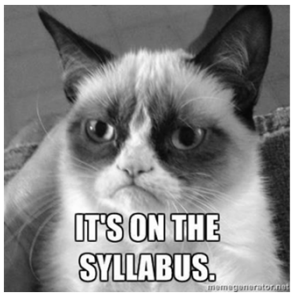
Modeling for Analytical Databases

(Chapter 11)

1



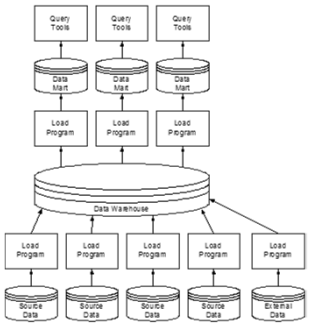
2

Data Warehouse (DW) Properties

- Organized around major subject areas of an organization (e.g. Product, Sales)
- Integrated from multiple operational (OLTP) data sources
- Periodically updated based on an established schedule (data lags behind operational sources)

3

Data Warehouse / Data Mart Architecture



4

Potential DW Benefits

- Competitive advantage
- Increased productivity of corporate decision makers
- Potential high return on investment through improved efficiency and/or profitability

5

DW Characteristics

- Data **integration**
  - With data from many operational databases
- **Loads** data rather than performs updates
- **Less predictable** database 'hits'
- **Complex queries** but a simple interface
- May emphasize **historical data**
- May **summarize** other data
- **Fewer** users, **fewer** queries, yet **many more** rows processed

6

DW Challenges

- Underestimation of required resources
- Hidden data quality issues in source data
- Omitting data only to find later that it is required
- Ever-increasing user demands
- Consolidating data from disparate sources
- High resource demands (storage and processsing)
- Ownership of the data
- Difficulty establishing genuine requirements
- "Big bang" projects that seem never-ending

7

Differences Compared with Operational Databases

- Different requirements
  - Historical vs. current data
  - Detailed and summarized data
  - Once captured, DW data usually does not change
- Database technology may not be relational
- Design techniques are different, but many OLTP design techniques can still apply
  - Normalization offers some efficiencies, but not as essential as with OLTP
  - Performance often trumps maintainability

8

Quality Criteria Revisited

- Completeness
- Non-redundancy
- Enforcement of (business) rules
- Data reusability
- Stability and flexibility
- Simplicity and elegance
- Communication and effectiveness
- Performance

9

Modeling / designing?

- Data warehouses feed data marts
  - Need a separate approach
  - Marts are more focused
  - Warehouses are more general and must handle all marts envisaged
- Consider each in turn

10

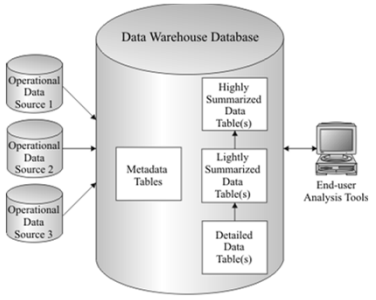
Modeling for Data Warehouses

1. May need an initial corporate model of the business
2. Need to understand existing (operational) data (bases)
3. Determine requirements of the warehouse
4. Determine sources and handling differences
5. Shaping data for data marts

Last two steps are more complex

11

Summary Table Architecture



12

Sources and Differences when Designing

- Minimize number of source systems
- Carefully judge source data item quality
- Reconcile multiple sources
  - Eg. Differences in timeframe and currency of item
- Handle compatibility of coding schemes for data items
- Unpack overloaded attributes
  - Eg. Address containing postcode where postcode becomes an important part of a warehouse

13

Shaping Data for Data Marts

- Need to **maximize flexibility**
- Cater for **common purposes between marts** and basic commonality (sorted out when handing requirements for the warehouse)
- If difficult to cater for both flexibility *and* common purpose opt for flexibility
- The rule: **Maximize Flexibility, Minimize Anticipation**

14

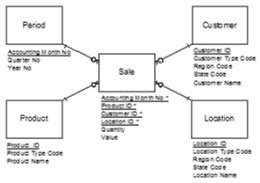
Modeling for Data Marts

- Modeling for general business people who
  - Have little technical knowledge
  - Often need special/complex queries
- Much simpler than operational databases
  - Facts vs. transaction handling and complex business rules
- Users of data marts need to move easily between marts

15

Basic Data Mart Architecture: Star Schema

- Fact table (only one)
- Dimension tables
  - To classify fact table into categories



