Data Science & Analytics

Hiren Deliwala & Dr. Jian ZHANG

Data Modeling & Normalization

Database Development Process

1. Collect user/business requirements.
2. Develop the conceptual E-R Model based on the user/business requirements.
3. Convert the E-R Model to a set of relations in the (logical) relational model
4. **Normalize the relations to remove any anomalies.**
5. Create a table for each normalized relation in a relational database management system.

Purpose of Normalization

- Normalization
  - is a process for assigning attributes to entities.
  - reduces data redundancies and helps eliminate the data anomalies – deletion, insertion & update anomalies
- Normalization works through a series of stages called normal forms:
  - First normal form (1NF)
  - Second normal form (2NF)
  - Third normal form (3NF)
  - Fourth normal form (4NF)
- The highest level of normalization is not always desirable.

Example

- Purchase Order

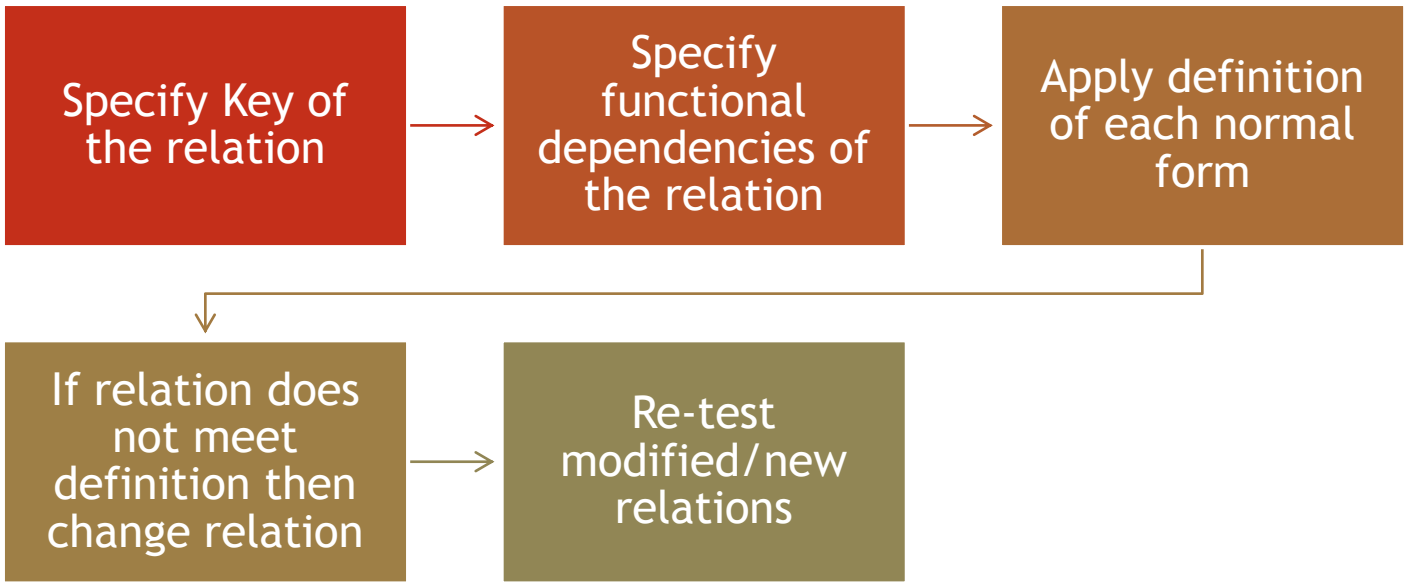
ER Model

- PO Header (PO_Number, PODate, Vendor, Ship_To, etc)
- PO Line Items (PO_Number, LineItemNum, PartNum, Description, Price, Qty)

Challenges

- Update, Insert & Delete Anamolies

Normalization Process



Keys

- A key is a data item that allows us to uniquely identify a row in a relation.
- A key functionally determines a row.
  - The Key → All other attributes

