CHAPTER 3

Data Warehouses and OLAP

3.Data Warehouses and OLAP

3.1 Data Warehouse

A data warehouse provides tools for executives and business managers to systematically organize, understand and use their data to make strategic decisions. It is a must have latest marketing weapon and a way to keep customers, by learning more about their needs.

Many possible definitions are there for a data warehouse. A data warehouse is a copy of transaction data specifically structured for query and analysis. Sometimes non-transaction data are stored in a data warehouse-through probably 95-99% of the data usually are transaction data . It is "query and analysis" because the main output from data warehouse systems are either tabular listings (queries) with minimal formatting or highly formatted"formal"reports.W.H.Inmon a leading architect in the construction of data warehouse systems defines it to be "A data ware house is a subject-oriented, integrated and time variant volatile data in collection of data in support of management's decision making process".

Subject oriented

The data warehouse is organized around major subjects such as a customer, product, supplier, sales etc.Data warehousing focuses on modeling and analysis of data for decision-makers.

Integrated

A data warehouse is made by collecting heterogeneous collection of data, which is, integrated e.g. flat files, relational databases etc.

Time variant

Data is stored from a historical perspective (over the last ten years, twenty years etc.). Every key structure in the database has an element of time embedded within it.

Non volatile

It is data stored that is physically separate from the optional database-it requires only two operations loading and access to data unlike a transaction processing system which requires concurrently control, processing and recovery mechanisms.

3.2 Differences between Operational Systems and Data Warehouses

The main feature of an online transaction processing system is its ability to perform transform and query processing. The systems are usually known as On Line Transaction Processing (OLTP) systems. They cover day to day operational and transactional data. Operational databases, historic, support large volumes of data (databases of size above 100 GB). Data warehouse on the other hand serve users or knowledge workers in the role of data analysis and decision making. These systems are known as Online Analytical Processing System (OLAP). This major distinguishing feature between OLAP & OLTP.

- **User and system orientation:** Clerks, clients and information technology professionals use OLTP systems and it is customer_ oriented whereas the OLAP is market-oriented used by knowledge workers, analysis and managers.
- **Data contents:** An OLTP system manages current data and is not used for decision making purposes. An OLAP manages large amounts of historical data with facilities for Summerton and aggregation.
- **Database design:** An OLTP system usually adopts an Entity-relationship model and an application oriented database design. An OLAP system uses a star or a snowflake model.

- **View:** An OLTP system rustics itself to data available within a department or an organization whereas OLAP spans versions of database schemes and it makes use of information that generated from organizations integrating information from many data stores.
- **Access Patterns:** Short, atomic transactions are made on OLTP systems and OLAP systems deal with read only operations and complex queries on historical data.

3.3 A Multidimensional Data Model:

The model views data in the form of a data cube. OLAP tools are based on multidimensional data model. Data cubes usually model n-dimensional data.

From Tables Spreadsheets to Data Cubes

A data cube allows data to be modeled and viewed in multiple dimensions. Dimensions are facts that define a data cube. Dimensions are the perspective or entities with respect to which organizations would like to keep records. For example National Bank may create a customer warehouse in order to keep records of the bank's customers with respect to the dimension time, transaction, branch and location. These dimensions allow the bank to keep track of things like monthly transactions, branches and locations where the transactions were made. Each dimension may have a table associated with it, called the dimension table. For example the dimension tables for a transaction might include amount, type of transaction etc.

A multidimensional data model is typically organized around a central; theme like transactions. A fact table represents this theme where facts are numerical measures. Facts are usually quantities, which are used for analyzing relationship between dimensions. The fact table contains then names of facts, or measures, as well as keys related dimensions.

Although we hand to visualize data cubes three-dimensional geometric structures in the data warehouse the data cube inn n-dimensional.

To gain a better understanding of data cubes let us look at a simple 2-D data cube: a spreadsheet from a ABC company. In particular we would like to look at the ABC Company sales data for items sold per quarter in the city of Hyderabad. These data are shown. 3.1

Table 3.1

Location="Hyderabad"				
Item(type)				
Time(quarter)	Home entertainment	Computer	Phone	Security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	35	502
Q4	997	1089	48	630

Table 3.2

Location ="Chennai**" location="**Bangalore**" location="**Calcutta**"**

Item Item Item

time	**Home ent.**	**Comp.**	**Phone sec.**	**Home**	**Ent.**	**Comp.**	**phone**	**sec.**	**Home**	**Ent.**	**Comp.**	**Phone sec.**	
Q1	845	828	90	623	1027	986	34	978	678	825	13	430	
Q2	943	870	5 6	645	1130	1024	45	940	645	765	25	421	
Q3	768	890	45	6 57	1002	940	54	945	789	876	54	543	

Now suppose we would like to view the sales data with a third dimension. For instance, according to time, item as well as location for the cities Calcutta, Bangalore and Chennai. This 3-D data is shown in Table 3.2. Conceptually, we may also represent the same data in the form of a 3-D data cube.

Suppose that we would like to view our sales data with the additional fourth dimension, such as supplier. Viewing these 4-D cubes becomes tricky. However, we can think of 4-D cube as being a series of 3-D cubes.

In data warehousing literature, the data cube such as of the above is referred to as a cuboids. Given a set of dimensions we can construct a lattice of cuboids, each showing data at a different level of summarization, or group by. The lattice of cuboids is then to as a data cube.

3.4Stars, snowflakes and Fact Constellations:

Schemas for Multidimensional Databases

Unlike an entity-relationship model used for relational databases a data warehouse requires a concise subject oriented schema that facilities on-line data analysis. The most commonly used data model for a data warehouse is a multidimensional model. Such a model can exist in the form of a star schema, a snowflake schema or a fact constellation schema.

Star Schema:

The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (fact table) containing the bulk of data with no redundancy, and (2) a set of smaller attunement tables (dimension tables), one for each dimension.

Snowflake:

The snowflake schema is the variant of the star schema model, where some of the dimension tables are normalized, thereby splitting the table into additional tables. The resulting schema graph forms a shape similar or a snowflake.

Fact Constellation:

Sophisticated applications may require multiple fact tables to share dimension tables. This kind of schema can be viewed as a collection of stars. This kind of schema can be viewed as a collection of stars, and hence is called as a galaxy schema or a fact constellation.

Example for defining Star, Snowflake and Fact Constellation Schema

Just as we use relational query languages like SQL, a data miming query language can be used to query a data-mining task DMQL, whi9ch contains language primitives for defining data warehouse and data marts. Data warehouse and data marts can be defined using two language primitives, one for cube definition and another for dimension definition.

The cube definition has the following syntax:

Define cube <cube_name> [(dimensional list)]:<measure list>

The dimension definition has the following syntax:

Define dimension<dimension_name> as (<attribute or sub-dimension list>)

3.5 Review questions

1 Expalin about Data Warehouse

2 list Differences between Operational Systems and Data Warehouses

3 Expalin aboutA Multidimensional Data Model

4 Discus about Stars, snowflakes and Fact Constellations:

3.6 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