# CSCI 4333 Section 2 Design of DB Systems

## 4/8/2024 (self - annotation)

**Introduction to Database Design**

K. Yue

**1. Introduction**

* Bad database/table designs result in unnecessary redundancy: redundancy with little or no benefit.
* However, redundancy can serve many purposes, and are employed frequently in computer science.
* Problems:
	1. Inefficient storage
	2. Anomaly: data inconsistency, loss of data integrity, difficulties in maintenance.

***Example:***

Consider the vastly simplified and poorly-designed relation/table:

Employee(EmpId, Name, DeptId, ManagerId).

Assumptions made:

1. Every employee works for only one department.
2. Every department has an unique DeptId.
3. Every employee is represented as a tuple in the Employee relation.
4. Every employee has an unique EmpId.
5. Every department has one manager, identified by the ManagerId

Thus, EmpId is a candidate key (CK). An instance of Employee:

|  |  |  |  |
| --- | --- | --- | --- |
| **EmpId** | **Name** | **DeptId** | **ManagerId** |
| 101 | Lady Gaga | *D123* | *M10* |
| 122 | Brad Pitts | *D123* | *M10* |
| 140 | Lebron James | *D123* | *M10* |
| 155 | Narendra Modi | D222 | M21 |
| 167 | Jennifer Lopez | D222 | M21 |
| 311 | John Smiths | D300 | M33 |

**Problem:**

* Unnecessary redundancy: e.g., the fact that manager M10 manages Department D123 is stored in three rows.

**1.1 Update Anomaly:**

(a) M44 becomes the new manager of Department D123

* Inefficiency in update: needs to update all three rows in this example.
* Potential (logical data) inconsistency.

|  |  |  |  |
| --- | --- | --- | --- |
| **EmpId** | **Name** | **DeptId** | **ManagerId** |
| 101 | Lady Gaga | D123 | *M10 -> M44* |
| 122 | Brad Pitts | D123 | *M10 -> M44* |
| 140 | Lebron James | D123 | *M10 -> M44* |
| 155 | Narendra Modi | D222 | M21 |
| 167 | Jennifer Lopez | D222 | M21 |
| 311 | John Smiths | D300 | M33 |

SQL solution: it works but is not efficient.

UPDATE Employee
SET ManagerId = 'M44'
WHERE DeptId = 'D123';

(b) Jennifer Lopez is reassigned to work for Department D300:

* Need to know the manager of D300.
* Potential inconsistency.

The table may become:

|  |  |  |  |
| --- | --- | --- | --- |
| **EmpId** | **Name** | **DeptId** | **ManagerId** |
| 101 | Lady Gaga | D123 | M10 |
| 122 | Brad Pitts | D123 | M10 |
| 140 | Lebron James | D123 | M10 |
| 155 | Narendra Modi | D222 | M21 |
| 167 | Jennifer Lopez | **D300** | M21 *(may not be updated to M33)* |
| 311 | John Smiths | D300 | M33 |

The intuitive SQL command:

UPDATE Employee
SET DeptId = 'D300'
WHERE Name = 'Jennifer Lopez';

will produce inconsistent result, as shown in the table above.

One needs to update both DeptId and ManagerId. However,

UPDATE Employee
SET DeptId = 'D300',
     ManagerId = (SELECT DISTINCT ManagerId FROM Employee WHERE DeptId = 'D300')
WHERE Name = 'Jennifer Lopez';

will not work in MySQL as one cannot include a SELECT clause on the same table in the SET clause of an UPDATE statement.

A possible solution using a session variable, @M\_NO:

SELECT DISTINCT ManagerId INTO @M\_NO
FROM Employee WHERE DeptId = 'D300';

UPDATE Employee
SET DeptId = 'D300',
     ManagerId = @M\_NO
WHERE Name = 'Jennifer Lopez';

**1.2 Insertion Anomaly:**

It is not possible creating a new Department D777, with manager M40 but no employee working for it yet. This is because, as the PK, EmpId cannot be null.

|  |  |  |  |
| --- | --- | --- | --- |
| **EmpId** | **Name** | **DeptId** | **ManagerId** |
| 101 | Lady Gaga | D123 | M10 |
| 122 | Brad Pitts | D123 | M10 |
| 140 | Lebron James | D123 | M10 |
| 155 | Narendra Modi | D222 | M21 |
| 167 | Jennifer Lopez | D222 | M21 |
| 311 | John Smiths | D300 | M33 |
| ***????*** | **????** | **D777** | **M40 *(this row cannot be added)*** |

