Tutorial for ER2000
Logical Design for Data Warehousing

Il-Yeol Song, Ph.D.
Professor
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104
Song@drexel.edu
http://www.cis.drexel.edu/faculty/song/
October 2000

Contents

• Introduction
• Lifecycles for Data Warehouse Developments
• Database Design for Data Warehouses
• Dimensional Modeling for ROLAP
  ▪ Star Schema and its Variations
  ▪ Model Development Methods
  ▪ Conceptual Models for Dimensional Modeling
  ▪ General Guidelines for Star Schema Design
• Conclusion
Data Warehouse (DW)

- A DW is an *integrated repository* of data that is put into a form that can be *easily understood, interpreted, and analyzed* by the people who need to use it *to make decisions*.
- Data are *extracted* from operational systems, then *integrated, cleansed, transformed, and aggregated*, into a read-only database that is optimized for decision-making.

Data Warehousing (DW) (cont.)

- A Data warehouse is a
  - Subject-oriented,
  - Integrated,
  - Time-variant,
  - Non-volatile
  - Collection of data in support of management’s decision-making process.  
    
    (W. H. Inmon, 1992)
Motivation for DW

- An enterprise-wide repository of data, information, knowledge, and meta-data
- Separation of user-accessible data for business intelligence from OLTP data
- Transform the data into information
- Perform sophisticated analysis of data
OLTP and OLAP/DM Queries

• A query that doesn't require OLAP:
  - How many shoes did we sell last month?

• Queries that require OLAP/Data Mining:
  - How many size 10 shoes in red did we sell last month in the midwest, the northeast, the southeast, compared with the same month last year, actual vs. Budget?
  - What are the top 25 brands, by products, styles, and regions, for this period for total US based on sales dollars?
  - How much promotional expenses did we spend on customers who purchased less than $100 worth of products?
  - What percent of the market do we own?
  - How are our defect rates improving?

OLTP and OLAP/DM Queries

• Queries that require OLAP/Data Mining: (cont.):
  - How much discount should we offer to boost the sales volume significantly?
  - Find the correlation between buying patterns of products of type A and those of type B. What are sales trends?
  - Are our profits are increasing or decreasing?
  - What are 10 best risks?

  - What is the profile of customers who are likely to respond to future mailings?
  - Which existing customers are likely to buy our next new products?
  - Which customers are likely to switch to the competition in the next six months?
  - Is this customer likely to be a good credit risk?
Overview of DW Technologies

- DW generation
- DW data management
- Data management at ROLAP/MOLAP
- Data access

DW generation
- Extract from multiple sources and transform data
- Filter the data to eliminate unnecessary data
- Clean data to eliminate incorrect and duplicate data
- Convert into OLAP format
- Load the data into the data warehouse
Overview of DW Technologies  
(cont.)

• **DW Management**
  - Manage metadata
  - Design the DW
  - Aggregate the data
  - Ensure data quality and consistency
  - Manage the system (security, performance tuning)

• **Data Management**
  - Database for ROLAP or MOLAP

Overview of DW Technologies  
(cont.)

• **Access Tools**
  - Key factors: ease-of-use and power
  - Predefined access: use 4GL
  - Ad-hoc access: query and reporting tools
  - Advanced tools: statistical analysis and visualization tools
  - Data mining tools
Strategies for Reducing Risks

• Facts
  ▪ The 80% of current corporate DW projects fail to meet organizational objectives while 40% fail completely. (Earl Hadden & Associates, Bersin 1996)

Strategies for Reducing Risks (cont.)

• Develop a business case
• Use incremental development methodology
• Plan a data warehouse, build a data mart
  ▪ Use conformed dimensions and conformed facts
• Use tools to automate warehouse population and management
• Keep metadata
  ▪ Manage the quality of data based on that metadata
• Build on an architecture
Strategies for Reducing Risks (cont.)

