CSPP 53017: Data Warehousing

Winter 2013

Lecture 5

Svetlozar Nestorov

Class News

- Class web page: http://bit.ly/WTWXV9
- Subscribe to the mailing list!
- Homework 3 is due on Feb 15 (11:59pm).
- Second 15 minute in-class quiz next time (6:30pm) on Feb 19.
  - Covers the first five lectures, project submissions and the Gradiance homework.
  - Open book/notes
- Last 15 minute in-class quiz will be on Mar 5.

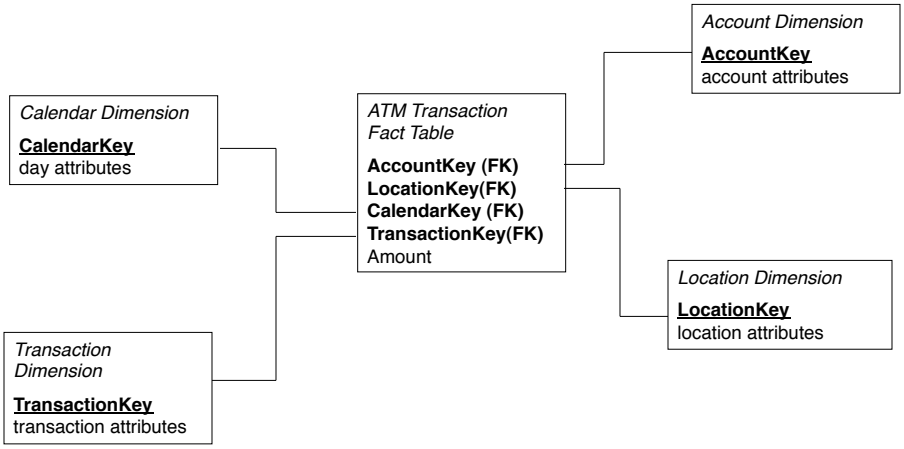
Dimension Hierarchies

- Many-to-one hierarchies within a dimension are often flattened and presented as a series of attributes in the dimension (i.e. opposite of snowflaking)
- Multiple well-defined hierarchies are often present in one table
  - E.g. marketing and finance departments may have incompatible and different views of product hierarchy – in that case all of the marketing-defined attributes and all of the finance-defined attributes must be present in the detailed master product table
- Hierarchies accommodate drill-paths
  - used for drilling-down and drilling up

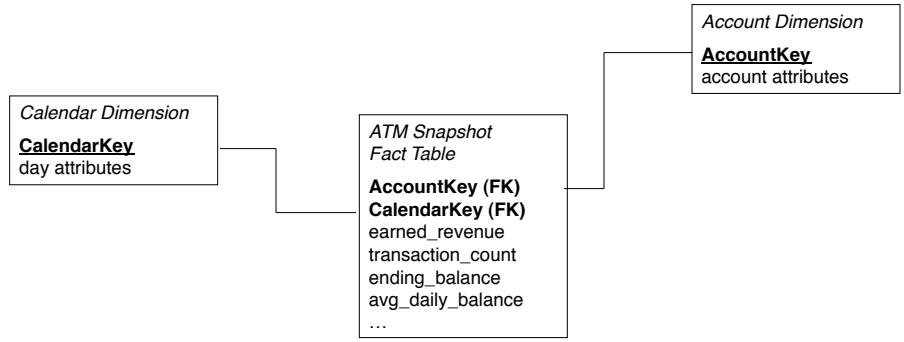
Transaction Schemas

- Usually contains a single "fact" – the value of the transaction
- Used to perform analysis in extreme detail
  - Behavior counts (e.g number of times ATM users make mortgage payments at ATM locations not near their homes)
  - Time-of-day analysis (e.g. number of ATM transactions that occur during the lunch hour)
  - Sequential behavior analysis (used for fraud detection, cancellation warning, etc.)
  - Basket analysis (what products sell in combination with other products)
- Excellent setting for data-mining

Transaction Schema Example



Snapshot Schema Example



Snapshot Motivation

- Not all related facts can be ***easily*** (i.e. quickly) derived from the transaction schema
  - E.g. account transactions vs. the revenue generated by the account
  - It is possible to derive the revenue by crawling through the transactions while taking into account the complex relationships between individual transactions and basic measure of revenue; but sometimes that is not a feasible approach
  - Alternative – *snapshot tables*

