Dimension table

A Dimension Table is a table in a star schema of a data warehouse. Data warehouses are built using dimensional data models which consist of fact and dimension tables. Dimension tables are used to describe dimensions; they contain dimension keys, values and attributes.

For example, the time dimension would contain every hour, day, week, month, quarter and year that has occurred since you started your business operations. Product dimension could contain a name and description of products you sell, their unit price, color, weight and other attributes as applicable. Attributes would be a customer's first and last name, age, gender etc.

Dimension tables are typically small, ranging from a few to several thousand rows. Occasionally dimensions can grow fairly large, however. For example, a large credit card company could have a customer dimension with millions of rows. Dividing a data warehouse project into dimensions, provides structured information for reporting purpose.

When you create a dimension, you logically create a structure for your projects. This dimension table can be utilized across for reports and it's about re-usability. If there are any changes to be made, it is evident that only a particular table will get affected. When a company wants to create a report, they can read the data from the dimension table since the table consists of necessary information.

What is a Fact Table?

A Fact Table is a central table in a star schema of a data warehouse. It is an important concept required for ***Data Warehousing and BI Certification***. A fact table stores quantitative information for analysis and is often denormalized. A fact table works with dimension tables and it holds the data to be analyzed and a dimension table stores data about the ways in which the data can be analyzed.

Thus, a fact table consists of two types of columns. The foreign keys column allows to join with dimension tables and the measure columns contain the data that is being analyzed.

Transaction Fact Tables

A Transaction table is the most basic and fundamental view of business operations. These fact tables represent an event that occurred at an instantaneous point in time. A row exists in the fact table for a given customer or product only if a transaction has occurred.

A given customer or product is likely linked to multiple rows in the fact table because the customer or product is involved in more than one transaction. Transaction data often is structured quite easily into a dimensional framework. The lowest-level data is the most natural dimensional data, supporting analyses that cannot be done on summarized data.

Granularity

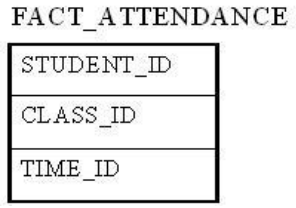
The first step in designing a fact table is to determine the **granularity** of the fact table. By **granularity**, we mean the lowest level of information that will be stored in the fact table. This constitutes two steps:

1. Determine which dimensions will be included.
2. Determine where along the hierarchy of each dimension the information will be kept.

A factless fact table:-

It is a fact table that does not have any measures. It is essentially an intersection of dimensions. On the surface, a factless fact table does not make sense, since a fact table is, after all, about facts. However, there are situations where having this kind of relationship makes sense in data warehousing.

For example, think about a record of student attendance in classes. In this case, the fact table would consist of 3 dimensions: the student dimension, the time dimension, and the class dimension. This factless fact table would look like the following:



The only measure that you can possibly attach to each combination is "1" to show the presence of that particular combination. However, adding a fact that always shows 1 is redundant because we can simply use the COUNT function in SQL to answer the same questions.

Factless fact tables offer the most flexibility in data warehouse design. For example, one can easily answer the following questions with this factless fact table:

- How many students attended a particular class on a particular day?
- How many classes on average does a student attend on a given day?

What is Metadata?

Metadata is simply defined as data about data. The data that is used to represent other data is known as metadata. For example, the index of a book serves as a metadata for the contents in the book. In other words, we can say that metadata is the summarized data that leads us to detailed data. In terms of data warehouse, we can define metadata as follows.

- Metadata is the road-map to a data warehouse.
- Metadata in a data warehouse defines the warehouse objects.
- Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.

Note − In a data warehouse, we create metadata for the data names and definitions of a given data warehouse. Along with this metadata, additional metadata is also created for time-stamping any extracted data, the source of extracted data.

Categories of Metadata

Metadata can be broadly categorized into three categories −

- **Business Metadata** − It has the data ownership information, business definition, and changing policies.
- **Technical Metadata** − It includes database system names, table and column names and sizes, data types and allowed values. Technical metadata also includes structural information such as primary and foreign key attributes and indices.
- **Operational Metadata** − It includes currency of data and data lineage. Currency of data means whether the data is active, archived, or purged. Lineage of data means the history of data migrated and transformation applied on it.

Role of Metadata

Metadata has a very important role in a data warehouse. The role of metadata in a warehouse is different from the warehouse data, yet it plays an important role. The various roles of metadata are explained below.

- Metadata acts as a directory.
- This directory helps the decision support system to locate the contents of the data warehouse.
- Metadata helps in decision support system for mapping of data when data is transformed from operational environment to data warehouse environment.
- Metadata helps in summarization between current detailed data and highly summarized data.
- Metadata also helps in summarization between lightly detailed data and highly summarized data.
- Metadata is used for query tools.
- Metadata is used in extraction and cleansing tools.
- Metadata is used in reporting tools.
- Metadata is used in transformation tools.
- Metadata plays an important role in loading functions.