- Build on an architecture
  - Maintain consistency with the existing architecture
  - Develop a technical blueprint
    - The overall system architecture
    - The sever and data mart architecture
    - The essential components of database design
    - The data retention strategy
    - The backup and recovery strategy
    - The capacity plan for hardware and infrastructure

DW Lifecycles

- Characteristics:
  - DW tends to grow exponentially
  - DW is not a system, but a process.

- Two examples:
  - Kimball and others (KRRT98)
  - Meyer and Cannon (MC98)
Kimball’s Business Dimensional Lifecycle Diagram

Methodology for designing, building, and deploying the data warehouse. (MC98)
Database Design for DW

- **Objective of a DW:**
  - Creating a database optimized for decision support

- **Views of Using ER model for DW Design**
  - Normalized data models support large numbers of transactions with very few records.
  - ER models tend to be very complex and difficult to navigate.
  - ER model identifies *first* entities, *then* relationships
• Basic requirements of a warehouse design
  ▪ The schema must be simple
  ▪ They must support clean, consistent, and accurate data representation
  ▪ They should support fast query processing
  ▪ They should support fast data loading
  ▪ Long term challenges: *flexibility over efficiency*

• Two types of data representation in DW
  ▪ Dimensional model for ROLAP
    ▪ Star schema
    ▪ Snowflake schema
    ▪ Starflake schema
    ▪ Multi star schema
  ▪ Data cube for MOLAP
OLAP: Typical Operations for OLAP

- Drill-down
- Roll-up
- Dice
- Slice
- Cross-tab
- Cube by
- Others (ranking, moving averages, etc.)

OLAP: Drill-down and Roll-up

- **Drill-down**
  - View the data in a more specialized level within a dimension.
  - Drilling down is adding new headers from the dimension tables.
- **Roll-up**
  - Converse of drill-down
  - Rolling up is removing row headers.
OLAP: Dice and Slice

- **Dice**
  - Rotate the cube to see the data from different perspectives.

- **Slice**
  - Slice data on columns (rows), eliminating the rest of the display. Keep a subset of dimensions for selected values.

OLAP: Cross-tab

- A matrix report with X and Y axis.

<table>
<thead>
<tr>
<th></th>
<th>1998</th>
<th>1999</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>By City</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HC</td>
<td>90</td>
<td>160</td>
</tr>
<tr>
<td>NY</td>
<td>70</td>
<td>240</td>
</tr>
<tr>
<td>BOSTON</td>
<td>60</td>
<td>200</td>
</tr>
<tr>
<td><strong>SUM</strong></td>
<td>180</td>
<td>600</td>
</tr>
</tbody>
</table>

| **By Product** | 200 | 800 | 270 |
| **SUM**        | 1500| 1800| 800 |
OLAP: Cube by (Multiple granularity Aggregates)

- **Traditional SQL**
  
  SELECT Supplier, Part, Customer, SUM(Dollars)
  FROM Sales
  GROUP BY Supplier, Part, Customer;

- **SQL with CUBE Operator** (Gray et. al 96)
  
  SELECT Supplier, Part, Customer, SUM(Dollars)
  FROM Sales
  CUBE BY Supplier, Part, Customer;

OLAP: Cube Operator (Multiple Granularity Aggregates)

- **One figure for each combination of:**
  
  - (supp-value, part-value, cust-value) GROUP BY S, P, C
  - (supp-value, part-value, ALL) GROUP BY S, P
  - (supp-value, ALL, cust-value) GROUP BY S, C
  - (supp-value, ALL, ALL) GROUP BY S
  - (ALL, part-value, cust-value) GROUP BY P, C
  - (ALL, part-value, ALL) GROUP BY P
  - (ALL, ALL, cust-value) GROUP BY C
  - (ALL, ALL, ALL) all data
OLAP: Cube Operator (Multiple Granularity Aggregates)

- One cube by of N attributes = \(2^N\) Group-bys
- Other advanced operations
  - Ranking and computed attribute definitions
  - Sampling, rolling averages, quantiles
- Challenge: provide efficient support for complex analysis queries on very large data sets.