Functional Dependencies

- Describes a relationship between *attributes* within a single relation.
- An attribute is *functionally dependent* on another if we can use the value of one attribute to determine the value of another.
- Example: Employee_Address is functionally dependent on Employee_Number because Employee_Number can be used to uniquely determine the value of Employee_Address.
- Use Arrow symbol → to indicate a functional dependency.
- X → Y is read *X functionally determines Y*
- X is known as *determinant*; X determines Y

Key Concepts

- Determinants
  - A determinant in a database table is any attribute that you can use to determine the values assigned to other attribute(s) in the same row.
- Candidate Key
  - A Candidate Key can be any column or a combination of columns that can **qualify as unique key in database**. There can be multiple Candidate Keys in one table. Each Candidate Key can qualify as Primary Key.
- Primary Key
  - A Primary Key is a column or a combination of columns that **uniquely identify a record**. Only one Candidate Key can be Primary Key.
- Foreign Key
  - A foreign key (FK) is a column or combination of columns that is used to establish and enforce a link between the data in two tables or entities

First Normal Form (1 NF)

- 1NF Definition
  - Each attribute (column) value must be a single value only.
  - All values for a given attribute (column) must be of the same type.
  - Each attribute (column) name must be unique.
  - The order of attributes (columns) is insignificant
  - No two rows in a relation can be identical.
  - The order of the rows is insignificant.

Normalization Example

Company	Ticker	Location	Date	Stock Price
Apple	AAPL	Cupertino, CA	1/1/2013	300
Apple	AAPL	Cupertino, CA	1/1/2014	500
Apple	AAPL	Cupertino, CA	1/1/2015	5000
Chipotle	CMG	Denver, CO	1/1/2013	400
Chipotle	CMG	Denver, CO	1/1/2014	500
Chipotle	CMG	Denver, CO	1/1/2015	750

Ticker & Date can uniquely determine the company, location & stock price.

1NF Anomalies

- Insertion: Cannot add new company if its stock price is not recorded yet
- Update: If company moves location we need to update multiple rows
- Deletion: Deleting the stock info will delete the company info

Second Normal Form (2 NF)

- 2NF Definition
  - A relation is in second normal form (2NF) if all of its non-key attributes are dependent on all of the *key*.
  - Another way to say this: A relation is in second normal form if it is free from partial-key dependencies
  - Relations that have a single attribute for a key are automatically in 2NF.
  - This is one reason why we often use artificial identifiers (non-composite keys) as keys.

Second Normal Form (2 NF)

- Functional Dependencies
  - Ticker, Date → Company, Location, Stock Price
  - Ticker → Company, location
- Ticker, Date is our key
- Ticker → location
  - violates the rule for 2NF in that a part of our key determines a non-key attribute
  - Partial key dependency
- Split in to two new relations
  - Company (company, ticker, location)
  - Stock Prices (ticker, date, stock price)

1NF Anomalies Resolved

- Insertion: Cannot add new company if its stock price is not recorded yet
  - Can now add company without a stock price
- Update: If company moves location we need to update multiple rows
  - Only one row needs to be changed
- Deletion: Deleting the stock info will delete the company info
  - Can delete stock info without impacting company information

2NF Anomalies

- Insertion: Cannot add new company if it does not have a ticket symbol
- Update: If company moves location we need to update multiple rows (rows with different ticket symbols)
- Deletion: Delisting the company from the stock market will remove the company location information too

Third Normal Form (3 NF)

- 2NF Definition
  - A relation is in third normal form (3NF) if it is in second normal form and it contains no *transitive dependencies*.
  - Consider relation R containing attributes X, Y and Z. R(X, Y, Z)
  - If X → Y and Y → Z then X → Z
- **Transitive Dependency:**
  - FD1: Ticker → Company
  - FD2: Company → Location
  - So therefore: Ticker → Location
- Solution – split into two new relations
  - Stock_Ticker (company, ticker)
  - Company_Location (company, location)

