UNIT-2

Learning Objective

- **Principles of dimensional modeling**
- **Dimensional Modeling Implementation**

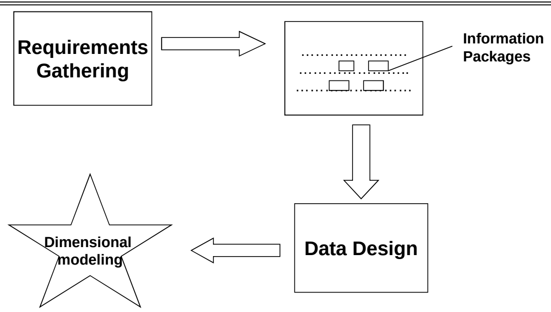
Principles of dimensional modeling

- From Requirements to data design
- STAR schema
- STAR Schema Keys
- Advantages of the STAR Schema

FROM REQUIREMENTS TO DATA DESIGN

- Requirements Definition completely drives the data design for the data warehouse.
- Data design consists of putting together the data structures.
- A group of data elements form a data structure.
- Logical data design includes determination of various data elements that are needed and combination of the data elements into structures of data. It also includes establishing relationship among the data structures.

Requirements Definition Document



FROM REQUIREMENTS TO DATA DESIGN

Requirements Definition Document Cont..

- In the previous diagram , the results of the requirements gathering phase is documented in detail in the requirements definition document.
- An essential component of this document is the set of information package diagrams .
- These are the information matrices showing the metrics ,business dimensions and the hierarchies within business dimensions .
- It forms the basis for the logical data design for the data warehouse .The data design process results in dimensional data model.

DESIGN DECISIONS

- **Choosing the process** : Selecting the subjects from the information packages for the first set of logical structures to be designed.
- **Choosing the grain**: Determining the level of detail for the data in the data structures.
- **Identifying and conforming the dimensions**: Choosing the business dimensions to be included in the first set of structures and making sure that each particular data element in every business dimension is conformed in one another.
- **Choosing the facts:** Selecting the metrics or units of measurements to be included in the first set of structures.
- **Choosing the duration of the database**: Determining how far back in time you should go for historical data.

Dimensional Modeling Basics

- It gets its name from the business dimensions that are required to be incorporated into the logical data model.
- It is the logical design technique to structure the business dimensions and the metrics that are analyzed along these dimensions.
- This model provides high performance for queries and analysis also.

Consider an example,

Example

Consider an example, the information package diagram for

Automaker sales, consists of three data entities:

1)Metrics or measurements

2) Business dimensions, and

3)Attributes .

Example cont..

Now to represent the above information we need to come up with the data structures to represent these three types of data entities.

Let us work with the measurements and analysis that are provided at the bottom of the information package diagram. These are the facts for analysis .The facts are:

Example cont..

- Actual sale price ,MSRP sale price, options price, full price, dealer
- add-ons ,dealer credits, dealer invoice, amount of deployment etc.
- We will group these facts in a single data structure .These facts will
- form the fact table.

Example cont..

Each fact item or measurement goes into the fact table as an attribute for automaker sales.

Now will move to the information package diagram ,taking the business dimensions one by one.

The product business dimension is used when we want to analyze the facts by products. The list of data items relating to the product dimension are : model name, model year ,package styling, product line, product category, exterior color ,interior color, first model year.

Next, we will move to other sections of the information package diagram ,taking the business dimensions one by one.

The product business dimensions is used when we want to analyze the facts by products .The various data items relating to the product dimension are :model name ,model year etc .All of these data items relate to the product in some way.We can combine them in one relational table and call it as product dimension table.

Example cont..

We can form the dimension tables for other business dimensions like dealer, customer demographics etc.

After creating the fact and dimensional tables, we must arrange these tables in the dimensional model.

The dimensional model should primarily facilitate queries and analysis, that will analyze the metrics across one or more dimensions using the dimension table attributes.

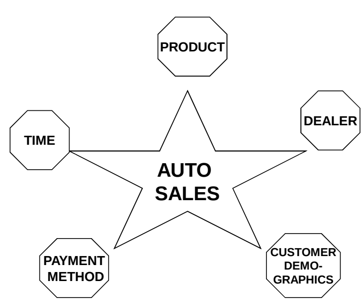
The criteria that should be used for combining tables into a dimensional model:

The model should provide the best data access.

