Data Modeling and Databases
Ch 7: Schemas

Gustavo Alonso, Ce Zhang
Systems Group
Department of Computer Science
ETH Zürich
Database schema

- A Database Schema captures:
  - The concepts represented
  - Their attributes
  - The constraints and dependencies over all attributes

- You now know everything needed to build good schemas:
  - ER modeling as the basis for capturing concepts
  - Functional dependencies for constraining the data
  - Normal forms for removing redundancy and anomalies
  - Algorithms for decomposing and deriving normalized tables
OLTP

- On-line Transaction Processing
- Refers to workloads with updates (UPDATE, INSERT, DELETE), often online (with response time requirements), and a high volume of, typically, small transactions (queries return few records, updates affect few records)
- Integrity is important, OLTP workloads are commonly run over 3NF schemas and many tables
- Examples: banking, e-shopping, sales, etc.
Example OLTP: TPC-C


The Company portrayed by the benchmark is a wholesale supplier with a number of geographically distributed sales districts and associated warehouses. As the Company's business expands, new warehouses and associated sales districts are created. Each regional warehouse covers 10 districts. Each district serves 3,000 customers. All warehouses maintain stocks for the 100,000 items sold by the Company.
TPC-C Schema
TPC-C Schema

TPC-C has 9 tables:
OLAP

- On-Line Analytical Processing
- Refers to workloads with heavy, complex queries (retrieving large number of records, often involving aggregation), data is updated but typically in batches (daily sales, weekly transactions, etc.)
- OLAP workloads commonly use de-normalized schemas, with some standard approaches such as star or snowflake schemas. For data mining, data cubes are used
- Examples: marketing analysis, reporting, data analysis
Example OLAP: TPC-H

- TPC-H queries address:
  - Pricing and promotions;
  - Supply and demand management;
  - Profit and revenue management;
  - Customer satisfaction study;
  - Market share study;
  - Shipping management.
Star schemas (SS benchmark)

- Star schemas are used in OLAP and data mining.
- Often referred as fact table + dimension tables
- Used when design centered around a very large collection of facts (sales, transactions, events, etc.)
Example query (SS benchmark)

- The query is intended to provide revenue volume for lineorder transactions by customer nation and supplier nation and year within a given region, in a certain time period.

- **Q3:**
  
  ```sql
  select c_nation, s_nation, d_year, sum(lo_revenue) as revenue from customer, lineorder, supplier, date where lo_custkey = c_custkey and lo_suppkey = s_suppkey and lo_orderdate = d_datekey and c_region = 'ASIA' and s_region = 'ASIA' and d_year >= 1992 and d_year <= 1997 group by c_nation, s_nation, d_year order by d_year asc, revenue desc;
  ```
Normalization and schemas

- Normalization tries to avoid redundancy and anomalies. Why do de-normalized schemas work?
- OLAP databases are not usually populated by hand or individual transactions, data is loaded in batches and only periodically.
- It is during the data loading process where the constraints and anomalies are controlled => responsibility is moved from the schema to the application doing the data loading.
- This makes sense since the data loading can perform many transformations over the original data (see data cubes).
Snowflake schema

- In star schemas, the dimension tables and the fact table are not normalized.
- A snowflake schema is a star schema where the dimension tables are normalized (some or all of them). Normalization is applied to low cardinality attributes to remove redundancy.
TPC-DS

Schema Overview

The TPC DS schema models the sales and sales returns process for an organization that employs three primary sales channels: stores, catalogs, and the Internet. The schema includes seven fact tables:

- A pair of fact tables focused on the product sales and returns for each of the three channels
- A single fact table that models inventory for the catalog and internet sales channels.

In addition, the schema includes 17 dimension tables that are associated with all sales channels.
A modern decision database

Operational Systems
- Store
- WEB
- Catalog
- Inventory
- Promotions

Refresh Process
Set of Files
ETL

Ad hoc and Reporting Queries
DSS Database TPC-DS

User Queries
Reports
TPC-DS

- Snow-storm schema: TPC-DS consists of multiple snowflake schemas, which are multiple star schema with dimensions linking to dimensions

- Logical schema
  - The actual implementation is flexible as long as views in accordance with the specification are provided

- Large number of tables (26)
  - Large number of columns per table (38)
  - Multiple fact tables to enable joins between large tables (fact to fact joins)
TPC-DS: Store sales
TPC-DS: promotions

- Promotions
  - Catalog Returns
    - Catalog Sales
  - Inventory
    - Web Sales
  - Web Returns
    - Web Sales
  - Store Returns
    - Store Sales
TPC-DS: Store Channel

- store
- store_sales
- item
- store_returns
- reason
- date
- time
- customer
- customer_address
- customer_demographics
- household_demographics
Modern trends

The trend today:

- Main memory databases (entire database or entire working set in main memory, no I/O for query processing)
- Column stores rather than row stores (at the physical design level not at the logical design level)
- OLTP and OLTP in one system
- Heavy specialization for some applications (de-normalization)
Data cubes

- Data cubes are used for analysis and reporting.
- They include pre-aggregated data across several dimensions and granularities
- Support operations such as:
  - Slicing: selecting over one dimension (sales of a product by year and region)
  - Dicing: selection over the dimensions of the original cube (sales in June, July, August, for a range of products, in shops within a particular region)
  - Drill up and down: group-by at different granularities (sales by shop, sales in a shop by department, sales by product in a department)
  - Roll up: aggregation along a dimension (sales by week, month, year)
Data cubes examples