Last week's leftover

- Review – steps in data modeling
- Assign keys
- Data normalization

1

Steps in Data Modeling

- Identify entities
- Identify attributes
- Apply naming conventions
- Identify relationships
- *Apply data model patterns*
- Assign keys
- Normalize to reduce data redundancy
- Denormalize to improve performance

This steps are performed iteratively until satisfactory.

2

Normalization & Denormalization

- Normalization is the process of reducing redundant data
  - Prevents data inconsistency and update anomalies
  - Avoids storing identical data in multiple tables
- Normalization slightly degrades database performance
  - More impacts on reads
  - Little impact on writes, which tend to be the bottleneck anyway
  - Denormalization is common on read-only databases on which high performance is required (e.g. Web read-only databases)
  - Database design and disk configuration (architecture) interact

3

Some Definitions

- Row or record: a fixed tuple (set) of attributes (fields) that describes an instance of an entity
- Key: a unique identifier for a row in a table, used to select the row in queries. It can be composed of several fields.
- Non-key: all the other fields in the row
- Entity: Object defined in system model about which data is stored in the database. A table in a relational database.

4

Source: MIT 1.264 lecture notes

Data normalization references

- http://en.wikipedia.org/wiki/Database_normalization
- http://support.microsoft.com/kb/283878

5

Five Normal Forms

1. All occurrences of an entity must contain the same number of attributes. No lists.
2. All non-key fields must be a function of the key.
3. All non-key fields must not be a function of other non-key fields.
4. *A row should not contain two or more independent multi-valued facts about an entity.*
5. *A record cannot be reconstructed from several smaller record types.*

6

Source: MIT 1.264 lecture notes

1st Normal Form

- All occurrences of an entity must contain the same number of attributes. No lists.
  - a) no repeating groups
  - b) entries be uniquely identifiable (i.e. must have PK)

Customer	Day	Amount
Pete	Monday	19.00 -28.20
Pete	Wednesday	-84.00
Sarah	Friday	100.00 150.00 -40.00

Customer	Day	Amount
Pete	Monday	19.00
Pete	Monday	-28.20
Pete	Wednesday	-84.00
Sarah	Friday	100.00
Sarah	Friday	150.00
Sarah	Friday	-40.00

7

http://en.wikipedia.org/wiki/First_normal_form

Customer	Day	Amount
Pete	Monday	19.00
Pete	Monday	-28.20
Pete	Wednesday	-84.00
Sarah	Friday	100.00
Sarah	Friday	150.00
Sarah	Friday	-40.00

Customer	Day	Amount
Pete	Monday	19.00
Pete	Monday	19.00

Person	Foods Not Eaten
Jim	Liver, Goat's cheese
Alice	Broccoli
Norman	Pheasant, Liver, Peas

Transaction ID	Customer	Day	Amount
1	Pete	Monday	19.00
2	Pete	Monday	19.00

8

2nd Normal Form (1/2)

Part	Warehouse	Quantity	WarehouseAddress
42	Boston	2000	24 Main St.
333	Boston	1000	24 Main St.
390	New York	3000	99 Broad St.

- All non-key fields must be a function of the full key
  - Example that violates second normal form:
    - Key is Part + Warehouse
    - Someone found it convenient to add Address, to make a report easier
    - WarehouseAddress is a fact about Warehouse, not about Part
  - Problems:
    - Warehouse address is repeated in every row that refers to a part stored in a warehouse
    - If warehouse address changes, every row referring to a part stored in that warehouse must be updated
    - Data might become inconsistent, with different records showing different addresses for the same warehouse
    - If at some time there were no parts stored in the warehouse, there may be no record in which to keep the warehouse's address.

9

Source: MIT 1.264 lecture notes

2nd Normal Form (2/2)

- Solution
  - Two entity types: Inventory, and Warehouse
  - Advantage: solves problems from last slide
  - Disadvantage: If application needs address of each warehouse stocking a part, it must access two tables instead of one. This used to be a problem but rarely is now.

Part	Warehouse	Quantity
42	Boston	2000
333	Boston	1000
390	New York	3000

Warehouse	WarehouseAddress
Boston	24 Main St.
New York	99 Broad St.