Multi-dimensional OLAP (MOLAP)

- Characteristics of multidimensional data
  - Sparsity: usually more than 90% of cells are empty
  - Use proprietary data store
  - Data must be copied or moved into these data stores
  - Pre-aggregation provides fast response time
  - Scalability problem
  - Evolved from query tools
Multi-dimensional Data Cube: logical

Figure: The Cube: A Metaphor for a Dimensional Model

Multidimensional DB: Physical Structure
(Dinter98)

distribution channel (dc)/
publication campaign (pc)
dc = shipping
pc = TV spot
dc = shipping
pc = bulk mail

Upper Level
Sparse Dimensions

Lower Level
Dense Dimensions (contain data)
Multi-dimensional OLAP (MOLAP) (cont.)

- OLAP Council (http://www.Olapcouncil.org)
  - Founded in January 1995
  - Provide OLAP MD-API
  - Founded by Arbor, Comshare, Pilot, IRI

- OLAP reports (http://www.olapreport.com)

Relational OLAP (ROLAP)

- Provides multi-dimensional analysis against data that remain in a RDBMS
- Typical schema: dimensional model
- No pre-aggregation needed
- Aggregate data are stored in separate tables
- Runs more slowly than MDDs
- Full security and administration provided through RDBMSs
- Problems of ROLAP products
  - Offer limited calculation functionality
  - Hard to maintain aggregate tables in the DW
Dimensional Modeling for ROLAP

- Introduction
- Star Schema and Its Variation
- Model Development Methods
- Additivity
- M:N Between Dimension and Fact table
- Other Dimensional Modeling Techniques
- Conceptual Models for Dimensional Modeling
- Clickstream data marts
- General Guidelines for Star Schema Design

A database schema for data warehousing (by Ralph Kimball)
- Known also as Star Schema and its variations
- Consists of a few central fact tables and many dimension tables
  - Simplify SQL queries
    - Reduce joins by OLAP/Relational engines with well-defined join paths
  - Provides a multidimensional analysis space within a RDB
    - Analyze facts by different dimensions
Example of Star Schema

- Fact table
  - Stores all transactions or factual data that are analyzed
  - Typically numeric measures
  - From millions to more than billion rows
  - Example: Sales, Orders, Inventory, Revenue, Budgets, Bookings, Claims, Treatments, Diagnosis, etc.

- Dimensions tables
  - Attributes about facts
  - Supports grouping, browsing, constraining
  - Example: Time, Customer, Promotions, Demographics, LifeStyles, Products, Stores, Markets

- You are analyzing facts by different dimensions

Star Schema
Time Dimension

- Time table at daily grain: (KRRT 98)
  - Time_key
  - Day_of_week
  - Day_number_in_month
  - Day_number_overall
  - Month
  - Month_number_overall
  - Quarter
  - Fiscal_period
  - Season
  - Holiday_flag
  - Weekday_flag
  - Last_day_in_month_flag

An Example of Star Schema for a credit card company

- Fact table: Financial Measure
  - 165 attributes with one for each year
  - MEASURE96: 80 million rows (60GB)
  - MEASURE97: 90 million rows (65GB)

- Dimensional Tables
  1. Customer Name (CUST_NAME)
     - 15 attributes of private and secure data
  2. Customer Profile (CUST_PROFILE)
     - 85 attributes of generic account data
     - A static table refreshed monthly
(3) **Life Style (LIFE_STYLE)**
   - 168 attributes of all aspects of lifestyle
   - Provided by an outside vendor, MetroMail
   - 20% match (>1.2 million records)

(4) **Demographics (DEMOGRAPHICS)**
   - 43 attributes of household income, age, housing, etc.
   - Acquired from May & Speh
   - 70% match (4.9 million records)

(5) **Credit Bureau**
   - 84 attributes of credit history
   - Used to identify risky customers
   - Provided by Equifax and Trans Union every quarter

(6) **Acquisitions**
   - 92 attributes of application data
   - 9 tables:
     - Status (Approved, Declined, Pended) *
     - Type (Account, Company, Scoring)

(7) **Promotions**
   - All marketing and risk-based customer contacts data

(8) **Personal Identification Table**
   - Unique identification number for each customer
   - To track across multiple accounts and products
   - Confidential table
Star Schema and Its variations

- Star Schema
- Snowflake Schema
- Starflake schema
- Multistar schema (PKB 98, Chapter 9)
- Multifacts schema
- Factless fact tables
Star Schema and Its variations

- Star Schema
  - 1:N between dimension and fact
  - The PK of a fact table consists of FKs from dimensional tables
  - All dimensions are denormalized
  - All dimensions have a surrogate key
  - M:N between the fact table and a dimensional table needs special treatment.

