either of them in a timely manner could lead to reduced system performance.

Taking the time to explore the most efficient OLAP cube generation path can reduce or prevent performance problems after the data warehouse goes live.

Front End Development
At this point, business requirements have been captured, physical environment complete, data model decided, and ETL process has been documented. The next step is to work on how users will access the data warehouse. Front end development is how users will access the data for analysis and run reports. There are many options available, including building your front end in-house or purchasing an off the shelf product. Either way, there are a few considerations to keep in mind to ensure the best experience for end users. Secure access to the data from any device - desktop, laptop, tablet, or phone should be the primary consideration. The tool should allow your development team to modify the backend structure as enterprise level reporting requirements change. It should also provide a Graphical User Interface (GUI) that enables users to customize their reports as needed. The OLAP engine and data can be the best in class, but if users are not able to use the data, the data warehouse becomes an expensive and useless data repository.

Report Development
For most end users, the only contact they have with the data warehouse is through the reports they generate. As mentioned in the front end development section, users' ability to select their report criteria quickly and efficiently is an essential feature for data warehouse report generation. Delivery options are another consideration. Along with receiving reports through a secure web interface, users may want or need reports sent as an email attachment, or spreadsheet. Controlling the flow and visibility of data is another aspect of report development that must be addressed. Developing user groups with access to specific data segments should provide data security and control. Reporting will and should change well after the initial implementation. A well-designed data warehouse should be able to handle the new reporting requests with little to no data warehouse system modification.

Performance Tuning
As we create separate development and testing environments, doing so allow organizations to provide system performance tuning on ETL, query processing, and report delivery without interrupting the current production environment.

Testing
Once the data warehouse system has been developed according to business requirements, the next step is to test it. Testing, or quality assurance will allow the data warehouse team to expose and address issues before the initial rollout.

Implementation
Deciding to make the system available to everyone at once or perform a staggered release, will depend on the number of end users and how they will access the data warehouse system. Another important aspect of any system implementation is end-user training.

5. CONCLUSION

A data warehouse is a relational data base that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data but it can include data from other sources. Transaction workload is separated from analysis workload and enables organization to consolidated data from several sources. In addition to relational database a data warehouse environmental include an extraction transportation transformation and loading solution (ETL), an online analytical processing (OLAP) engine analyst is tool for client and other applications that manage the process or gathering data & developing it to business users.

6. REFERENCES

[1]. Inmon, W.H., Building the Data Warehouse. John Wiley, 1992.

[2]. Codd, E.F., S.B. Codd, C.T. Salley, "Providing OLAP (On-Line Analytical Processing) to User Analyst: An IT Mandate."

[3].Kimball, R. The Data Warehouse Toolkit. John Wiley, 1996.

[4]. Barclay, T., R. Barnes, J. Gray, P. Sundaresan, "Loading Databases using Dataflow Parallelism." SIGMOD Record, Vol. 23, No. 4, Dec.1994.

[5]. Blakeley, J.A., N. Coburn, P. Larson. "Updating Derived Relations: Detecting Irrelevant and Autonomously Computable Updates." ACM TODS, Vol.4, No. 3, 1989.

[6].Gupta, A., I.S. Mumick, "Maintenance of Materialized Views: Problems, Techniques, and Applications." Data Eng. Bulletin, Vol. 18, No. 2, June 1995.

[7].9 Zhuge, Y., H. Garcia-Molina, J. Hammer, J. Widom, "View Maintenance in a Warehousing Environment, Proc. Of SIGMOD Conf., 1995.

[8]. Roussopoulos, N., et al., "The Maryland ADMS Project: Views R Us." Data Eng. Bulletin, Vol. 18, No.2, June 1995.

[9]. O'Neil P., Graefe G. "Multi-Table Joins through Bitmapped Join Indices" SIGMOD Record, Sep 1995.

[10]. Harinarayan.V., Rajaraman.A., Ullman.J.D. " Implemen ting Data Cubes Efficiently" Proc. of SIGMOD Conf., 1996.