10

Source: MIT 1.264 lecture notes

3rd Normal Form (1/2)

Employee	Department	DepartmentLocation
234	Finance	Boston
223	Finance	Boston
399	Operations	Washington

- Non-key fields cannot be a function of other non-key fields
  - Example that violates third normal form
    - Key is employee
    - Someone found it convenient to add department location for a report
    - Department location is a function of department, which is not a key
  - Problems:
    - Department location is repeated in every employee record
    - If department location changes, every record with it must be changed
    - Data might become inconsistent
    - If a department has no employees, there may be nowhere to store its location

11

Source: MIT 1.264 lecture notes

3rd Normal Form (2/2)

- Solution
  - Two entity types: Employee and department

Employee	Department
234	Finance
223	Finance
399	Operations

Department	DepartmentLocation
Finance	Boston
Operations	Washington

12

Source: MIT 1.264 lecture notes

Moral of Data Modeling Story

- Systems are ephemeral (short lived)
- Data is permanent
- If you mess up a system, you rewrite it and it's fixed
- If you mess up the data, it's usually irretrievable
- Real business have subtle business rules
  - Care in data modeling and business rules is needed to achieve good data quality
  - Care in data normalization is needed to preserve data quality

13

Source: MIT 1.264 lecture notes

Reference

- http://support.microsoft.com/kb/283878
- http://phlonx.com/resources/nf3/
- www.agiledata.org/essays/dataModeling101.html
- http://ocw.mit.edu/
- http://www.utexas.edu/its/windows/database/datamodeling/dm/design.html
- http://db.grussell.org/section004.html

- Free DB design tool:
- http://www.fabforce.net/dbdesigner4/

14

Lecture 8

Database development, SQL (I)

15

Basic SQL Operations

- Database Creation & Connection
- Table & Index Creation
- Changing/Deleting Objects
- Basic Data Manipulation
  - Data insertion
  - Data retrieval
  - Data deletion and updating
- Advanced Select

16

Introduction to SQL

- SQL
  - Structured Query Language
  - an ANSI standard computer language for accessing and manipulating databases.
  - Retrieve, insert, delete, update data against a database
  - CRUD – Creation, Retrieval, Updating, and Deletion.
- Not a complete language like Java, C++, …
  - SQL is a sub-language of about 30 statements
  - Usually embedded in other languages or tool for database access
  - Portable across operating systems
  - Somewhat portable among DBMS vendors

17

Source: MIT 1.264 lecture notes

Database Creation & Connection

- Create a new database
  - CREATE DATABASE *databaseName;*
- Use a database
  - USE *databaseName*;
- Show available databases; (MySQL)
  - SHOW DATABASES;

18

Table & Index Creation

- Create a table
  - CREATE TABLE *tableName* (
    field1 datatype [(length)] [NULL, NOT NULL],
    field2 datatype [(length)] [NULL, NOT NULL],
    …);
  - DESCRIBE *tableName; (MySQL)*
- Create an index
  - CREATE [UNIQUE] INDEX *indexName* ON *tableName (columnName)*

19

Advanced Table Creation

CREATE TABLE *tableName* (
field1 datatype [(length)] [NULL, NOT NULL]
[PRIMARY KEY] [UNIQUE]
[DEFAULT …], [CHECK …]
…);

- PRIMARY KEY: specify the column is used as a primary key
- UNIQUE: the values in the column should be unique
- DEFAULT: specify the default value if not defined
- CHECK: constrains the value of the column
  - CHECK (title_id LIKE '[A-Z][A-Z][0-9][0-9][0-9][0-9]')

20

Index

- Why use index?
  - Fast data retrieval and sorting
  - Unique index ensures the values of the indexed column are unique
- Why not use index?
  - Performance penalty on data insertion, deletion, update
- When?
  - On columns with frequent retrievals
  - On columns used to join other tables (to be covered)
  - On columns accessed in sorted orders

21

Changing and Deleting Objects

- Change table definition
  - ALTER TABLE `tableName` [ADD `fieldName datatype` …][DROP `fieldName`]
- Delete a table
  - DROP TABLE `tableName`
- Delete a database
  - DROP DATABASE `databaseName`
- Delete an index
  - DROP INDEX `tableName.indexName`
  - DROP INDEX `indexName on tableName` (MySQL)

22

Data Insertion

- Inserting data
  - INSERT INTO `tableName` *`[(col1, col2, …)]`* VALUES `(val1 [, val2, …])`;
  - Not all data attributes (field values) need to present
  - DBMS will use default values for absent data fields
- Show data in a table
  - SELECT * FROM `tableName`

- *Demonstration with MS-SQL through SQL Server Management Studio Express.*

23

Review

- SQL – Structured Query Language
  - Database Creation & Connection
  - Table & Index Creation
  - Changing Deleting Objects
  - Basic Data Manipulation
    - Data insertion
    - Data retrieval

24

Microsoft SQL Server 2008 Express

- Home
  - http://www.microsoft.com/sql/default.mspx
- Express version download
  - **Microsoft® SQL Server® 2008 Express with Tools**
  - http://www.microsoft.com/download/en/details.aspx?displaylang=en&id=22973

25

MySQL References

- Manual Download
  http://dev.mysql.com/doc/
- Software Download (5.x is now recommended)
  http://dev.mysql.com/downloads/

26

SQL References

- Reference book:
  - The Practical SQL Handbook, Judith et al. Addison-Wesley, ISBN 0201447878
- Reference web sites:
  - http://www.w3schools.com/sql/default.asp
  - http://www.1keydata.com/sql/sql.html
  - http://sqlzoo.net/
  - http://sqlcourse.com/
  - http://www.firstsql.com/tutor.htm