Example of Star Schema

- Snowflake Schema
  - All dimensions are normalized into 3NF
Star Schema and Its variations

- Starflake Schema
  - Combinations and star and snowflake schema

Example of Starflake Schema with M:N Relationship

- Multistar schema (PKB98)
  - A star schema where the concatenated FKs might not provide a PK for the fact table
  - A fact table contains a degenerated dimension (e.g., Rec_no and Rec_line_Item)
Star Schema and Its variations

- Multifacts schema (KRRT98)
  - Many fact tables with one or more dimensions in common
  - Use conformed dimensions
  - Conformed dimensions support drill-across operations
  - Cases:
    - Aggregation
    - Partitioned facts for a single fact table
    - Multiple facts for different business processes (Supply chain analysis)
      - Constellation schema
    - Heterogeneous products schema
    - Transaction and Snapshot schema

Example of Multifacts schema
Star Schema and Its variations

- Factless fact tables (KRRT98)
  - Fact tables with no meaningful measures
  - Cases:
    - Event Tracking System
      - Ex: Student Attendance System
        Dimensions: (Time, Course, Facility, Student, Teacher)
    - Coverage System
      - Promotion Coverage
        Dimensions (Time, Store, Product, Promotion)
      - Ex: Which products on promotion did not sell?
        » (Find list of products in promotion from the coverage factless fact table) - (Find the list of products that did sell)

Model Development Methods

- Conversion from OLTP
- Development from Template
- Custom Development
Model Development Methods

- Conversion from OLTP
  - Convert the enterprise ERD into a set of Star Schema
    - Separate the ERD into separate business processes
    - Select M:N or N-ary relationship with numeric measures (Fact)
    - Denormalize all other tables into dimensions
    - Assign a surrogate key for each dimension
  - Can leverage the existing design
  - Will work when:
    - The models are new, well-designed, and normalized
    - Familiar with the source system OLTP models
    - Well understand the data in OLTP systems

- Development from a Template
  - Get a generic model and customize it
  - Will work when:
    - The organization is small
    - Initial info. Requirements are difficult to define
    - The subject area is clear
    - The template is well-defined and well-documented
  - Good sources (Kimball 96, Adamson98)
Model Development Methods

- Custom Development
  - Build from scratch
  - Requires tremendous efforts and expertise
  - Will work when:
    - Has in-house expert modelers
    - The subject area is small
    - A full implementation is not needed
    - Uses a RAD methodology with JAD
  - Good source (KRRT98)

Additivity

- Additivity: add the facts along all of the dimensions
- Numerical measures of intensity are not always additive.
- Three types:
  - (Fully) additive
  - Semi-additive
  - Non-additive
Additivity: Fully additive

- Can add across any dimensions
  - Examples: sales_dollars, cost, qty_sold, margin_dollars

Example of Star Schema (Grocery Store)

Additivity: Semi-additive

- Can add across some dimensions, but not all
  - Semi-additivity occurs when measures represent levels
    - Example: Account_balance, inventory_level, head_count
  - Snapshots of a value taken at a time; a flow past a point
  - Semiadditive facts must be averaged across time when used across time

Example of Star Schema (Grocery Store)
Additivity: Non-additive

- Can’t add across any dimensions
  - Percentages (or rates) are non-additive
    - Example: Interest_rate, margin_rate, temperature
  - Break down into additive components
    (Keep Margin_dollars and Sales_Dollars, instead of Margin_rate)

M:N Between Dimension and Fact

![Diagram](image)