Metadata Repository

Metadata repository is an integral part of a data warehouse system. It has the following metadata −

- **Definition of data warehouse** − It includes the description of structure of data warehouse. The description is defined by schema, view, hierarchies, derived data definitions, and data mart locations and contents.
- **Business metadata** − It contains has the data ownership information, business definition, and changing policies.
- **Operational Metadata** − It includes currency of data and data lineage. Currency of data means whether the data is active, archived, or purged. Lineage of data means the history of data migrated and transformation applied on it.
- **Data for mapping from operational environment to data warehouse** − It includes the source databases and their contents, data extraction, data partition cleaning, transformation rules, data refresh and purging rules.
- **Algorithms for summarization** − It includes dimension algorithms, data on granularity, aggregation, summarizing, etc.

Challenges for Metadata Management

The importance of metadata can not be overstated. Metadata helps in driving the accuracy of reports, validates data transformation, and ensures the accuracy of calculations. Metadata also enforces the definition of business terms to business end-users. With all these uses of metadata, it also has its challenges. Some of the challenges are discussed below.

- Metadata in a big organization is scattered across the organization. This metadata is spread in spreadsheets, databases, and applications.
- Metadata could be present in text files or multimedia files. To use this data for information management solutions, it has to be correctly defined.
- There are no industry-wide accepted standards. Data management solution vendors have narrow focus.
- There are no easy and accepted methods of passing metadata.

The detailed information about the data itself is called metadata in the system. There are many tools used to manage the metadata and to make the information readily available to the users. These Metadata Management Tools help to know the data well and to manage them according to the users' needs. Data analysis, data management, and data governance can be done easily with the tools as they manage better than the humans themselves. The tools are used to know the data well and to use them when they are asked for. Various rules are incorporated while using these tools.

Importance of Metadata Management Tools

- Any data, be it basic or metadata, should be managed properly to make future forecasts and to see the past performance. Only then, companies can make proper decisions to know the growth and to make necessary arrangements to analyze the business graph. The tools in metadata management help in this aspect.
- Most of the organizations should be GDPR compliant and hence it should know the origin and details of data. Metadata can give all the details regarding the data and can save time rather than toiling through other folders or registries.
- It is not possible always to know the user's details while logging into the system. They cannot be monitored always and this creates problems when the system is attacked by malware or virus. Origin of the virus cannot be traced and now metadata comes to rescue in the form of login details and the time spent in the system and the links checked by them.
- All the data elements should be correlated to know the data well and this helps in collecting the linked elements to form insights about the

data. Metadata management helps in performing these tasks well and to make them available whenever needed.

- If the data is not managed well, it will be difficult to trace data and to make them trustable as a useful source in the organization. Data by itself is critical and it becomes complicated if they are not managed well.
- Many tools available today are based on vendors and hence data management based on the user cannot be done. Time should be spent to check the data, discover the patterns, and to make them work whenever decisions have to be made on data.
- Metadata management makes the work of BI analysts easier than ever due to the volume of data they have to analyze from the huge data lake. They can filter the data easily from the tools and use them to create BI data reports. Without tools, the task is time-consuming and complicated with the amount of data.
- It is difficult to locate the data manually with keywords. Also, it is difficult to check data when a calculation is involved. With the proper management of tools, this task can be done creatively and productively.

- BI people can focus more on the data and the intelligence derived from the data can be used for the analysis. This helps data analysis in a faster manner.

Types of Metadata Management Tools

- Collibra tool is used mostly in data governance to handle huge data in the entire enterprise. All the data can collaborate so that the metadata can be managed well. Interactions with the data can be managed well and can be collaborated with most of the digital technologies such as artificial intelligence, internet of things. This makes data management easier and flexible.
- The Alation tool helps to automate data and to create a catalog so that the data can be managed easily. Data scientists find it easy to approach the tool and use it as data inventory in the system. It provides proper data warnings and collaborations so that the data can be managed and cataloged well.
- IBM has a data management tool called InfoSphere Information Governance Catalog that has different tools to create an environment

for authoring the documents and to create a central catalog to manage the data. Filters and data relationships are known and managed in the sites to keep informed about the data developments. Different industries and data domains can be checked with the help of this tool.

- The Windows registry serves as an actual registry in real life where all the information is kept for future use or for past reference. Here, the data is stored and settings are kept so that it can be checked in between and modified if needed. This makes the system to work efficiently with all the data in hand.
- ASG enterprises have developed a tool called Enterprise Data Intelligence that provides a platform for almost all the data related stuff such as reference data, cataloging, governance, analysis, and data delivery. A portal is provided to access data and the vendor makes sure that the data is secure in the portal. System performance is monitored and workloads are automated so that performance is not affected in the system. Data compliance is monitored regularly so that users need not worry about the same.

- Through *Informatica*, different views of metadata are provided with proper display of relationships and the number of users. Technical and operational management of data can be done with this tool. All the data can be connected end-to-end and encrypted properly so that the data could be analyzed. Also, data relationships can be identified in the tool. The governance approach is flexible and analytics methods are used to manage the data.
- Oracle has different approaches to manage data such as Oracle Enterprise Metadata Management, Oracle Data Relationship Management, and Oracle Enterprise Data Management which is a cloud platform. The data requirements are identified and checked into the tool. They act as a repository to store data and also as a tool to integrate data.