CHAPTER 4

Data Warehouse Architecture

4. Data Warehouse Architecture

4.1 Steps for the design and construction of warehouse:

A data warehouse provides competitive advantage by presenting relevant information from which to measure performance and make critical judgments in order to win in the competitive market space. A data warehouse can enhance productivity since it can quickly gather information about the activities of the organization. Thirdly, a data warehouse provides enhanced customer relationship management since it provides consistent view of the customers and items along all lines of the business, all departments and all markets. Finally, it is possible that a data warehouse brings about cost reduction by tracking trends, patterns and exceptions over long periods of time in a reliable and a consistent manner.

The business analysis framework must be understood in order to create an efficient data warehouse. Four different views of the data warehouse must be consisted: the top-down view, the data source view, the data warehouse view and the business query view.

- **The top-down view** allows the selection of relevant information necessary for the data warehouse. The information matches the current and the forthcoming needs of the business.
- **The data source view** exposes the way in which the data is captured, stored and managed by a data warehouse system. For individual source tables to integrated source tables information may be detailed at various levels of accuracy. Normally CASE tools or entity relationship models hold the traditional data sources.
- **The data warehouse view** includes the fact tables and dimension tables. It represents the information that is stored with in the data warehouse as well as information regarding the data, time of origin of the source data to provide the historical context.

- Finally, the **business query view** is the perspective of data in the data warehouse from the point of view of the end user.

A variety of skills are required for building a data warehouse as it is a complex task requiring business skills, technology skills and program management skills .

A data warehouse may be built using a top-down approach or a bottom up approach or a combination of both. The top down approach overall plans are made and it is usually works well where the technology is well known and mature. The bottom up approach starts with experiments and prototypes –this is usually successful in early stages of business modeling. In the combined approach both the planned, strategic nature of the top down approach and the opportunistic application of the bottom up approach provides the organization advantage.

In general the warehouse designs consist of the following steps:

1. Choose the **business process** to model, for example orders, invoices, shipments etc. If the business process is multiple and involves many parts of the organization a data warehouse model should be followed. On the other hand if the business process involves only one kind of process a data mart should be made.

2. Choose the **grain** of the business process. The grain is the fundamental atomic level at which a fact in a fact table is to be represented.

3. Choose the **dimensions** that will apply to each fact table record. Typical dimensions are time, item, customer, supplier, warehouse transaction type and status.
4. Choose the **measures** that will populate each fact table record. Typical measures are numeric additive quantities like dollars sold and units sold.

Since data warehouse construction is a difficult and a long term task, its implementation scope should be clearly defined in the beginning. The goals of an initial data warehouse should be specific, achievable and measurable

4.2 Three-tier data warehouse architecture

Data warehouses normally adopt three-tier architecture:

1. The bottom tiers is a warehouse database server that is almost always a relational database stsyem.Data from operational databases and from external sources are extracted using application program interfaces known as **gateways**. A gateway is supported by the underlying DBMS and allows client programs to execute code.

2. The middle tier is an **OLAP server** that is typically implemented using a relational OLAP (ROLAP) model.

3. The top tier is a **client**, which contains query and and reporting tools, analysis tools and/or data mining tools.

From the architecture point of view there are three data warehouse models: the enterprise warehouse, the data mart, and the virtual warehouse.

1. **Enterprise Warehouse:** An enterprise warehouse collects all details comprising of all information about subjects spanning the entire organization. It provides corporate wide data

integration, usually from one or more operational systems and from external information providers. It takes extensive business modeling and it takes many years to design and build.

2. **Data Mart:** A data mart consists of a subset of corporate wide data that is of value to specific group of users. The scope is confined to specific selected subjects. The data contained in a data mart tend to be summarized.

3. **Virtual Warehouse:** A virtual warehouse is a set of views over operational databases. For efficient query processing, only some of the possible summary views may be materialized. A virtual warehouse is easy to build and it requires excess capacity on the operational database servers.

4.3 Types of OLAP servers: ROLAP versus MOLAP versus HOLAP

Normally business users are presented with multidimensional data from data warehouses or data marts without them being aware of the way in which i. e. how or where the data are stored. However, the physical architecture and implementation of OLAP servers need to take into consideration the issues regarding data storage. Various implementations are possible:

1. **Relational OLAP(ROLAP)servers:** These are intermediate servers that stand in between a relational back-end server and client front-end tools. They use a relational or an extended relational DBMS to store and manage warehouse data, OLAP middle ware to support missing pieces.

2. **Multidimensional OLAP (MOLAP) servers:** These servers support multidimensional view of data through array based multi dimensional storage engines. They map multidimensional views directly to data cube array structures.