Figure 1. Healthcare Billable Encounter Schema [KRRT98]
Note: only diagnosis dimension is illustrated
Method A: Bridge Table

Figure 2. Solving Multiple Diagnosis With a Bridge Table [KRRT98]
Note: only diagnosis dimension is illustrated

Method C-1: One-to-One Relationship between Dimension and Fact tables

Figure 7. Denormalized (Non-Positional) Diagnosis Dimensional Table With Positional-Attributes
One-to-One Relationship Between Dimension and Fact Tables
(Note: secondary attributes 3 - 19 omitted for brevity.)
Query with Concatenated Attributes

- Queries

```sql
SELECT BPE.Billed_to_payer_amount
FROM DD, BPE
WHERE DD.diagnosis_key = BPE.diagnosis_key
AND DD.concatenated_diagnosis__sorted__key LIKE '%heart%';
```

Method C-2: One-to-Many Relationship Between Dimension and Fact Tables

```
<table>
<thead>
<tr>
<th>Diagnosis Dimension</th>
<th>Billable Patient Encounter fact table</th>
</tr>
</thead>
<tbody>
<tr>
<td>diagnosis_key (PK)</td>
<td>time_key (FK)</td>
</tr>
<tr>
<td>primary_diagnosis</td>
<td>patient_key (FK)</td>
</tr>
<tr>
<td>primary_diagnosis_description</td>
<td>provider_key (FK)</td>
</tr>
<tr>
<td>secondary_diagnosis1</td>
<td>location_key (FK)</td>
</tr>
<tr>
<td>secondary_diagnosis1_description</td>
<td>payer_key (FK)</td>
</tr>
<tr>
<td>secondary_diagnosis2</td>
<td>procedure_key (FK)</td>
</tr>
<tr>
<td>secondary_diagnosis2_description</td>
<td>diagnosis_key (FK)</td>
</tr>
<tr>
<td>secondary_diagnosis3</td>
<td>billed_to_payer_amount</td>
</tr>
<tr>
<td>secondary_diagnosis3_description</td>
<td></td>
</tr>
<tr>
<td>secondary_diagnosis20</td>
<td></td>
</tr>
<tr>
<td>secondary_diagnosis20_description</td>
<td></td>
</tr>
<tr>
<td>concatenated_diagnosis_sorted_key</td>
<td></td>
</tr>
<tr>
<td>AsthmaFlag</td>
<td></td>
</tr>
<tr>
<td>DiabetesFlag</td>
<td></td>
</tr>
</tbody>
</table>
```

Figure 8. Denormalized (Non-Positional) Diagnosis Dimensional Table With Postional-Attributes

One-to-Many relationship between Dimension and fact tables
(Note: secondary attributes 3 - 19 omitted for brevity.)
Method D: Lowering the Grain of the Fact Table

![Diagram of Fact Table and Diagnosis Dimension]

Figure 9. Method D: Lowering The Grain Of The Fact Table

Other Dimensional Modeling Techniques

- See KRRT98 for:
  - Slowly changing dimensions
  - Large dimensions
  - Junk dimensions
  - Heterogeneous products schema
  - Transaction and snapshot schemas
  - Many-to-One-to-many traps
  - Role playing dimensions
  - Part hierarchies
  - Multinational currency tracking
Conceptual Models for Dimensional Modeling

- Dimensional fact model (GR98)
- StarER Model (TBC99)
- M/ER Model (SBHD98)
- Kimball and others (1998)
Conceptual Models for Dimensional Modeling: StarER Model (TBC99)

![StarER Model Diagram](image)

Conceptual Models for Dimensional Modeling: M/ER Model (SBHD98)

![M/ER Model Diagram](image)

Fig. 7. The ME/R diagram for the analysis of vehicle repairs
Conceptual Models for Dimensional Modeling: Kimball and others (KRRT98)

- Diagrams
  - DW Bus Architecture Matrix of data marts and dimensions
  - Fact table diagram
  - Dimension table detail
  - Fact table detail

- Kimball’s four steps to design a star schema
  1. The data mart
  2. The fact table grain
  3. The dimensions
  4. The facts

