JACKIE's Forum

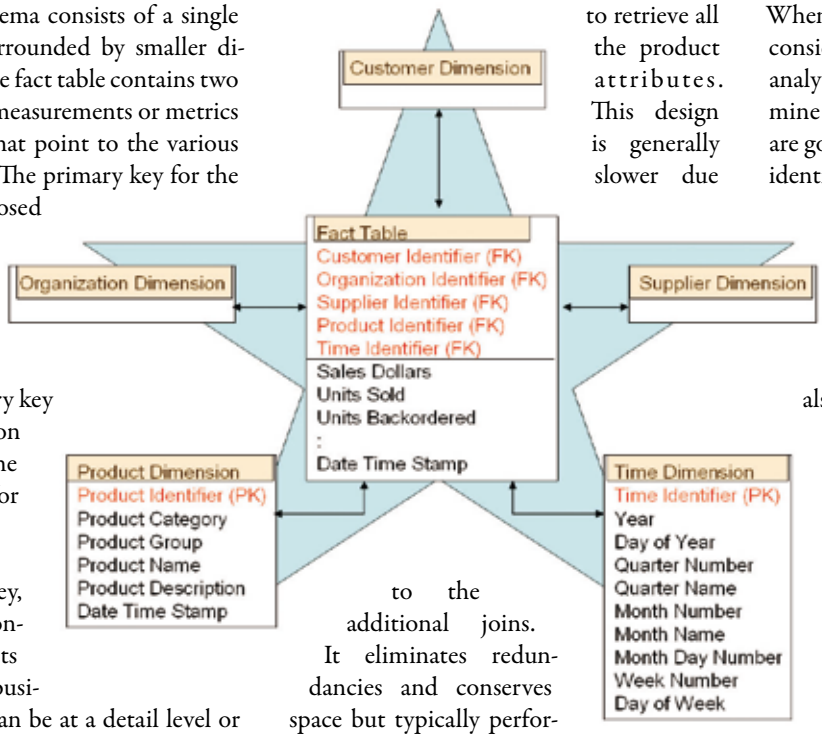
Star Schema

Jackie Jansen

Many of you will have heard the term "star schema" when it comes to database design. For some of you, that may be all you know.

A star schema is a specialized data model for business analysis. It is a design that allows for multi-dimensional database functionality but is implemented in a relational database such as DB2.

Basically a star schema consists of a single large fact table surrounded by smaller dimension tables. The fact table contains two types of columns: measurements or metrics and foreign keys that point to the various dimension tables. The primary key for the fact table is composed of all the foreign keys for the dimension tables. In the example, notice that the primary key of each dimension table is part of the composite key for the fact table.



In addition to the key, the fact table contains measurements or facts about the business. These facts can be at a detail level or they can already be summarized. For example your fact table may contain sales data by day or sales data by month. You want to try and keep the same level of aggregation for all the metrics in the fact table. The sample schema shown here can answers questions such as: sales dollars for a specific product over any time period, YTD sales by supplier, units backordered for a customer last month.

The Dimension tables contain attributes or information about the specific dimension. A customer dimension might include the customer's name, city, state, and country information. In a star schema the hierarchy of the dimension is stored in the dimension table. There is an off shoot of this design called a snowflake schema. In a snowflake model the dimension table is normalized to 3rd normal form instead of 2nd normal form and each hierarchy levels becomes a separate table. In this example Product Categories and Product Groups would each become separate tables that could be joined to a Product Dimension to retrieve all the product attributes. This design is generally slower due to the additional joins. It eliminates redundancies and conserves space but typically performance is a larger concern than disk space in data warehouse installations today. The hierarchy in a dimension table is used both to identify different aggregation levels and also drill down paths. For example, in the product dimension shown here you could ask for sales totals by individual product or by product group or product category. If you were looking at product category you could drill down to see the various groups that made up that category and from there you could drill down to the individual products themselves.

A star query is a join between the fact table and one or more dimension tables. Dimension tables do not join to each other. On the System i, join performance can be enhanced if you build both encoded vector indexes (EVIs) and binary radix indexes over each of the foreign keys in the fact table.

When designing a star schema you need to consider the business process you wish to analyze, such as sales. You need to determine the various metrics or facts that you are going to keep about a sale. You want to identify the dimensions and their attributes. You also need to decide on the level of detail or granularity that you want to store in your fact table. For performance reasons, you may choose to have a fact table with daily information and you may also create a summary table or a materialized query table (MQT - V5R3+) at the monthly aggregation level.

Star schemas are an optimal design for business analysis purposes. Their direct and intuitive mapping is understandable to the end user. Star schemas are optimized for multi-dimensional functionality and are widely supported by most of the Business Intelligence tools on the market. Some tools actual require your data to be in this format before you can access or query the data.

***Jackie Jansen** is a Senior Consulting IT Specialist. She currently works in the IBM Americas Advanced Technical Support Solutions Centre. Jackie is a frequent speaker at iSeries Technical Conferences and User Group meetings. Contact her at jjansen@ca.ibm.com*