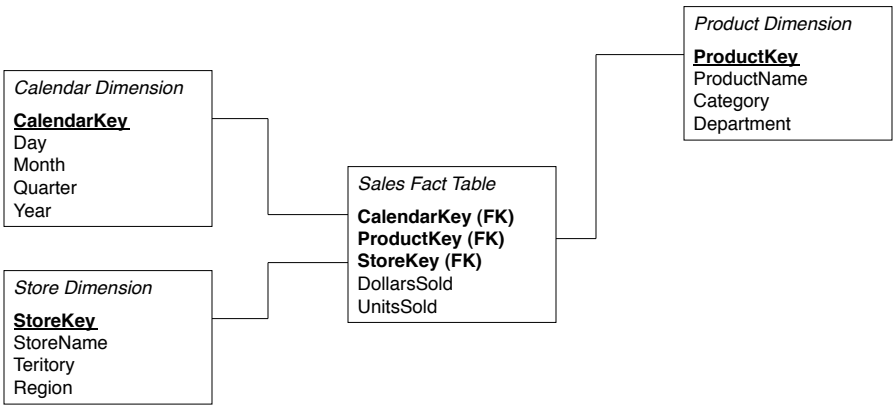
Snapshot Tables

- Records added at the end of specific reporting time period (monthly, daily, hourly)
- Records still created by "crawling" through transactions, but at least this is done only once (for each record)
- *Current rolling snapshot* – the one being created currently
- Facts
  - Completely additive (e.g. Earned Revenue, Total Transaction Count – can add across all dimensions)
  - Semi additive (e.g. Ending Account Balance, Average Daily Balance – can add across all dimensions except time)
- Fact Structure
  - Open-ended: snapshot table can include a variety of facts, counts, and summaries

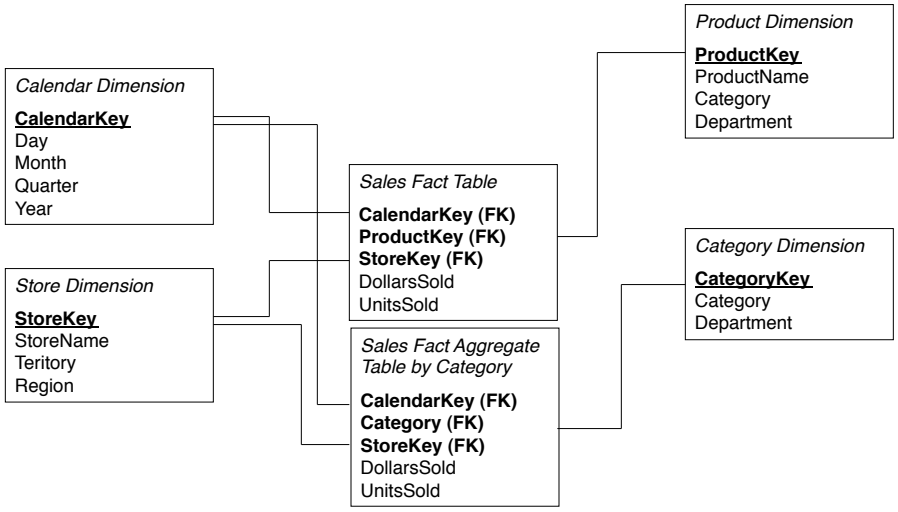
Alternative Schemas

- Accumulating Snapshots
  - Special purpose snapshots (for narrow very specialized use – lots of Type 1 overwriting in the fact table for calendar keys)
- N-way Aggregates
  - Simple snapshot schemas that keep all original dimensions from the transaction schemas and add no newly derived facts
  - Product of simple reduction of the grain level in one or more associated dimensions