**1.3 Deletion Anomaly**

John Smiths no longer works here. Result: the information that M33 is the manager of Department D300 is also lost.

|  |  |  |  |
| --- | --- | --- | --- |
| **EmpId** | **Name** | **DeptId** | **ManagerId** |
| 101 | Lady Gaga | D123 | M10 |
| 122 | Brad Pitts | D123 | M10 |
| 140 | Lebron James | D123 | M10 |
| 155 | Narendra Modi | D222 | M21 |
| 167 | Jennifer Lopez | D222 | M21 |
| *~~311~~* | *~~John Smiths~~* | *~~D300~~* | *~~M33~~* |

**1.4 Decomposition**

A standard way of resolving unnecessary redundancy is by proper **decomposition**: breaking down a relation into two or more component relations.

***Example:*** the decomposition of the relation Employee into two relations:

1. Emp(EmpId, Name, DeptId)

|  |  |  |
| --- | --- | --- |
| **EmpId** | **Name** | **DeptId** |
| 101 | Lady Gaga | D123 |
| 122 | Brad Pitts | D123 |
| 140 | Lebron James | D123 |
| 155 | Narendra Modi | D222 |
| 167 | Jennifer Lopez | ~~D222~~ D300 |
| ~~311~~ | ~~John Smiths~~ | ~~D300~~ |

2. Department(DeptId, ManagerId)

|  |  |
| --- | --- |
| **DeptId** | **ManagerId** |
| D123 | ~~M10~~ M44 |
| D222 | M21 |
| D300 | M33 |
| D777 | M40 |

To obtain the original relation Employee(EmpId, Name, DeptId, ManagerId) from

Emp(EmpId, Name, DeptId)
Department(DeptId, ManagerId)

Relational algebra: using natural join, |x|.

Employee = Emp |x| Department

This decomposition is said to be a *lossless* decomposition.

SQL:

SELECT Emp.\*, Department.ManagerId
FROM Emp INNER JOIN Department ON (Emp.DeptId = Department.DeptId);

or

SELECT \*
FROM Emp NATURAL JOIN Department;

1. There is no loss of information: the definition of *lossless* decomposition.
2. No previously mentioned redundancy and anomaly.

**2. Methods for good database designs**

Two main tools:

1. Integrity Rules:  data constraint rules for avoiding data inconsistency.
2. Normal Forms:  a set of rules for designing good relation schemas.

**2.1 Integrity Rules**

* Most of the integrity rules are *application* dependent.
* Need to analyze the semantic of the applications to find out the integrity rules.
* These are known as *Database-Specific Integrity Rules*, or Application-Specific Integrity Rules.
	+ Specific to the given application
	+ Not universally applicable.

***Examples***: some database-specific integrity rules.

1. Student Id should be a seven-digit number.
2. Date of Birth should be greater than 1900.
3. The room number of Delta Building should start with a 'D'.
4. A student cannot take more than 24 credits in any semester.
5. A student must show proof of a meningitis shot before registration for the first semester.

**2.2 General Integrity Rules**

* They should be satisfied by *every* database.
* However, they are not necessarily enforced by the DBMS.
* Two general integrity rules in relational databases:
	1. *Entity Integrity Rule*: based on the concepts of primary keys (PK) and candidate keys (CK)
	2. *Referential Integrity Rule*: based on the concept of foreign keys (FK).

**2.3 Entity Integrity Rule**

* Entity Integrity: no*component* of a *candidate key* of a relation can have a null value.
* Meaning: In a relational database, a row that cannot be identified by ites CK will not be stored.

***Example:***

Employee(EmpId, Name, DeptId, Salary)

|  |  |  |  |
| --- | --- | --- | --- |
| **EmpId** | **Name** | **DeptId** | **Salary** |
| 101 | Lady Gaga | D123 | 55000000 |
| 122 | Brad Pitts | D123 | 10100000 |
| 140 | Lebron James | D123 | 50000000 |
| 155 | Narendra Modi | @: null | @ |
| *@* | Jennifer Lopez | D222 | 20000000***(this row cannot be added)*** |
| *@* | John Smiths | D300 | 70000***(this row cannot be added)*** |

* If EmpId is a candidate key, this Employee instance does not satisfy the entity integrity rule.
* Conversely, if we accept the relation instance above as valid, EmpId cannot be a candidate key.