Example cont..

- The whole model must be query-centric.
- It must be optimized for queries and analyses.
- The model must show that the dimension tables must interact with fact table.
- It should be structured in such a way that every dimension can interact equally with the fact table.
- The model should allow drilling down and rolling up along dimension hierarchies.

Example cont..



Example cont..

With the requirements listed before we can find the dimensional model with the fact table in the middle and the dimensions tables arranged around the fact table satisfies the conditions .The dimensional tables have the direct relationship with the fact table in the middle.

Example cont..

Such an arrangement in the dimensional model looks like a star formation ,with the fact table at the core of the star and the dimension tables along the spikes of the star. The dimensional model is therefore called a **Star schema.**

Each dimension table is associated with fact table in one-to-many relationship , for one row in the product dimension table, there are one or more related rows in the fact table.

E-R MODELING VERSUS DIMENSIONAL MODELING

E-R MODELING

Used primarily for OLTP systems

- **OLTP systems capture details of events or transactions .**
- **OLTP systems focus on individual events.**
- **An OLTP system is a window into micro-level transactions.**
- **Picture at detail level necessary to run the business.**
- **Suitable only for questions at transaction level.**

Removes data redundancy

Ensures data consistency

Expresses microscopic relationships.

DIMENSIONAL MODELING

Most suitable for modeling the data for the data warehouse.

- DW meant to answer questions on overall process.
- DW focus is on how managers view the business.
- DW reveals business trends.
- Information is centered around a business process.
- Captures critical measures

Views along dimensions

Intuitive to business users.

USE OF CASE TOOLS

Many case tools are available for data modeling.

- These tools can be used to define the tables ,the attributes and the relationships.
- After creating the initial model ,we can change fields characteristics ,add fields , create new relationships ,and make any number of revisions with utmost ease.
- Case tools helps in forward engineering the model and generate the schema for the target database system.
- For modeling the data warehouse ,we are interested in the dimensional modeling technique.

Star Schema

Star Schema contains a fact table in the middle connected to a set of dimension tables

Let us examine the features of star schema dimensional model with the help of an example

Example : Order Analysis

Consider a schema for a manufacturing company

The marketing department is interested to know about its performance regarding the orders it receives

Metrics/Facts to be analyzed are:

- Cost
- Quantity sold
- Order amount
- Profit margin

These constitute the fact table

Example

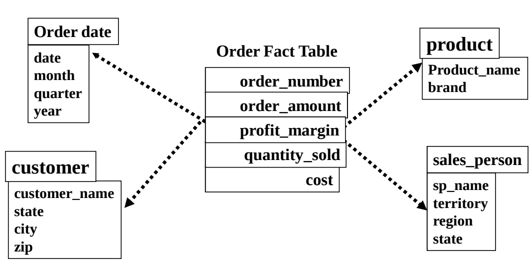
The metrics are analyzed along four business dimensions:

- Product
- Customer
- Sales Persons
- Date

The star schema looks as follows for order analysis:

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Shivendra Goel, Lecturer. U2.‹#›

Example of Star Schema



© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Shivendra Goel, Lecturer. U2.‹#›

Example contd.

For any metric, the star schema implicitly answers the questions of

- What-what product
- When-what date
- By whom-sales person
- To whom-customer

For example, for a given order amount? what product was sold? when was the order placed? who was the customer? which salesperson brought the order?

Example contd.