3. **Hybrid OLAP (HOLAP) servers:** The hybrid OLAP approach combines ROLAP and MOLAP technologies providing greater scalability and faster computation of MOLAP. For example, a HOLAP server may allow large volumes of data to be kept in a relational database and aggregations to be kept in separate MOLAP store.

Data warehouse Implementation

Data warehouses contain huge volumes of data. OLAP servers demand that decision support queries be answered in the order of seconds.

Efficient computation of data cubes

Efficient computations of aggregation across many sets of dimensions remain at the core of multidimensional data analysis. In SQL terms these aggregations are referred to as **"group-by's".**

The compute cube operator and Its Implementation

One approach to cube computation is to include compute cube operator. The compute cube operator computes aggregates over all subsets of the dimensions specified in the operation.

4.4 Further development of Data Cube Technology

Discovery driven exploration of data cubes

Data can be summarized and stored in a variety of ways in a multidimensional cube of an OLAP system. A user or analyst can search for interesting patterns in a cube by specifying a number of OLAP operations, such as drill down, roll up, slice, and dice. While tools are there to help the discovery process is not automated. The user follows his or her own intuition or hypotheses, tries to recognize exceptions or anomalies in the data. Discovery driven exploration is an alternative approach in which pre-computed measures indicating data exceptions are used to guide the user in the data analysis process. Exception indicators indicate cell values that are significantly different from the anticipated structural model. Three measures are used as exception indicators to help identify data anomalies. These measures indicate the degree of surprise that the quantity in a cell holds, with respect to its expected value. The3 measures are computed and associated with every cell, for all levels of aggregation. They are

SelfExp: This indicates the degree of surprise of the cell value, relative to other cells t the same level of aggregation.

InExp: This indicates the degree of surprise somewhere beneath the cell, if we were to drill down from it.

PathExp: This indicates the degree of surprise for each drill-down path from the cell.

4.5 Complex Aggregation at Multiple Granularity: Multi-feature cubes

Data cubes facilitate the answering of data mining queries as they allow the computation of aggregate data at multiple levels of granularity. Multifeature cubes compute complex queries involving multiple dependent aggregates at multiple granularities. These cubes are very useful in practice. Many complex data mining queries can be answered be answered by multifeature cubes without any significant increase in computing cost, in comparison to cube computation for simple queries with standard data cubes.

From Data Warehousing to Data Mining

In this section we study the usage of data warehousing for information processing, analytical processing and data mining. We also introduce on-line analytical mining (OLAM), a powerful paradigm that integrates OLAP with data mining technology.

Data warehouses and data marts and data marts are used in a wide range of applications. Business executives in almost every industry use the data collected, integrated, preprocessed and stored in data warehouses. Data warehouses are used extensively in banking and financial services, consumer goods and retail distribution sectors and controlled manufacturing, such as demand based production.

The more a data warehouse has been in use the more it would have evolved. This evolution takes place through a number of stages. Initially data warehouses are used for generating reports and answering predefined queries. Progressively, it used to analyze summarized and detailed data. Later, the data warehouses are put to strategic use. Finally the data warehouse may be put to use for strategic decision-making and knowledge discovery using data mining tools.

Business users need to know what exists in the data warehouse (through metadata), how to access the contents of the data warehouse, how to examine the contents using analytical tools and how to present the results of such an analysis. There are three kinds of data warehouse applications:

- **Information Processing** supports querying, basic statistical analysis, and reporting using cross tabs, tables charts and graphs current trend in information processing is to construct low-cost Web based accessing tools that are integrated with the web browsers.
- **Analytical Processing** supports basic OLAP operations, including slice and dice, drill-down, roll up and pivoting. It generally operates on historical data in both summarized and detailed forms.

The major strength of on-line analytical processing over information processing is the multidimensional analysis of data in a data warehouse.

- **Data Mining** supports knowledge discovery by finding hidden patterns and associations constructing analytical models, performing classifications and predictions, and finally presenting the mining results using visualization tools.

4.6 Review questions

1 Give theData Warehouse Architecture

2Expalin aboutThree-tier data warehouse architecture

3 List theTypes of OLAP servers: ROLAP versus MOLAP versus HOLAP

4.Explain about Complex aggregation?

4.7 References

[1]. Data Mining Techniques, Arun k pujari 1st Edition

[2] .Data warehousung,Data Mining and OLAP, Alex Berson ,smith.j. Stephen

[3].Data Mining Concepts and Techniques ,Jiawei Han and MichelineKamber

[4]Data Mining Introductory and Advanced topics, Margaret H Dunham PEA

[5] The Data Warehouse lifecycle toolkit , Ralph Kimball Wiley student Edition