27

Google: sql tutorial

Supplement

28

4th Normal Form (1/4)

Employee	Skill	Language
Brown	Cook	English
Smith	Type	German

- A row should not contain two or more independent multi-valued facts about an entity.
  - Example that violates fourth normal form:
    - An employee may have several skills and languages
  - Problems
    - Uncertainty in how to maintain the rows. Several approaches are possible and different programmers may take different approaches, as shown on next slide

29

Source: MIT 1.264 lecture notes

4th Normal Form (2/4)
Problem 1

- Disjoint format. Effectively same as 2 entity types.
  - Blank fields ambiguous. Blank skill could mean:
    - Person has no skill
    - Attribute doesn't apply to this employee
    - Data is unknown
    - Data may be found in another record (as in this case)
  - Programmers will use all these assumptions over time, as will data entry staff and users

Employee	Skill	Language
Brown	Cook	
Brown	Type	
Brown		French
Brown		German
Brown		Greek

30

Source: MIT 1.264 lecture notes

4th Normal Form (3/4)
Problem 2

- Cross product format.
  - Repetitions: updates must be done to multiple records and there can be inconsistencies
  - Insertion of a new skill may involve looking for a record with a blank skill, inserting a new record with possibly a blank language or skill, or inserting a new record pairing the skill with some or all of the languages.
  - Deletion is worse: It means blanking a skill in one or more records, and then checking you don't have 2 records with the same language and no skill, or it may mean deleting one or more records, making sure you don't delete the last mention of a language that should not be deleted.

Employee	Skill	Language
Brown	Cook	French
Brown	Cook	German
Brown	Cook	Greek
Brown	Type	French
Brown	Type	German
Brown	Tye	Greek

Source: MIT 1.264 lecture notes

4th Normal Form (4/4)
Solution

- Two entity types: Employee-skill and employee-language
- Note that skills and languages may be related, in which case the starting example was ok:
  - If Smith can only cook French food, and can type in French and Greek, then skill and language are not multiple independent facts about the employee, and we have not violated fourth normal form.

Employee	Skill
Brown	Cook
Brown	Type

Employee	Language
Smith	French
Smith	German
Smith	Greek

Source: MIT 1.264 lecture notes

5th Normal Form (1/4)

Agent	Company	Product
Smith	Ford	Car
Smith	GM	Truck

- A record cannot be reconstructed from several smaller record types.
- Example:
  - Agents represent companies
  - Companies make products
  - Agents sell products
- Most general case (allow any combination):
  - From the above table, however, Smith does not sell Ford trucks nor GM cars

Source: MIT 1.264 lecture notes

5th Normal Form (2/4)

- A variation is where the problem occurs–If an agent sells a certain product and she represents the company, then she sells that product for that company.

Agent	Company	Product
Smith	Ford	Car
Smith	Ford	Truck
Smith	GM	Car
Smith	GM	Truck
Jones	Ford	Car

- We can reconstruct all true facts from 3 tables instead of the 1:

Agent	Company
Smith	Ford
Smith	GM
Jones	Ford

Agent	Product
Smith	Car
Smith	Truck
Jones	Car

Company	Product
Ford	Car
Ford	Truck
GM	Car
GM	truck

Source: MIT 1.264 lecture notes

5th Normal Form (3/4)

- Problems with the 1 table form
  - Facts are recorded multiple times. E.g., the fact that Smith sells cars is recorded twice. If Smith stops selling cars, there are 2 rows to update and one will be missed.
  - Size of this table increases multiplicatively, while the normalized tables increase additively. With big operations, this is a big difference.
    - 100,000 x 100,000 is a lot bigger than 100,000 + 100,000

Source: MIT 1.264 lecture notes

5th Normal Form (4/4)

- An example with a subtle set of conditions

Agent	Company	Product
Smith	Ford	Car
Smith	Ford	Truck
Smith	GM	Car
Smith	GM	Truck
Jones	Ford	Car
Jones	Ford	Truck
Brown	Ford	Car
Brown	GM	Car
Brown	Toyota	Car
Brown	Toyota	Bus

Agent	Company
Smith	Ford
Smith	GM
Jones	Ford
Brown	Ford
Brown	GM
Brown	Toyota

Agent	Product
Smith	Car
Smith	Truck
Jones	Car
Jones	Truck
Brown	Car
Brown	Bus

Company	Product
Ford	Car
Ford	Truck
GM	Car
GM	Truck
Toyota	Car
Toyota	Bus

Jones sells cars and GM makes cars, but Jones does not represent GM
Brown represents Ford and Ford makes trucks, but Brown does notsell trucks
Brown represents Ford and Brown sells buses, but Ford does not make buses

Source: MIT 1.264 lecture notes