Example: E-commerce Warehouse Design (Song & Levan-Schlutz98)

- DW Bus Architecture Matrix

<table>
<thead>
<tr>
<th>Data Mart</th>
<th>Date</th>
<th>Time</th>
<th>Customer</th>
<th>Product</th>
<th>Promotion</th>
<th>Website</th>
<th>Navigation</th>
<th>Advertising</th>
<th>Warehouse</th>
<th>Vendor</th>
<th>Communication</th>
<th>Cust/Comp</th>
<th>Cust/Ship</th>
<th>Ship/Com</th>
<th>Ship/Mode</th>
<th>Duration</th>
<th>Equipment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer Billing</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sales</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Promotions</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advertising</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Customer Service</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Network Infrastructure</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inventory</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shipping</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Receiving</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accounting</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finance</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labor &amp; Payroll</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profit</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example: E-commerce Warehouse Design (cont.) (Song & Levan-Schlutz98)

(1) The data mart: web sales system

(2) The fact table grain: Grain of the Fact Table
   - Sale Line item level

(3) The dimensions: Dimension table detail

(4) The facts: Fact table detail

---

Figure 6. Dimension Detail Diagram for Website Dimension
Fact Table Detail Diagram for E-commerce Sales (Grain: Line Item)

- Date_key
- Time_key
- Customer_key
- Product_key
- Promotion_key
- Website_key
- Navigation_key
- Advertising_key
- Warehouse_key
- Customer_Shops_key
- Ship_mode_key
- Location_key
- Order_number
- Order_line_number
- line_item_product_price
- line_item_quantity
- line_item_total_amount
- line_item_tax
- line_item_discount_amount
- line_item_product_total
- line_item_shipping
- line_item_product_code
- line_item_product_description
- line_item_product_name
- line_item_product_image
- line_item_product_category
- line_item_product_subcategory
- line_item_product_brand
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_department
- line_item_product_category
- line_item_product_subcategory
- line_item_product_size
- line_item_product_color
- line_item_product_package
- line_item_product_sku
- line_item_product_departme
Example: E-commerce Warehouse Design (cont.) (Song & Levan-Schutz98)

Query star schema for brands sold by fiscal_month by city, by ship_mode_type, by promotion_type.

Figure 10. Aggregate Sales Star Schema

Clickstream Data Mart (Kimball 2000)

- Grain types
  - Complete session
    - One record for each competed customer session
    - Can perform extensive demographic and web site effective analyses
  - Page label
    - One record for each page event
    - Can perform detailed maps and trajectories of every web visit
    - Generates too many records
Clickstream Data Marts:
Session level (Kimball 2000)

Dimensions:
  ● Web server day
  ● Web server time (86400 records at the second level)
    ● Visitor day
    ● Visitor time
    ● Visitor
    ● Starting page
    ● Ending page
    ● Referring context
    ● Session diagnosis

Facts:
  ● Number of pages visited
  ● Total dwell time
  ● Orders placed
  ● Units_ordered
  ● Order_dollars

General Guidelines for Star Schema Design

• Develop the Data Warehouse Bus Architecture Matrix
• Start with a single-source data mart, not multiple-source one.
• Use conformed dimensions and conformed facts
  ▪ Take time to design conformed dimensions
  ▪ Define at the most granular level possible
• Start from a star schema and denormalize a dimension when you see the clear reasons not to do so. Examples could be reporting requirements.
• Fact tables are deep and narrow, while dimensions are wide
General Guidelines for Star Schema Design

- **Fact tables**
  - One fact table per business process
  - Every fact must be at the same *grain* level
  - Do not store aggregated columns within fact tables
  - Fact tables should store facts that will not vary over time
  - The PK of a fact table is a concatenation of FKS (and a degenerated dimension)
  - Minimize the column size
  - Check the additivity of measures
  - Partition the fact tables
  - Avoid text fields