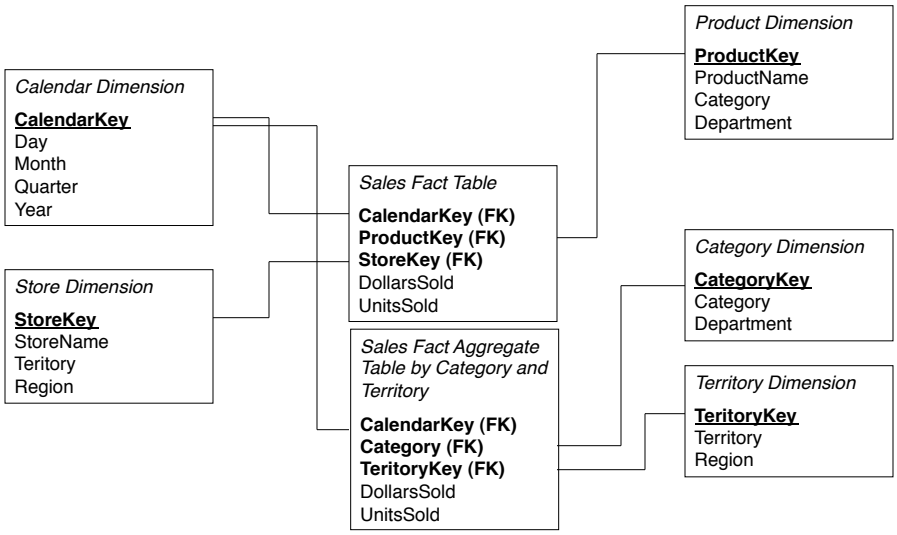
Starting Schema



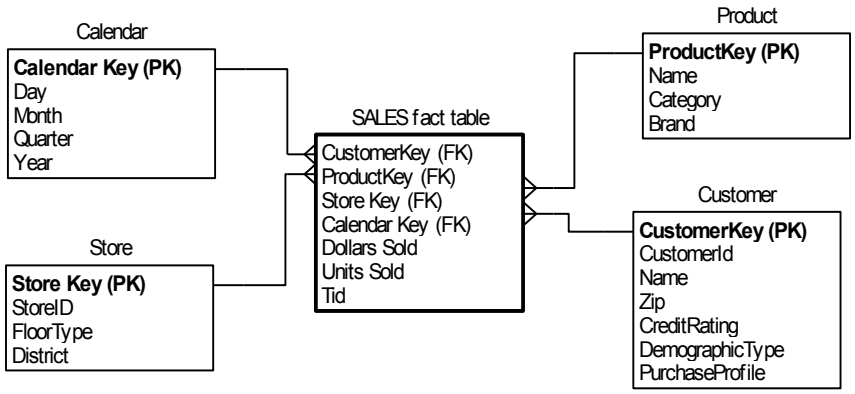
One-way Aggregate



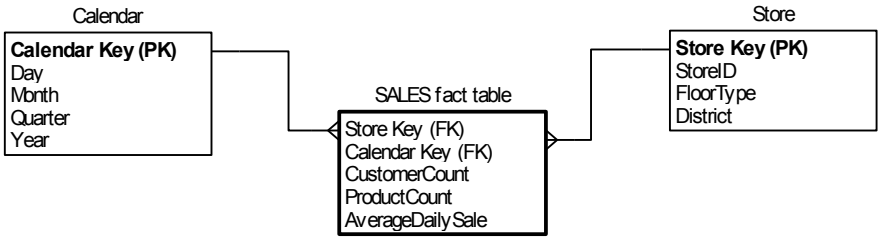
Two-way Aggregates



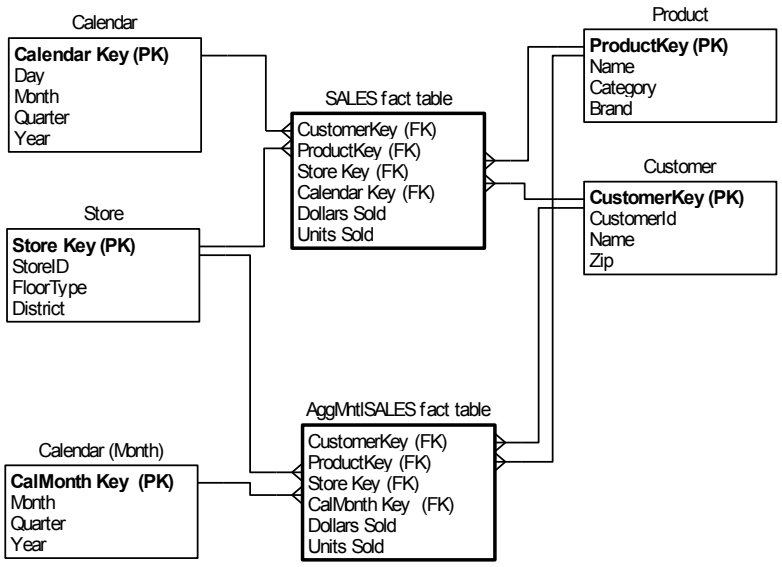
Retail Star Schema



Retail Snapshot Schema



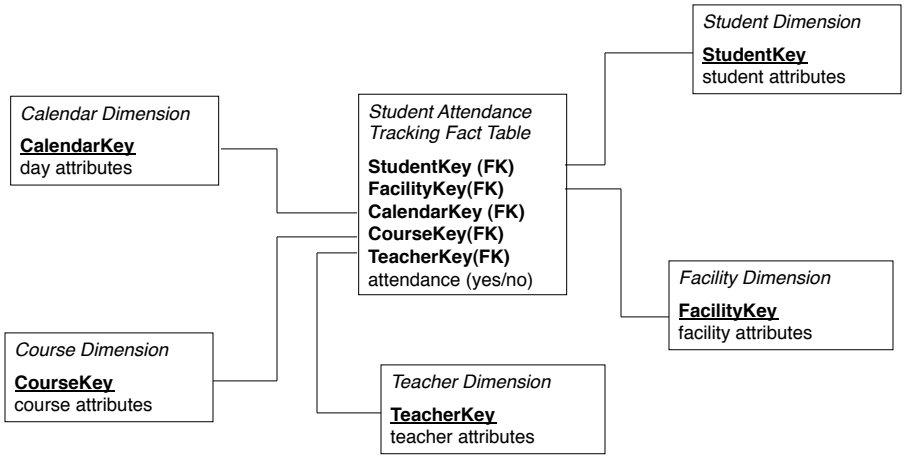
Retail One-way Aggregate



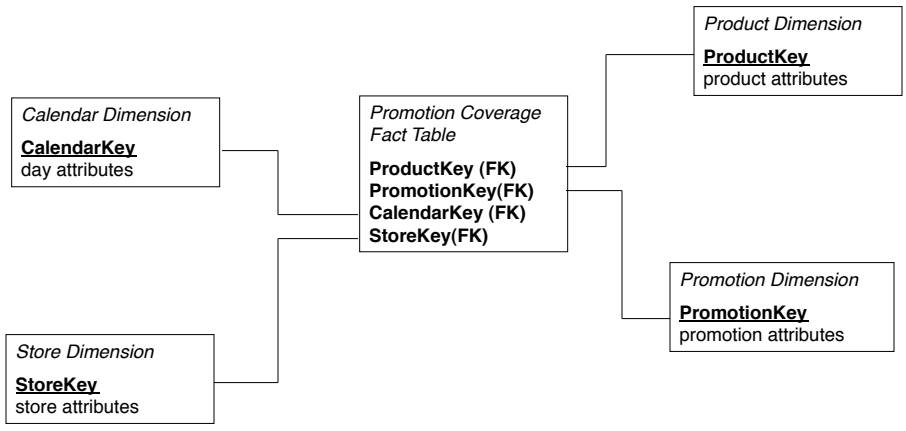
Additional Concepts

- Facts of Differing Granularity
  - Two choices
    - Allocating the higher level facts to a lower granularity
    - Separate fact tables for each granularity
- Multiple Currencies and Units of Measure
  - Sometimes fact measurements need to be expressed in multiple currencies or unit so measure