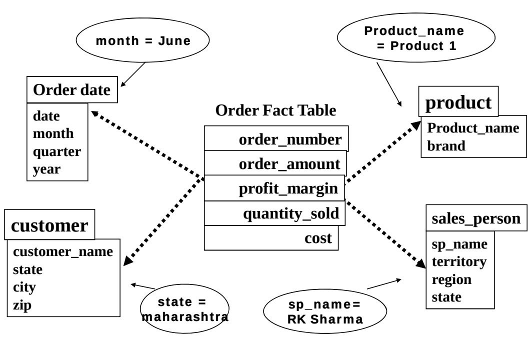
When such a query is made against a DW, then the results of the query are produced by joining one or more dimension tables with the fact table

Example of a simple query:

- The marketing deptt. Wants the metrics quantity sold and order amount for product “Product 1” relating to customers in the state of Maharashtra, obtained by salesperson RK Sharma, during the month of June

The following figure shows how the query is formulated from the star schema:

Query Formulation from the Star Schema



Drill Down Analysis

Analyzing the DW to get lower level of details

The given query is:

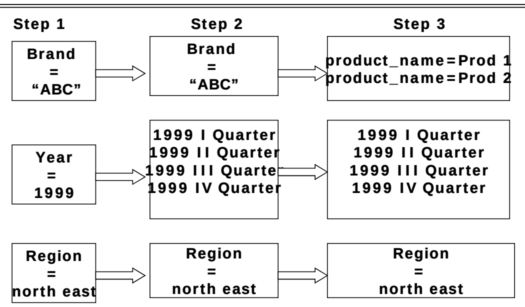
- Give the total quantity sold of the product brand = “ABC” , by the salespersons in the north-east region

In the next step of analysis, the query is

- Drill down to the level of quarters in 1999 for the north-east region for the same product brand

Next, the analysis goes down to the level of individual products in that brand

Drill Down Analysis cont..



Characteristics of a dimension table

The **primary key** of a dimension table uniquely identifies each row in the table

The table is **wide**. It can contain as many as 50 attributes

Attributes are **textual**, i.e. the attributes do not contain any numerical values for calculations. The attributes are of textual format that is of descriptive nature

Characteristics of a dimension table cont..

Attributes in a dimension table are **not normalized;** an attribute in a dimension table acts as a constraint in a query and is directly applied to a fact table for efficient query processing rather than going through any intermediary tables, which we need to create, if we normalize; in that case, the processing will not be fast and efficient

Characteristics of a dimension table cont..

Drill down and Roll up

- Attributes in a dimension table provide the ability to get to the details from higher level of aggregation to lower level of details.
- As an example, 3 attributes zip, city, and state, form a hierarchy of address dimension.
- We may get the Quantity sold by state, then drill down the sales by city, and, then by zip.
- Going the other way round, we may get the total sales by zip, then roll up to sales by city, and then by state.

A dimension table has **less number of records** than a fact table

Components of fact table

Concatenated key

- A row in a fact table is identified by primary keys of all dimension tables. So, the primary key is formed by the concatenation of the primary keys of all dimension tables.
- The star schema for order analysis with keys of fact table and dimension tables is represented as follows.

Components of fact table cont..

Data grain

- It is the level of details for measurements/facts
- For example, quantity ordered relates to the quantity of a particular product on a certain date. It is at the **most detailed level.** But, if we keep the quantity ordered as quantity of a product for each month, then the data grain is different and is at the higher level of granularity

Components of fact table cont..

Fully Additive Measures

- For a query that asks for totals for product = "Prod 1", in month of June for all customers in state = "Delhi", we need to find all the rows in the fact table, relating to all customers in that state and add order amount, cost, and quantity sold to get the totals.

Components of fact table cont..

Fully Additive Measures

- The values of these attributes can be summed up by simple additions.
- Such measures (that is, order amount, cost, and quantity sold) are called fully additive measures.

Semi-Additive Measures

- Suppose the order amount is Rs. 120/- and cost to the manufacturer is Rs. 100/-. Then profit margin amount is Rs. 20/-. This is a calculated metric derived from order amount and cost. Such derived attributes are not fully additive. These are called semi-additive measures

We must distinguish from semi-additive measures from fully-additive ones while performing aggregations (i.e., totals) in queries.

Components of fact table cont..

Table Deep, Not Wide

- A fact table contains fewer number of attributes than a dimension table, but the number of records in a fact table is very large in comparison.

Degenerate dimensions

- There can be some attributes which are neither facts nor dimensions. But, they may be useful in some types of analysis. These are called degenerate dimensions and are kept as attributes of fact table.
- For example, order number in the order analysis star schema. If we want to know the average number of products per order, then, we have to relate the products to order number to calculate average

Fact-less Fact table

A fact table may contain just a concatenated key and not any facts or measures. This happens when we use a fact table to represent events. This is a fact-less fact table.