- **Dimension tables**

  *A data mart is only as good as its dimension tables.*
  
  (Ralph Kimball, 1996)
General Guidelines for Star Schema Design

- **Dimensional attributes** should be: (KRRT98)
  - *Verbose* (full words, not codes or abbreviations)
  - Descriptive
  - Complete (no missing values)
  - Quality assured (no misspellings, impossible values, obsolete values)
  - Indexed (B-tree for high card. and bitmap for low-card.)
  - Equally available (in a single flat-denormalized dimension)
  - Documented

- **Dimension tables**
  - Don’t use PKs from source database
  - Use a single system-defined PK for each dimension
  - Use *conformed* dimensions among data marts or multi fact tables
  - Model *wide* dimensions
  - Denormalize frequently accessed dimensions
  - Normalize only when denormalized schema require too much disk space or denormalized data are not used together
  - Normalize only one or two largest dimension tables: Starflake schema
General Guidelines for Star Schema Design

- **Dimension tables (Contd.)**
  - An ideal dimension contains many readable text fields
    - Short phrases, not codes or abbreviations
    - A long field must be presented as a series of separate fields with complete words.
  - The best dimensions are the single-valued descriptors attached to a set of measurements
  - The granularity of a dimension could be equal or higher than the fact table granularity

- **Aggregation strategy**
  - Create aggregation in such a way that 70% of queries respond in a reasonable amount of time
  - Summary tables should retain all the dimensions that have not been aggregated
  - Avoid creating too many summary tables
  - Index the summary tables
  - Store each set of pre-aggregated data in a separate summary table
  - Dimension tables of an aggregation table should be shrunken versions of the dimension tables of the base fact table
  - Use materialized view facility
General Guidelines for Star Schema Design

- **Pre-aggregation and dynamic aggregation**
  - Consider dynamic aggregation if the fact table is large but structurally simple
  - Consider pre-aggregation if the fact table is small but structurally complex
  - Pre-aggregate when:
    - The compression ratio (CR) is low:
      - \( CR = \frac{\#Rows\_Displayed}{\#Rows\_Retrieved} \)
    - The dimension structure is very complicated
    - The computations are very complex

---

Top 100 Data Warehouse vendors
(http://www.dmreview.com/top100/)

1. IBM
2. SAS
3. Oracle
4. PLATINUM technology
5. NCR
6. Cognos
7. Compaq
8. Prism Solutions
9. Sybase
10. Business Objects
11. Seagate Software
12. Microsoft
13. Hewlett-Packard
14. Informatica
15. MicroStrategy
16. Brio Technology
17. Sun Microsystems
18. Hyperion Solutions
19. Information Advantage
20. Information Advantage
Top 100 Data Warehouse vendors (Cont’d)
(http://www.dmreview.com/top100/)

21. Red Brick Systems
24. Comshare
25. Syncsort
27. Information Builders
28. SPSS
29. Ardent Software
30. INTERSOLV
31. Sagest Technology
33. Seagate Technology
34. Computer Associates
49. Pilot Software
54. Informix
55. Cayenne
62. Innovative Systems
63. SILVERRUN Technologies
68. Sterling Software
69. Actuate Software
71. LSI Logic Storage Systems
76. Excalibur Technologies
77. Unisys
81. SELECT Software Tools
86. Sequent Computer Systems

Trends in DW Design & Developments

• The use of materialized views
• Merging ROLAP & MOLAP
• Merging OLAP into DB engines
• Building webhouses
• Availability of packaged data mart solutions
• Buying DW solutions, rather than building them
• DWs will be housed on federated database systems
Conclusion

• DW design is a process
  ▪ Efficiency vs flexibility
• Big gap between research and practice
• Research topics
  ▪ Framework for multidimensional models
  ▪ Selection and maintenance of materialized views
  ▪ Mixed design with ROLAP and MOLAP
  ▪ Automatic generation of multidimensional schema
  ▪ OO Approach
  ▪ Automated tools
  ▪ Design for scalability

References


References


References


[SRME00] Song, I.-Y., Rowan, B., Medsker, C., and Ewen, E., An Analysis of Many-to-Many Relationships Between Fact and Dimension Tables in Dimensional Modeling.