- Factless Fact Tables
  - No facts in the fact table
  - Used as a method for recording events in a data warehouse when there is no natural numeric measurement associated with the event
  - 2 types of factless fact tables
  - EVENT and COVERAGE

Event-type Factless Table



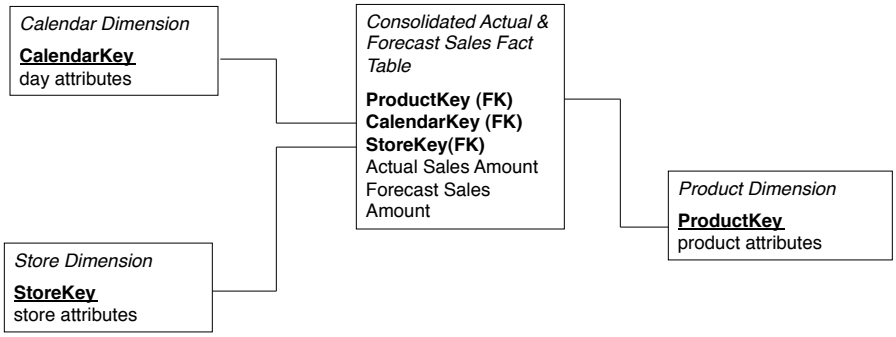
Coverage-type Factless Table



Consolidated Fact Tables

- Measurements from multiple processes can be merged into a single-fact table (if those processes correspond to exact same combinations of dimensions)
- Usually done with aggregated data
- Result in faster query response and less complicated presentation