For example, to analyze students' attendance, dimensions can be date, course, student, room, and professor.

In every fact table row, attendance will be indicated by number 1 (one), whether absent or present. 1 represents that attendance has been taken for that row. So, there is no need to record 1 in every fact table row.

Hence, there is no measure. No fact has been recorded. So, the table is fact-less.

Primary Key

How to form primary keys for dimension tables?

2 general principles are applied while choosing primary keys for dimension tables

Avoid built-in meanings in the primary key

- For example, the primary key of a product can't include the warehouse code, where a product is stored, since a product can be moved to a different warehouse.

Primary Key cont..

Do not use production system key as the primary key

- As an example, the operational system of a firm assigns a key to each customer. If the customer leaves that firm, same number may be assigned to a new customer.
- If the operational system-generated key is used as the primary key of the customer dimension table, same key may represent rows for the new as well as the retired customer, creating data conflict

Primary Key cont..

How to form primary keys for dimension tables?

Use surrogate keys as primary keys for dimension tables

- A surrogate key is a system generated sequence number that does not have any built-in meaning.
- Surrogate keys will be mapped to operational system keys, yet they are different.
- The operational system keys are kept as additional attributes in dimension table
- For example, product key is surrogate primary key for product dimension.

Foreign keys

Each dimension table has a one-to-many relationship with central fact table

So, the primary key of each dimension must be a foreign key in the fact table that references the corresponding dimension table

For example, if there are 4 dimension tables, namely, product, order_date, customer, and salesperson, then, primary key of each of these tables must be present in the order's fact table as foreign keys.

Primary Keys of a Fact Table

Use a single **compound primary key** whose length is the total length of the keys of the individual dimension tables

- In this case, both the compound key and foreign keys are kept in the fact table
- This option hence increases the size of the fact table

Use a **concatenated primary key**– i.e. – concatenation of all primary keys of dimension tables

- In this case, we don't need to keep separate foreign keys. The individual parts of primary keys themselves serve as foreign keys
- This is the mostly used option. It easily relates fact table rows with dimension table rows

Primary Keys of a Fact Table cont..

Use a primary key **independent** of the keys of dimension tables

- In this case, foreign keys must be kept in the fact table
- This option also increases the size of the fact table

Advantages of Star Schema

Easy for users to understand

- The users interact with DW directly through query tools. Users themselves formulate the queries, for which they must understand the data structure and its organization
- Star schema represents the data in a form needed by users for querying and analysis
- Fact table contains the metrics which needs to be analyzed along the business dimensions, and
- Dimension tables contain the attributes along which the users usually query and analyze

Advantages of Star Schema contd..

Optimizes navigation

- Star schema optimizes navigation even when we look for a query result that seems complex. The navigation is still simple and straight-forward
- For example, show the total quantity sold for product A of brand ABC by salespersons in north-east region for first 2 quarters of year 1999
- This query can be easily understood using a star schema as follows

Conclusion

- The Star Schema used for data design is a relational model consisting of Fact and Dimension tables.
- The components of the dimensional model are derived from the information packages in the requirements definition.
- Star Schema advantages are: easy for users to understand, optimizes navigation, most suitable for query processing, and enables specific performance schemes.

Dimensional Modeling

- Updates to the Dimension tables
- Miscellaneous dimensions
- Snowflake schema
- Aggregate fact tables
- Families of STARS

Updates to dimension tables

Changes to dimension tables may be classified into 3 categories:

1. Type 1:Correction of errors
2. Type 2:Preservation of history
3. Type 3:Tentative soft Revisions

Miscellaneous Dimensions

- Large Dimensions
- Rapidly Changing Dimensions
- Multiple Hierarchies
- Junk Dimensions

Large Dimensions

- A large dimension is very deep;has a large number of rows.
- It may be very wide & may have a large number of attributes.
- Large dimensions call for special considerations.

Multiple Hierarchies

Large dimensions often have multiple hierarchies.