**2.4 Referential Integrity Rule**

* Referential integrity rule: relations do not contain any *unmatched* *non-null foreign key* values.
* Any non-null value of a foreign key K must appear in the parent (referenced) relation where K is a candidate key.

***Example:***

EMP(EmpId, Name, DeptId)

|  |  |  |
| --- | --- | --- |
| **EmpId** | **Name** | **DeptId** |
| 101 | Lady Gaga | D123 |
| 122 | Brad Pitts | D123 |
| 140 | Lebron James | *@* |
| 155 | Narendra Modi | D222 |
| 167 | Jennifer Lopez | D222 |
| 311 | John Smiths | D300 |

Department(DeptId, ManagerId)

|  |  |
| --- | --- |
| **DeptId** | **ManagerId** |
| D123 | M10 |
| D222 | M21 |
| D300 | M33 |

* DeptId is a foreign key in the table EMP, referencing DeptId in the table Department. EMP(DeptId) references Department(DeptId).
* The referential integrity rule is satisfied.
* Note that DeptId may be null in EMP.

***Example:***

EMP(EmpId, Name, DeptId)

|  |  |  |
| --- | --- | --- |
| **EmpId** | **Name** | **DeptId** |
| 101 | Lady Gaga | D123 |
| 122 | Brad Pitts | D123 |
| 140 | Lebron James | D123 |
| 155 | Narendra Modi | D222 |
| 167 | Jennifer Lopez | D222 |
| 311 | John Smiths | D300 |
| **350** | **Bun Yue** | **D119*(should not be added)*** |

Department(DeptId, ManagerId)

|  |  |
| --- | --- |
| **DeptId** | **ManagerId** |
| D123 | M10 |
| D222 | M21 |
| D300 | M33 |

* The referential integrity rule is not satisfied in the example above.

Note:

1. In practical DBMS, pay attention to where the referential integrity rule is enforced.
2. For example, in MySQL, only the INNODB data/storage engine may enforce the referential integrity rule.
3. If the DBMS does not enforce the referential integrity rule, it will be the task of the DB developers to do so.

**Introduction to Functional Dependency and Normalization**

by K. Yue

**1. Introduction to Normalization**

* *Normal forms* (NF): a set of rules to detect poor database design so they can possibly improved by decompositions.
* Require the concepts of various kinds of *data dependency*: constraints or restrictions between two *sets*of *attributes*.
	1. Functional dependency (FD, most important: used to define NF, up to BCNF)
	2. Multi-valued dependency (MVD for defining 4NF)
	3. Join dependency (JD fordefining 5NF)
* Some common normal forms in ascending order: 1NF, 2NF, 3NF, BCNF, 4NF, 5NF, DKNF, 6NF.
* Higher normal forms are more restrictive and signify better designs.
* A relation in a higher normal form implies that it is in a lower normal form, but not *vice versa*.

***Example***

If a relation R1 is in 4NF, then R1 is also in BCNF, 3NF, 2NF and 1NF. Refer to the diagram below for R1, R2, and R3.

If a relation R2 is in 2NF, then

1. It is in 1NF,
2. it may or may not be in 3NF, and
3. it may or may not be in BCNF.

If a relation R3 is not in 3NF, then

1. It is not in BCNF.
2. It may or may not be in 1NF or 2NF.

The relations R1 and R3 in this example are depicted in the Venn's Diagram:



**General Overview**

* In general, the higher the normal form a relation is in, the better the logical design of the relation (in terms of avoiding redundancy and inconsistency).
* However, it may be necessary to consider other issues, especially performance.
	1. Higher normal forms may be achieved by *decompositions*, resulting in more relations.
	2. More joins may then be needed to provide the data for a query, decreasing performance.
* 1NF is usually assumed. However, there are relations not in 1NF in both theory and practice.
	1. For example, a composite data type may be supported by a specific DBMS vendor.
* 2NF is more interesting for *historical* reasons.
* 4NF or above involves data dependency that is hard to understand and use. They are usually not used in practice.
* Based on the concept of *functional dependencies* (FD), the most important normal forms are
	1. *3NF* and
	2. *BCNF* (*Boyce-Codd Normal Form*).

