# SCHEMA IN DATA WAREHOUSE

Smita Srivastava, Shweta Kurda, Priya Dubey
*Information Technology*
*Dronacharya College of Engineering*

***Abstract*- This paper is concerned about the schema of Data Warehouse. A schema is a collection of database objects, including tables, views, indexes, and synonyms. There is a variety of ways of arranging schema objects in the schema models designed for data warehousing. However, there are other schema models that are commonly used for data warehouses. The most prevalent of these schema models is the Third Normal schema. The star schema is perhaps the simplest data warehouse schema. It is called a star schema because the entity-relationship diagram of this schema resembles a star, with points radiating from a central table. The center of the star consists of a large fact table and the points of the star are the dimension tables. The snowflake schema is a more complex data warehouse model than a star schema, and is a type of star schema. It is called a snowflake schema because the diagram of the schema resembles a snowflake. Snowflake schemas normalize dimensions to eliminate redundancy. That is, the dimension data has been grouped into multiple tables instead of one large table.**

***Index Terms*- schema, star schema, snowflake schema.**

## I. HISTORY

The concept of data warehousing dates back to the late1980 when IBM researchers Barry Devlin and Paul Murphy developed the "business data warehouse". In essence, the data warehousing concept was intended to provide an architectural model for the flow of data from operational systems to decision support environment. The concept attempted to address the various problems associated with this flow, mainly the high costs associated with it. In the absence of a data warehousing architecture, an enormous amount of redundancy was required to support multiple decision support environments. In larger corporations it was typical for multiple decision support environments to operate independently. Though each environment served different users, they often required much of the same stored data. The process of gathering, cleaning and integrating data from various sources, usually from long-term existing operational systems (usually referred to legacy system), was typically in part replicated for each environment. Moreover, the operational systems were frequently reexamined as new decision support requirements emerged. Often new requirements necessitated gathering, cleaning and integrating new data from "data marts" that were tailored for ready access by users.

## II. DEFINITION

A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.

Subject-Oriented: A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.

Integrated: A data warehouse integrates data from multiple data sources. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.

Time-Variant: Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.

Non-volatile: Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

## III. STAR SCHEMA

The star schema architecture is the simplest data warehouse schema. It is called a star schema because the diagram resembles a star, with points radiating from a center. The center of the star consists of fact table and the points of the star are the dimension tables. Usually the fact tables in a star schema are in third normal form(3NF) whereas dimensional tables are de-normalized. Despite the fact that the star schema is the simplest architecture, it is most

commonly used nowadays and is recommended by oracle.

Fact Tables

A fact table typically has two types of columns: foreign keys to dimension tables and measures those that contain numeric facts. A fact table can contain fact's data on detail or aggregated level.

Dimension Tables

A dimension is a structure usually composed of one or more hierarchies that categorizes data. If a dimension hasn't got a hierarchies and levels it is called flat dimension or list. The primary keys of each of the dimension tables are part of the composite primary key of the fact table. Dimensional attributes help to describe the dimensional value. They are normally descriptive, textual values. Dimension tables are generally small in size then fact table.

**Benefits :**

Star schemas are demoralized, meaning the normal rules of normalization applied to transactional relational databases are relaxed during star schema design and implementation. The benefits of star schema demoralization are:

- Simpler queries - star schema join logic is generally simpler than the join logic required retrieving data from a highly normalized transactional schemas.
- Simplified business reporting logic - when compared to highly normalized schemas, the star schema simplifies common business reporting logic, such as period-over-period and as-of reporting.
- Query performance gains - star schemas can provide performance enhancements for read-only reporting applications when compared to highly normalized schemas.
- Fast aggregations - the simpler queries against a star schema can result in improved performance for aggregation operations.
- Feeding cubes - star schemas are used by all OLAP systems to build proprietary OLAP cubes efficiently; in fact, most major OLAP systems provide a ROLAP mode of operation which can use a star schema directly as a source without building a proprietary cube structure.

**Example:**

For example, the following query answers how many TV sets have been sold, for each brand and country, in 1997:

```
SELECT
        P.Brand,
        S.Country AS Countries,
        SUM(F.Units_Sold)
FROM Fact_Sales F
INNER JOIN Dim_Date D    ON F.Date_Id = D.Id
INNER JOIN Dim_Store S   ON F.Store_Id = S.Id
INNER JOIN Dim_Product P ON F.Product_Id =
P.Id
WHERE
        D.YEAR = 1997
AND P.Product_Category = 'tv'
GROUP BY
        P.Brand,
        S.Country
```

- 

**Stars: A Pattern Language for Query Optimized Schema**

For years organizations have deployed and used large on-line transaction processing (OLTP) systems to automate and record their business activity. The challenge now is to allow business analysts, those individuals in the organization chartered to support decision makers by producing reports, to access this data in an ad hoc manner. While OLTP schema are optimized for data entry, they seldom provide an acceptable solution for data analysis.

The star schema concept presented here is the product of many discussions over the years between the author and consultants specializing in Decision Support Systems (DSS). The star schema has also been called a star-join schema, data cube, data list, grid file, and multidimensional schema by practitioners in the field. The name star schema comes from the pattern formed by the entities and relationships when represented as an entity-relationship diagram (ERD); Typically there is a business activity in the center of the star surrounded by the people, places, and things that come together to perform this activity. These people, places, and things can be thought of as the points of the star.

The star schema pattern language presented here is an attempt to provide a method for business analyze to develop a schema which is easy to query. Seven

patterns are presented which are part of a larger pattern language that support the field of DSS, a branch of Business Data Processing Systems. The seven patterns are split into two sections. The first section addresses the issue of finding and organizing the relevant factors of the business that need to be analyzed.

**OLTP vs. OLAP :**

- OLTP (On-line Transaction Processing) is characterized by a large number of short on-line transactions (INSERT, UPDATE, DELETE). The main emphasis for OLTP systems is put on very fast query processing, maintaining data integrity in multi-access environments and an effectiveness measured by number of transactions per second. In OLTP database there is detailed and current data, and schema used to store transactional databases is the entity model (usually 3NF).

- OLAP (On-line Analytical Processing) is characterized by relatively low volume of transactions. Queries are often very complex and involve aggregations. For OLAP systems a response time is an effectiveness measure. OLAP applications are widely used by Data Mining techniques. In OLAP database there is aggregated, historical data, stored in multi-dimensional schemas (usually star schema).

**Reference**

[1] http://datawarehouse4u.info/Data-warehouse-schema-architecture-star-schema.html
[2] http://www.1keydata.com/datawarehousing/data-warehouse-definition.html
[3] http://en.wikipedia.org/wiki/Star_schema
[4] http://www.datafactz.com/blog/wp-content/uploads/2014/07/Data-warehousing-applications.png
[5] http://c2.com/ppr/stars.html
[6] http://datawarehouse4u.info/OLTP-vs-OLAP.html