Eg:Product dimension of a large retailer

One set of attributes may from the hierarchy for the marketing department.Users from that department use these attributes to drill up & drill down.In the same way the finance department may need to use their own set of attributes from the same product dimension to drill up & drill down.

Rapidly Changing Dimensions

With Type 2 change,we can create additional dimension table row with the new value of the changed attribute.By doing so you are able to preserve the history .If same attribute changes a second time ,you can create one or more dimension table row with the latest value.

Junk Dimensions

Some of the flags & textual data may be too obscure to be real value.These may not be included as significant fields in the major dimensions.At the same time these flags & texts cannot be discarded either.Some of the choices are:

- Exclude & discard all flags & texts.
- Place the flags & texts unchanged in the fact table.
- Make each flag & text a separate dimension table on its own.
- Keep only those flags & texts that are meaningful.

SNOWFLAKE SCHEMA

Snowflaking is a method of normalizing the dimension tables in a STAR schema.

Options to Normalize:

Partially normalize only a few dimension tables,leaving the others intact.

Partially or fully normalize only a few dimension tables,leaving the rest intact.

Fully normalize every dimension table.

Advantages & Disadvantages

Advantages

- Small savings in storage space.
- Normalized structures are easier to update & maintain.

Disadvantages

- Schema is less intuitive & end users are put off by the complexity.
- Ability to browse through the contents difficult.
- Degraded query performance because of additional joins.

Aggregate Fact Tables and Family of STARS

Aggregate Fact Table

Aggregate tables contain information that has a coarser granularity (fewer rows).

An *aggregate table* coexists with the base fact table, and contains pre-aggregated measures build from the fact table.

Why Aggregate Tables?

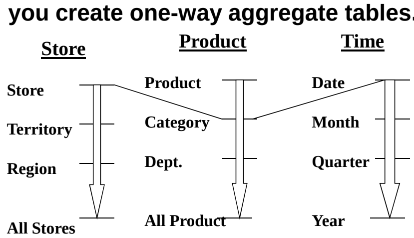
•When a user performs various forms of analysis, he or she must be able to get results comprising of a variety of combinations of individual fact table rows.
• If we do not keep details of individual stores, we can't retrieve results for the product by individual stores. On the other hand, if we do not keep details by individual products, we can't retrieve results for stores by individual product.

Aggregating Fact Tables

- **One-Way Aggregates**
- **Two-Way Aggregates**
- **Three-Way Aggregates**

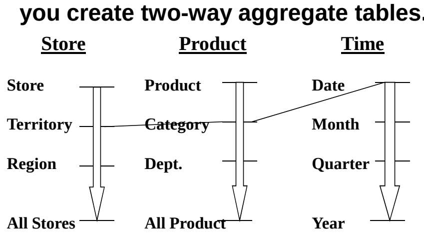
One-Way Aggregate

When you rise to higher levels in the hierarchy of one dimension and keep the level at the lowest in the other dimensions, you create one-way aggregate tables.



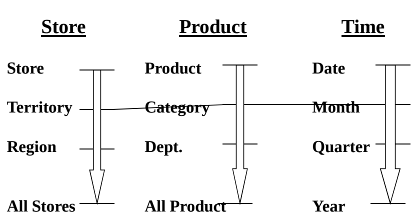
Two-Way Aggregates

When you rise to higher levels in the hierarchies of two dimensions and keep the level at the lowest in the other dimension, then you create two-way aggregate tables.



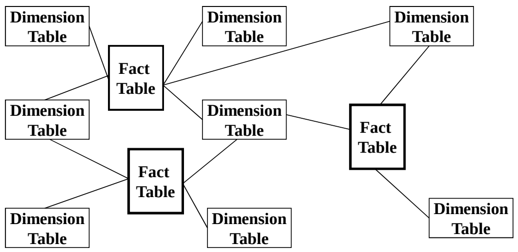
Three-Way Aggregate

When you rise to higher levels in the hierarchies of all the three dimensions, you create three-way aggregate tables.



Families of STARS

Almost all data warehouses contain multiple STAR schema structures.



Families of STARS cont..

- Each STAR serves a specific purpose to track the measures stored in the fact table.
- The fact tables of the STARS in a family share dimension tables.
- Usually, Time dimension is shared by most of the fact tables in the group.