**2. Functional Dependencies (FD)**

* Each attribute in a database represents certain data information in the application.
* There can be dependency between data.
* For example, types of dependency and relationship between two *sets* of *attributes* can be:
	+ Many to one (0..\* to 0..1): Functional Dependency (FD)
	+ Many to many (0..\* to 0..\*): Multi-Valued Dependency (MVD)
* These relationships are the results of *assumptions* we made about the application requirements.

***Example:***

A student, {StudentId: 1233457} can be associated with *only one* GPA: {GPA: 3.00}
However, several students can have the same GPA: {GPA: 3.00}

StudentId -> GPA    (FD)
(many)         (one)

On the other hand, a student {StudentId: 1233457} can take*many* courses: {cid: CSCI4333}, {cid:CSCI 2315}, {cid: CSCI3331},...
Many students can take the same course: {StudentId: 1233457}, {StudentId: 2233490}, {StudentId: 3333457},... take {cid: CSCI4333}.

StudentId ->-> cid    (MVD: not covered in this course)
(many)            (many)

**2.1 Many to one relationships**

***Example***

For *many* applications, the relationship between SSN and FName are many to one in a relation R(..,SSN, FName, ...)

SSN        ->     FName
(many)                 (one)

**Assumptions:**

1. A SSN uniquely identifies a person.
2. Given a SSN, there can only be one first name associated with it (not allowing/storing alias, etc.)
3. Many different SSN's (persons) may have the same first name.
4. There should not be two tuples with the same SSN, but different FName in*all instances* of R.

**Terms:**

1. SSN uniquely *determines* FName.
2. FName is *functionally determined* by SSN.
3. There is a *functional dependency* SSN -> FName.
4. Hence, a functional dependency specifies a many to one relationship between two*sets*of attributes.

For example, the relation instance:

|  |  |  |  |
| --- | --- | --- | --- |
| **SSN** | **FName** | **PHONE** | ... |
| ***123456789*** | ***Peter*** | 123-456-7890 |   |
| ***123456789*** | ***Paul*** | 713-283-7066 |   |
| 222229999 | Mary | 713-283-7066 |   |

is *not* allowed if we assume SSN -> FName.

***Example***

In a university, there may be a many-to-one relationship between {CourseId, StudentId} and {Grade}.

Interpretations:

1. A student may have only one grade for a course.
2. We say that there is a FD:
	* CourseId, StudentId -> GRADE, or
	* {CourseId, StudentId} *determines* Grade.
3. Note that under a different set of assumptions, the functional dependency may not be true.
4. For example, if a student is allowed to retake a course, then he may have two grades for the same course (in different semesters), then CourseId, StudentId -> Grade  is *false*.
5. We may instead have {CourseId, StudentId, Semester, Year}-> {Grade}
* Hence, a *functional dependency is a result of the requirements and business logics of the applications*.
* There is no universally true *non-trivial* functional dependency.
* In other words, FD depend on the *semantic* of the problems.

Note that AB->CD is a shorthand notation for {A,B} -> {C,D}

FD such as AB-> A, AB->B, AB->AB, A->A are *trivial*. They are always mathematically true but do not capture any data requirements.

***Example***:

In most applications, we have

SSN -> FName             (i.e.  a person has only one SSN.)

However, in a criminal database, several bad guys may use the same fake SSN, and thus

SSN -> FName  may not be true.

Or, if you are dealing with an international database with many countries, each country may has its own SSN.  Two countries may issue the same SSN.  Hence,

SSN -> FName   is not true.

We may instead have  SSN, CountryId -> Name.

**2.2 Definition of FD**

* A relation *schema* R is said to *satisfy* the *functional dependency* X -> Y if for *any* relation *instance* r that uses R, if there exists two tuples s and t ∈ r such that s[X] = t[X], then s[Y] = t[Y].
	1. (∃s, t ∈ r) (s[X] = t[X]) => s[Y] = t[Y]
	2. i.e.  same value in X implies same value in Y.

