IS660J

Lecture 3

Professor K.M. Burns

Agenda

4 step Dimensional Design Process

Visualizing real cases

→ Characterisitics

- Retail Sales
- E-Commerce

Coming weeks

– Inventory

– Procurement

– Order Management

– Customer Relationship Management

4 Step Dimensional Design Process

1. Select the business process to model (high impact and accessible data)
2. Declare the grain of the business process (atomic data)
3. Choose the dimensions that apply to each fact table row – discrete, textlike attributes
4. Identify the numeric facts that will populate each fact table row – continuously valued

Retails Case Study

Background

- UPCs & SKUs
- Data Collection pts – POS, Vendor deliveries
- Logistics of ordering, stocking, and selling products while maximizing profit (sales price less costs)
- Pricing, Volumes and Promotion

Dimensional Model Design

Business Process

- High impact - Management wants to understand what products are selling in which stores on what days under what promotional conditions
- Data is accessible in POS system

Dimensional Model Design

Declare the grain

- Individual line item (implies store and product) on a POS transaction (implies daily)
  - Broadest users
  - Slicing and dicing

Dimensional Model Design

Dimensions

  - Date/Time – found in most data marts
  - Product
  - Store
  - Promotion

See Figure 2.2

Conformed Dimensions

- EDW is comprised of many data marts
- In order for DMs to use the same dimensions (ie. product) in joins with fact tables, the dimension must be the same – consistent keys, consistent attribute names, consistent attribute definitions, and consistent attribute definitions.

Dimensional Model Design

Identify the facts

- Sales Quantity - Additive
- Sales Dollar Amount - Additive
- Cost Dollar Amount - Additive
- Gross Profit Dollar Amount – Additive
- Gross Margin % (GP$/Sales$) – Nonadditive
- Unit Price – Nonadditive

See Figure 2.3

Est. the number of rows in the fact table

Primer on SQL queries

Assignment 3 example

Constraints

SELECT time.Date, products.Brand, products.Description, stores.Region, stores.Store_Location_Desc, sales_facts.Actual_sales_price, sales_facts.Quantity_sold, sales_facts.Sales_revenue FROM time, products, sales_facts, stores WHERE sales_facts.Product_key=products.Product_key AND products.Brand LIKE '%Chef Ronaldo%' AND sales_facts.Time_key=time.Time_key AND (sales_facts.Time_key BETWEEN '0' AND '9999') AND sales_facts.Store_key=stores.Store_Key AND stores.Region LIKE '%%' ORDER BY stores.Region, products.Brand

More about Dimensions

- Date – Why store date attributes in a separate table when we can use SQL date functions? - see Figures 2.4 & 2.5
- Products – Pulled from master; Hierarchy of attributes – see Figure 2.7 → constraints and headers
- Stores – Multiple hierarchies; no master
- Promotions – causal dimension – price reductions, ads, displays and coupons – 1 or 4 addtl dimensions? Tradeoff. Typically 1.
  - "No Promotion in Effect" row; avoid null values in fact table, especially if it compromises referential integrity

Factless Fact Tables

What products were on promotion but not sell?

- Don’t want to overpopulate our fact table
- Promotion Coverage fact table - 2nd table to track what didn’t happen – same key, different grain (day or week)
- Factless because there are no measurements; it merely captures the relationships
- Chapter 12

Degenerate Dimensions

- POS transaction number → Grouping key (DD) → looks like a dimension key but doesn’t carry the attribute info of a dimension (empty); hence, degenerate
- Typically part of the primary key of the fact table
- See Figure 2.10
- Other examples: Order numbers, invoice numbers, bill of lading numbers

Reporting

- What are the weekly sales $ volume by promotion for the snacks category during Jan 2002 for stores in the Boston district?
- See Figure 2.10
- High quality dimension attributes are crucial because they are the source of query constraints and result set labels.
- Cross tabular reports are better.

Extending the Schema

Example: Figure 2.11

Possible alterations:

- New dimension attributes
- New Dimensions
- New Measured Facts – Decision: ALTER TABLE or create a new fact table
- Dimensions become more granular
- Addition of a completely new data source → new fact table

Resist Temptations to …

- Snowflake Dimensions – save space and/or maintenance time at the expense of ease of use and performance – see Figure 2.12
  - Dimension tables are relatively small compared to fact tables.
- Denormalize the fact table –ie. including fields from the Date dimension in Fact tables.
  - 25 or more dimensions may require us to find ways to combine correlated dimensions into a single dimension. Typical fact tables have 15 or less.

Surrogate keys

- Every join between dimension and fact tables should be based on meaningless integer surrogates. Avoid the natural operational product codes.
- Allow the data warehousing team to be buffered from operational changes.

Market Basket Analysis

Which products were sold in the same market basket together? Combinations – Market Basket Analysis, Affinity Grouping

Best done by Data Mining tools and/or OLAP. Why? Constraints and groupings need to be done in the fact table.

Can be done by creating an new fact table – see figure 2.14

Semiadditive measures – don't double count!!!

Summary

- 4 step process for designs
- Loading fact tables with atomic data provides the greatest flexibility because we can summarize at will, in many different directions

E-Commerce

ClickStream

- Every page event recorded by each the company's web servers. Information is valuable.
- + Dimensions – Page, Session, Referrer
- Extremely large data set of text and numbers
- How does it integrate with the rest of the data warehouse? Can its dimensions be "conformed in a data warehouse bus architecture?
  - Bus Architecture is composed of a master suite of conformed dimensions and standardized definitions of facts. Business process data marts throughout an enterprise can "plug into" this bus to receive the dimension and fact tables they need. The Bus thus supports the various processes and associated data marts that measure key aspects of the processes. See figure 3.7

Clickstream

- Many log files, many websites (ISPs, Ad Referrers, Search Engines, Web watcher services, ect.) are involved in 1 session for a user → How do we integrate these sources? How do we sync them in time?
- XML has the potential to make web pages far more expressive

Challenges

- Identifying the Visitor Origin
- Identifying the Session
- Identifying the Visitor
- Proxy Servers
- Browser caches

+ Dimensions

- Page
- Event
- Session
- Referral
- Existing dimensions – Date, Time, ect.

Clickstream Fact Tables

- Complete Sessions
  - Grain – 1 row for each completed user session
  - Dimensions – Date, Time, Customer, Page, Session, Referrer
  - Facts – Session seconds, pages visited, orders placed, order quantity, order $$$ amt
  - See figure 14.3

Clickstream Fact Tables

- Individual Page Events (2nd fact table)
  - – Grain – individual page event in each customer session
  - – Dimensions – Date, Time, Customer, Page, Event, Session, Session Id (degenerate), Product, Causal, Referrer
  - – Facts – Page seconds, order quantity, order $$$ amt
  - – See figure 14.4

Clickstream Fact Tables

- Aggregate (3rd fact table)
  - – Source is session grained fact table
  - – Grain – individual session
  - – Dimensions – Month, Demography, Entry Page, Session Outcome
  - – Facts – Number of Sessions, Session Seconds, Pages Visited, Orders Placed, Order Qty, Order $$$ Amt
  - – See figure 14.5

Integrating into the EDW

- See figure 14.6

Web Profitability Data Mart

- Allecate activity costs and infrastructure costs down to each sales transaction
- View of profitability over all our sales channels, not just the web
- New data mart
  - Grain – individual product sold on a sales ticket to a customer at a point in time (like Retail example)
  - See figure 14.7

Conclusion

Clickstream

- Identifying the Visitor Origin
- Identifying the Session
- Identifying the Visitor
- Proxy Servers
- Design and Integration