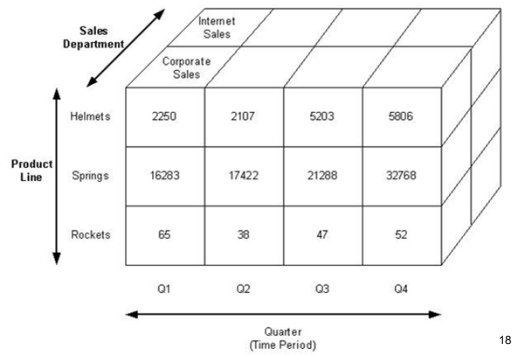
16

Figure 12-3: Four-Column Fact Table for Acme Industries

Product Line	Sales Department	Quarter	Quantity
Helmets	Corporate Sales	1	2250
Helmets	Corporate Sales	2	2107
Helmets	Corporate Sales	3	5203
Helmets	Corporate Sales	4	5806
Helmets	Internet Sales	1	1607
Helmets	Internet Sales	2	1812
Helmets	Internet Sales	3	4834
Helmets	Internet Sales	4	5150
Springs	Corporate Sales	1	16283
Springs	Corporate Sales	2	17422
Springs	Corporate Sales	3	21288
Springs	Corporate Sales	4	32768
Springs	Internet Sales	1	12
Springs	Internet Sales	2	24
Springs	Internet Sales	3	48
Springs	Internet Sales	4	48
Rockets	Corporate Sales	1	65
Rockets	Corporate Sales	2	38
Rockets	Corporate Sales	3	47
Rockets	Corporate Sales	4	52
Rockets	Internet Sales	1	2
Rockets	Internet Sales	2	1
Rockets	Internet Sales	3	6
Rockets	Internet Sales	4	9

17

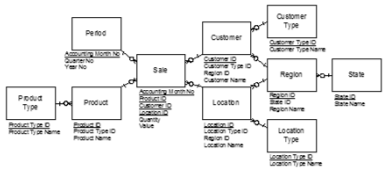
F12-04: Three-Dimensional Cube



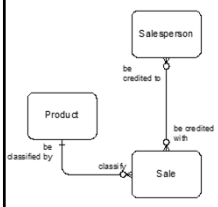
18

Alternative Architectures: Snowflake

- One fact table
- Dimensions are hierarchical
  - Collapse 1:many relationships through de-normalization



Snowflakes and Many-to-Many Relationships



Cannot be handled without action:

1. Ignore less common cases
   - But include data in fact (eg. Include Sales_Person_Count in fact table)
2. Use a repeating group in the dimension table
3. Treat sale-by-salesperson as the fact table

Whatever you do, involve the business users in the decision-making process

Time-dependent Data

- History and time are common in data marts and you must be able to
  - Handle different granularities of time
  - Cater for overlapping periods
  - Consider hierarchies of time periods
- Slowly changing dimensions are common (eg. People may move customer categories over time)
  - Speed of dimension data change
  - Speed of moving fact data from one dimension to another

Dimension Change Example

- Customers can change group
- Solutions
  - Two group foreign keys (now, and at time of purchase / transaction)
  - Ignore if change is slow and cost of ignoring it is low
  - Hold a history of each customer's membership of groups

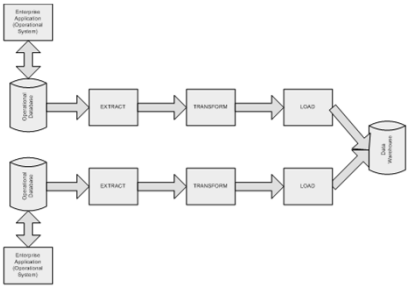
22

Data Integration Methods (+)

- ETL: Extract, Transform and Load.
  - Periodic (schedule) bulk process
  - Good for loading/refreshing data warehouses and data marts
  - Commercial packages (e.g. IBM [Ascential] Datastage or custom developed).
  - Common transformations are summarization, categorization, recoding
  - The target for the data is a centralized database

23

ETL



24

EAI

- EAI: Enterprise Application Integration
  - Framework of integrating data among disparate applications
  - Usually accomplished with push technology that is event-driven
  - Message queues are a common implementation method
  - The target for the data is an application

25

EAI

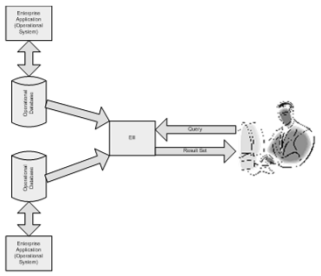
26

EII

- EII: Enterprise Information Integration
  - Real-time integration of disparate data sources
  - As queries are run, data is gathered from the various sources to satisfy the request
  - The target for the data is a person

27

EII



28

Concluding Word

- Data warehouses and data marts are complex
- Specific design challenges and limitations exist
- Patterns are useful here
- Do further reading about the area if you're interested

29

Dimensional Modeling Demo

30