***Example***:

This instance r of R violates X->Z.

|  |  |  |
| --- | --- | --- |
| **X** | **Y** | **Z** |
| *'A'* | 1 | *110* |
| *'A'* | 1 | *123* |
| 'A' | 1 | 345 |
| 'B' | 2 | 232 |
| 'C' | 1 | 110 |
| 'C' | 2 | 212 |

This instance r of R *does not violate* X->Y.

|  |  |  |
| --- | --- | --- |
| **X** | **Y** | **Z** |
| 'A' | 1 | 110 |
| 'A' | 1 | 123 |
| 'A' | 1 | 345 |
| 'B' | 2 | 232 |
| 'C' | 1 | 110 |
| 'C' | 1 | 211 |

However, this instance r *does not prove* that X->Y.

In order to have X-> Y, *all* instances r of R must not violate the conditions.

***Examples***:

DeptId -> ManagerId:

There are no two tuples with the same DeptId but different ManagerId.  Meaning: a department can have only one manager.

CourseId, StudentId, Semester -> Grade

There are no two tuples with the same CourseId, StudentId and Semester, but different Grade.  Meaning: any student taking a course in a semester has an unique grade. Note that it may not be true for a different university. Instead, the following may be true:

CourseId, StudentId, Year, Semester -> Grade

***Example***

Consider the following relation:

Supply(SupplierId, SupplierName, ProductId, ProductDesc, Quantity, ArrivalTime)

The relation stores the quantities and arrival times of shipments of products (identified by ProductId) from suppliers (Identified by SupplierId). A supplier may*not* have a unique name. Furthermore, the product description, ProductDesc, may be the same for two products. A supplier may supply the same product many times, each with a different ArrivalTime.

The functional dependencies (FD) of the relation may be:

SupplierId -> SupplierName
ProductId -> ProductDesc
SuplierId, ProductId, ArrivalTime -> Quantity

Decomposition:

Supplier(SupplierId, SupplierName) {SupplierId -> SupplierName}
Product(ProductId, ProductDesc) {ProductId -> ProductDesc}
Supply(SuplierId, ProductId, ArrivalTime, Quantity) {SuplierId, ProductId, ArrivalTime -> Quantity}

***Example (from Spring 2019 HW):***

Consider the following relation GO:

GO(GroupId, GroupName, GroupEMail, GroupChairId, GroupChairLName, GroupChairFName, GroupMemberId, GroupMemberMajor)
The relation stores information about student groups, their chair persons and members. Chair persons and members are students with unique student ids (stored as values in GroupChairId and GroupChairLName respectively). GroupId uniquely identifies a group, and a group has a unique name, and an email address (that may not be unique.) For example, three tuples are shown below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **GroupId** | **GroupName** | **GroupEMail** | **GroupChairId** | **GroupChairLName** | **GroupChairFName** | **GroupMemberId** | **GroupMemberMajor** |
| G1 | Biology | bio@uhcl.edu | 12345 | Lee | Bryan | 23323 | Biol |
| G1 | Biology | bio@uhcl.edu | 12345 | Lee | Bryan | 24990 | Biol |
| G1 | Biology | bio@uhcl.edu | 12345 | Lee | Bryan | 38879 | Phys |

Bryan Lee is the chair student of the group G1 Biology. The three tuples also store information of three members of group G1

(a) List all applicable functional dependencies. (Make reasonable assumptions if necessary.)

GroupId -> GroupName, GroupEMail, GroupChairId
GroupName -> GroupId
GroupChairId -> GroupChairLName, GroupChairFName
MemberId -> GroupMemberMajor

Assumptions:

1. A group has a unique chairperson.
2. A student may be a chairperson or a member for multiple groups.
3. A student has a unique major (e.g., no double majors).

(b) What are the candidate keys?

{GroupId, MemberId} and {GroupName, MemberId}

(c) What is the highest normal form? Why?

1NF. For example, GroupId -> GroupEMail violates 2NF.

(d) If the highest normal form is not BCNF, can you decompose the relation GD losslessly into component relations in BCNF while preserving functional dependencies? If yes, how. If no, why?

1. Group(GroupId, GroupName, GroupEMail, GroupChairId) {GroupChairId references Student(StudentId)}
2. Membership(GroupId, MemberId) {MemberId references Student(StudentId)}
3. Member(StudentId, StudentLName, StudentFName, Major, …)

Note that changes of attribute names in the member tables. For example, StudentLName is more appropriate than ChairLName since a student may not be a chair.