Conclusion

- Slowly changing Dimensions may be classified into three different types based on the nature of changes.
- Snow flaking is a method of normalizing the Star Schema.
- Aggregate or summary tables improve performance. Formulate a strategy for building aggregate tables.

Summary

The ER modeling technique is not suitable for data ware houses; the dimensional modeling is appropriate.

A set of related STAR schemas make up a family of STARS. Examples are snapshot and transaction tables, core and custom tables, and tables supporting a value chain or a value circle. A family of STARS relies on conformed dimension tables and standardized fact tables.

Review Questions

Objective Questions:

1)A star schema has what type of relationship between a dimension and fact table?

a) Many-to-many
b) One-to-one
c) One-to-many
d) All of the above

2)Fact tables are which of the following?

a) Completely demoralized
b) Partially demoralized
c) Completely normalized
d) Partially normalized

Review Questions cont..

3)A snowflake schema is which of the following types of tables?

a) Fact
b) Dimension
c) Helper
d) All of the above

4) Which consists of a fact table with a single table for each dimension:

a) star schema
b) snowflake schema
c) roll-up display
d) drill-down display

Review Questions cont..

5)Pick the correct statement(s):
a) Snowflaking affects the fact table
b) Outriggers affect the fact table
c) Mini-dimensions affect the fact table
d) Role-playing dimensions affect the fact table

6)A dimension can be added to an existing star schema when it is at:
a) A finer granularity than the fact table
b) A coarser granularity than the fact table
c) The same granularity as that of the fact table
d) Granularity has nothing to do with adding a dimension table

Review Questions cont..

7)Pick the odd one out:
a) Data vault
b) Business dimension modeling
c) ER modeling
d) Dimensional modeling

8)Snowflaking is not recommended in a data warehouse because:
a) It makes browsing difficult
b) It prohibits the use of bitmap indexes
c) Queries require more joins
d) All of the above

Review Questions cont..

9)Rapidly changing monster dimension can be handled using:
a) Outrigger
b) Mini-dimension
c) Snow-flaking
d) Vertical splitting

10)Pick the correct statement(s) about fact tables
a) Natural keys can appear in the fact table
b) The same dimension can appear many times in a fact table
c) Base level & summarized data can appear in the same fact table
d) Null values can appear in a fact table

Review Questions cont..

Short answer type Questions

1. Describe Slowly Changing Dimensions.
2. Compare and contrast Type2 and Type3 changing dimensions
3. What are junk dimensions?
4. How does a snowflake schema differ from a STAR schema.
5. Differentiate between slowly and rapidly changing dimensions.
6. Discuss data granularity in a data warehouse.
7. Differentiate between fully additive and semi additive measures
8. A dimension table is wide; the fact table is deep. Explain.
9. What is fact less fact table?
10. Can you think of any disadvantages of STAR schema?

Review Questions cont..

Long answer type Questions

1. What is the STAR schema? What are the component tables?
2. Design a simple STAR schema with a factless fact table to track patients in a hospital by diagnostic procedures and time.
3. Why is the ER modeling technique not suitable for the data warehouse? How is dimensional modeling different?
4. Explain in detail various advantages of STAR schema.
5. Explain all the types of keys used in STAR schema.

Review Questions cont..

6. Explain with the help of an example Family of STARS
7. Discuss the Process of normalization in Dimension modeling.
8. Explain various types of miscellaneous dimensions in detail.
9. What do you mean by “UPDATES TO THE DIMENSION TABLES”. Explain the complete process with the help of example.
10. Can you treat rapidly changing dimensions in the same way as Type 2 slowly changing dimensions? Discuss.

Suggested Reading/References

[1]. Paul Raj Poonia, "Fundamentals of Data Warehousing", John Wiley & Sons, 2003.

[2]. Sam Anahony, "Data Warehousing in the real world: A practical guide for building decision support systems", John Wiley, 2004

[3]. W. H. Inmon, "Building the operational data store", 2nd Ed., John Wiley, 1999.

[4]. Kamber and Han, "Data Mining Concepts and Techniques", Hartcourt India P. Ltd.,2001