Consolidate Fact Table Example



Building Dimensional Models

- Getting Started: MATRIX METHOD
  - Identifying all the possible fact tables
  - Identifying all the dimensions implied by those fact tables
- MATRIX: Identifying the Fact Tables
  - FIRST: List all single source fact tables
    - E.g. DELIVERIES FORM SUPPLIERS, CUSTOMER BILLING STATEMENTS, MARKETING PROMOTIONS TRACKING, CUSTOMER SERVICE INQUIRIES
  - SECOND: List multiple source fact tables that combine the single source designs into a broader view of business
    - E.g. CUSTOMER RELATIONSHIP MANAGEMENT (which combines CUSTOMER BILLING STATEMENTS, MARKETING PROMOTIONS TRACKING, and CUSTOMER SERVICE INQUIRIES)
- MATRIX: Identifying the Dimensions
  - Can be done after the list of fact tables was created
  - Deciding if a dimension is reasonable to consider for a fact table, regardless of whether there is a clear example of a production data source that ties the dimension into the data

The Intersection Matrix

FACT TABLES \ DIMENSIONS	Calendar	Customer	Service	Rate Category	Local Service Provider	Calling Party	Called Party	Long Distance Provider	Employee	Location	Equipment Type	Weather	Account Status
Customer Billing	X	X	X	X	X			X		X			X
Service Orders	X	X	X		X			X	X	X	X	X	X
Trouble Reports	X	X	X		X	X		X	X	X	X	X	X
Customer Inquiries	X	X	X	X	X	X		X	X	X		X	X
Billing Call Detail	X	X	X	X	X	X	X	X		X	X	X	X
Customer Relatinship Management	X	X	X	X	X	X	X	X	X	X	X	X	X
Customer Profit	X	X	X	X	X	X	X	X	X	X	X	X	X

Building Dimensional Models

- Using the four-step method for designing the fact tables
- Managing the dimensional method project
  - Communicating the design between the people involved with the DW project
  - Graphical tools
    - Data Warehouse Bus Architecture Matrix
    - Fact Table Diagram
    - Fact Table Detail
    - Dimension Table Detail

Data Warehouse Bus Architecture Matrix

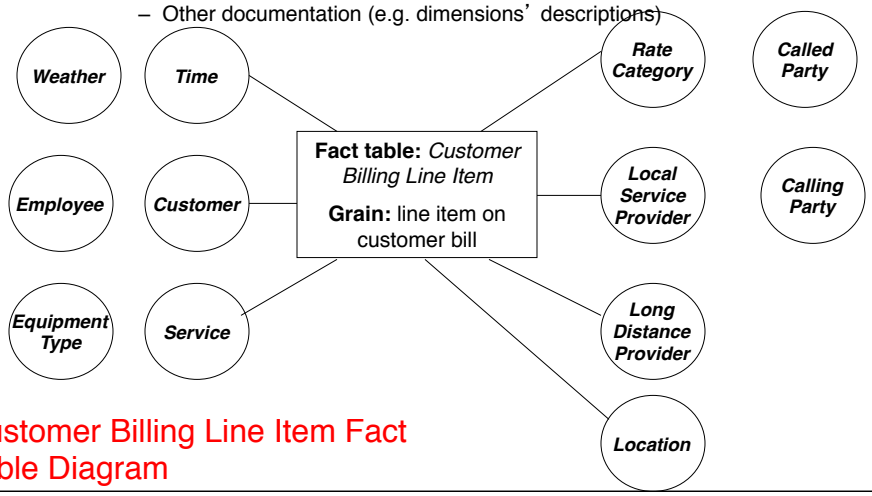
- Useful as a high-level introduction to the design
- Gives each audience a view of what the eventual scope of the data warehouse will become

Telecom. Company Data Warehouse Bus Architecture Matrix

DIMENSIONS / FACT TABLES	Time	Customer	Service	Rate Category	Local Service Provider	Calling Party	Called Party	Long Distance Provider	Employee	Location	Equipment Type	Weather
Customer Billing	X	X	X	X	X			X		X		
Service Orders	X	X	X		X			X	X	X	X	X
Trouble Reports	X	X	X		X	X		X	X	X	X	X
Customer Inquiries	X	X	X	X	X	X		X	X	X		X
Billing Call Detail	X	X	X	X	X	X	X	X		X	X	X
Customer Relatinship Management	X	X	X	X	X	X	X	X	X	X	X	X
Customer Profit	X	X	X	X	X	X	X	X	X	X	X	X