2NF Anomalies Resolved

- Insertion: Cannot add new company if it does not have a ticket symbol
  - Resolved – since you can add new company without a ticker
- Update: If company moves location we need to update multiple rows (rows with different ticket symbols)
  - Resolved – only one record needs to be changed
- Deletion: Delisting the company from the stock market will remove the company location information too
  - Resolved – only one record needs to be deleted

Boyce-Codd Normal Form (BCNF)

- Most 3NF relations are also BCNF relations.
- A relation is in BCNF if every determinant is a candidate key
- Example
  - Funds Example

Normalize a relation into BCNF

- List all of the determinants.
- See if each determinant can act as a key (candidate keys).
- For any determinant that is *not* a candidate key, create a new relation from the functional dependency. Retain the determinant in the original relation.

Fourth Normal Form (4NF)

- A relation is in fourth normal form if it is in BCNF and it contains no *multivalued dependencies*.
- **Multivalued Dependency**: A type of functional dependency where the determinant can determine more than one value.

Intuitive Normalization

- **1NF** Tables represent entities
  - Each entity has well defined unique key
- **2NF** Each table represents only one entity
  - No partial key dependencies
- **3NF** Tables do not contain attributes from embedded entities
  - No transitive relationships
- **BCNF**
  - Every determinant is a candidate key
- **4NF** Triple relationships should not represent a pair of dual relationships
  - no multivalued dependencies

Normalization Challenges

- Normalization splits database information across multiple tables.
- Performance Impact due to Joins

De-Normalization

- *De-normalize* the relations to achieve a performance improvement.
- De-normalization presents a trade-off between performance and modification anomalies / data redundancy.
- Query speed is improved at the expense of more complex or problematic data manipulation for updates, deletions and insertions.

Normalization Examples

Patient #	Surgeon #	Surg. date	Patient Name	Patient Addr	Surgeon	Surgery	Postop drug	ug side effec
1111	145 311	Jan 1, 1995; June 12, 1995	John White	15 New St. New York, NY	Beth Little Michael Diamond	Gallstones s removal; Kidney stones removal	Penicillin, none-	rash none
1234	243 467	Apr 5, 1994 May 10, 1995	Mary Jones	10 Main St. Rye, NY	Charles Field Patricia Gold	Eye Cataract removal Thrombosis removal	Tetracycline none	Fever none
2345	189	Jan 8, 1996	Charles Brown	Dogwood Lane Harrison, NY	David Rosen	Open Heart Surgery	Cephalosporin	none
4876	145	Nov 5, 1995	Hal Kane	55 Boston Post Road, Chester, CN	Beth Little	Cholecystectomy	Demicillin	none
5123	145	May 10, 1995	Paul Kosher	Blind Brook Mamaroneck, NY	Beth Little	Gallstones Removal	none	none
6845	243	Apr 5, 1994 Dec 15, 1984	Ann Hood	Hilton Road Larchmont, NY	Charles Field	Eye Cornea Replacement Eye cataract removal	Tetracycline	Fever

Normalization Examples

PlayerName	Team	TeamColor	Coachname	Coach#	Player#	PlayerPosition	TeamCaptain
Peyton	Denver	Orange	Fox	123	10	QB	Peyton
John	Denver	Orange	Fox	123	10	QB	Peyton
Jack	Denver	Orange	Fox	123	10	QB	Peyton
Richard	Denver	Orange	Fox	123	10	QB	Peyton

Normalization Examples

BookISBN	Title	First_Author	CoAuthors	Publisher	Pub_street	Pub_city	Pub_Contact	Pages	Price
123	Harry Potter 1	J K Rowling	Other	Bloomsbury	Wizard St	London	Dumbeldore	300	5
567	Harry Potter 2	J K Rowling	Other	Bloomsbury	Wizard St	London	Dumbeldore	500	10
109	Harry Potter 3	J K Rowling	Other	Bloomsbury	Wizard St	London	Dumbeldore	1000	15