Fact Table Diagram

- Logical diagram of each completed fact table which contains:
  - Fact table name
  - Fact table grain
  - The dimensions to which the fact table is connected
  - All other dimensions proposed for the fact table
  - Other documentation (e.g. dimensions' descriptions)



Customer Billing Line Item Fact Table Diagram

Fact Table Detail

- Provides a complete list of all facts available through the fact table
  - Actual facts
  - Derived facts (presented through DBMS views)
  - Other documentation including fact descriptions and aggregation rules (depicting which facts are non-additive or semi-additive across which dimensions)

Fact Table: ***Customer Billing Line Item***

TimeKey
CustomerKey
ServiceKey
RateCategoryKey
LocalServiceProviderKey
LongDistanceProviderKey
LocationKey

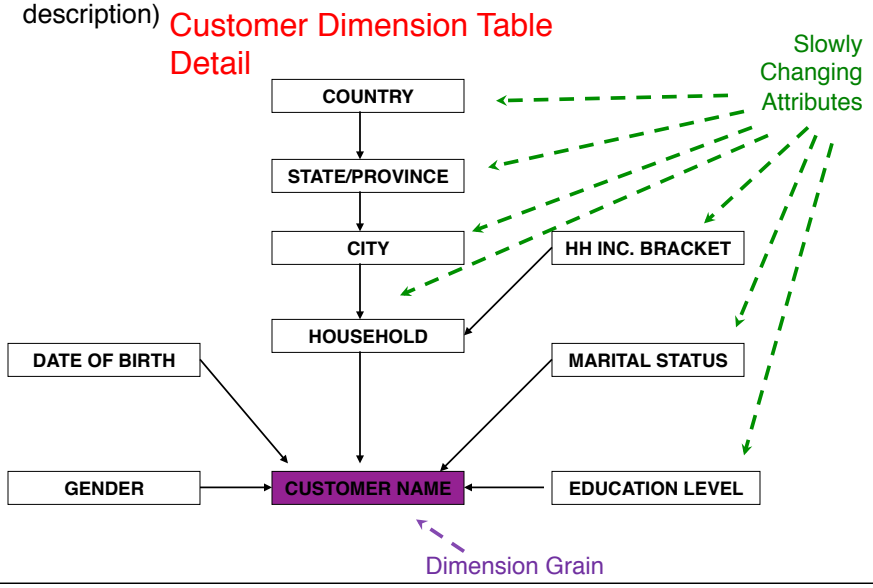
line_item_amount
line_Item_quantity
line_item_discount*
average_line_item_price*

* denotes derived fact

Customer Billing Line Item Fact Table Detail

Dimension Table Detail

– Shows individual attributes (and their approximate cardinality) within a single dimension as well as other documentation (e.g. attribute description)



Sources for Fact and Dimension Tables

- Selecting Data Sources
  - Formal Sources – supported by IS
  - Informal Sources – importing data from PC databases, from newly designed processes, etc.
- Understanding Data Sources
  - Answering the questions
    - What are the sources?
    - What is their availability?
    - Who is responsible for them?
  - Understanding the details of the sources
- Data Source Definition
  - Source Name
  - Business Owner
  - IS Owner
  - Platform
  - Location
  - Description

Selecting Data Sources

- Issues to consider:
  - Data Accessibility
    - In case of more than one equally feasible candidate data source, choose the on with easiest access
  - Feed Longevity
    - Do not choose sources scheduled for phasing out
  - Data Accuracy
  - Project Scheduling
    - In case data sources switch to a new system while being used as DW data sources, DW must switch to receiving data from the new system
- Sources that were considered, but not selected, must be documented as well

Customer Matching

- Classical Data Warehousing Problem:
  - **Combining data** about customers (or some other entity for that matter) from **disparate sources** (A.K.A. de-duplicating, householding)
  - The problem is often not trivial
  - Sophisticated commercial software available for this special purpose

Additional Concepts

- Estimate table sizes
  - Should be done
  - Largest theoretical size: multiply number of rows for each dimension
  - For more accuracy: examine data sources
  - Must consider future as well
- Designing for aggregation
  - Aggregates stored in their own fact tables separate from the base-level data
  - Each stored aggregation level occupies its own fact table