

## CS352 Lecture -The Relational Model; Relational Algebra

last revised December 8, 2020

### *Objectives:*

1. To briefly introduce the entity-relationship model
2. To introduce fundamental concepts of the relational model, including terminology, keys, nulls, views
3. To introduce the relational algebra

### *Materials:*

1. Projectable of an ER diagram example drawn two ways
2. Projectable of figure 2.9 from book
3. On Canvas example of library example database used in class (scheme and instance) - also projectable
4. Projectable of natural language query processing
5. Projectable of query language classification
6. "Cut and Paste" RA symbols on Canvas
7. Projectable of RA queries against example database
8. Note-taking handout for the above on Canvas

### **I. Introduction**

A. We have seen that a database management system typically describes a database at three levels of description.

1. The physical level - how the data is stored in files
2. The conceptual level - the "big picture"
3. The view level - individual views of the database for each application

B. In order to be able to describe a database, we need some system of notation and representation - a data model. This is true at all levels; but, we are

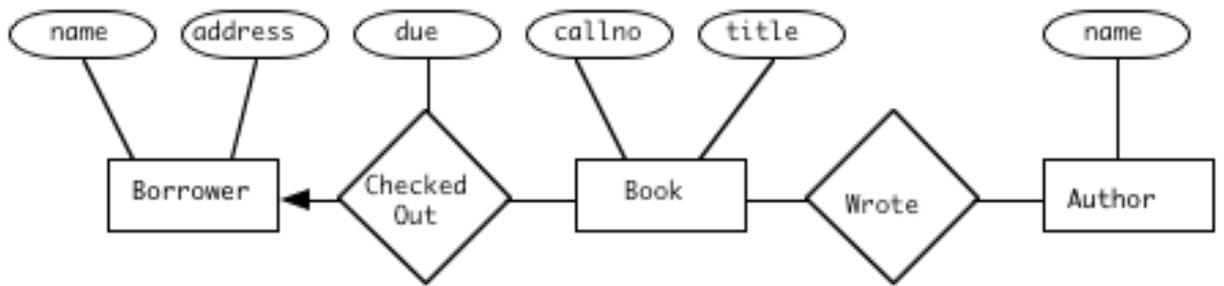
particularly concerned with description at the conceptual and view levels.  
We must describe two things:

1. Data objects
2. Relationships between data objects

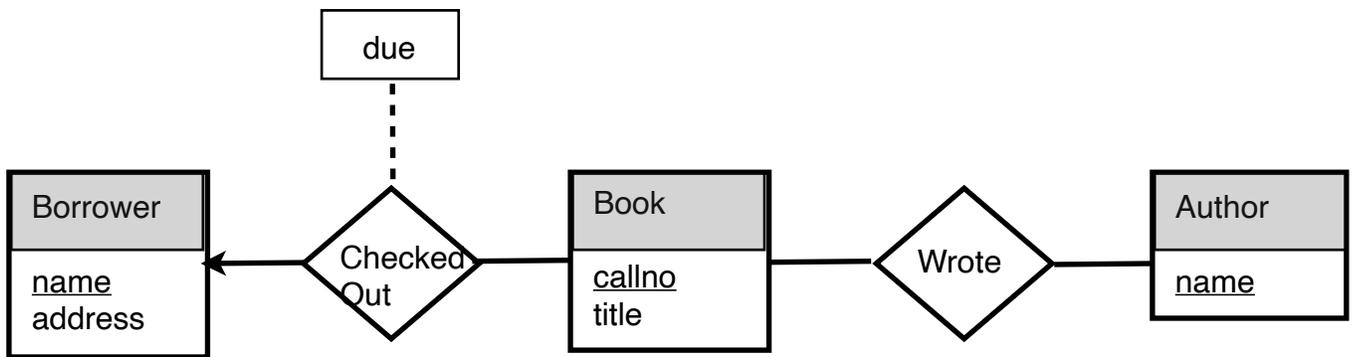
C. We noted last time that, historically, there have been five major types of data model - all of which are represented by DBMS's in use today, though two are mostly in legacy systems:

1. The hierarchical model (largely obsolete, but still used in some legacy systems -we will not look at)
2. The network model (largely obsolete, but still used in some legacy systems -we will not look at)
3. The relational model (the dominant model, and the focus of much of the course). We will begin talking about it today.
4. Various object-oriented and object-relational models (we will look at briefly in this course)
5. Various semi-structured models used in situations where performance is critical (we will look at briefly in this course)

D. There is another model, called the entity-relationship model, which is not, per se, a basis for commercial products; but it is a very useful tool for DESIGNING databases. The ER model utilizes a diagrammatic notation, known as an ER diagram. Here is an example, using one of several different notations that are used for ER diagrams:



Can also be drawn this way:



PROJECT

1. An ER diagram incorporates the following elements:
  - a) Entity - an entity is an object that we wish to represent information about.  
  
Example in the above: Borrower, Book, Author
  - b) Relationship - a relationship is some connection between two or more entities:  
  
Examples: In the above, the “Checked Out” relationship between a borrower and a book; the “Wrote” relationship between a book and its authors.
  - c) Attribute - Individual facts that we store concerning an entity or relationship.
    - (1) Example: In the above, we record for a Book entity its call number and title. (In practice, we'd record a lot more as well)
    - (2) Example: Relationships do not always have attributes, but sometimes they do. CheckedOut has a DateDue attribute, while Wrote has no attributes.
2. Note two ways of drawing. (There are actually more we will talk about later!)
3. While all the data models represent entities in roughly the same way - as a sequence of records in a file - each representing one entity - one of the key distinctions between the various models is how they represent relationships.

## II. The Relational Model

- A. The Relational Data Model is the predominant data models used for commercial DBMS's. Although it was first proposed in the 1970's, it has been the subject of intensive research that continues to today, and has become the dominant model in actual use. It will, therefore, be the major focus of this course.
  
- B. One of the major advantages of the relational model is that it has a solid mathematical foundation. The relational model is based on the mathematical theory of relations. As a result, a large body of mathematical tools is available to support work with this model.
  
- C. The Relational Model represents both entities and relationships the same way, via tables.
  - 1. This means that the relational model draws no distinction between entities and relationships; both are simply tables as far as the relational model is concerned, and are processed in exactly the same way.
  
  - 2. However, a distinction between entities and relationships can be useful in the process of designing a database scheme. Thus, one can use the entity-relationship model as a design tool, with the relational model serving as a means of implementation. The conversion from design to implementation is straight-forward since both entities and relationships can easily be represented by tables. (We will look at this more formally when we talk about database design.)
  
  - 3. One important requirement in the relational model is that the relational model requires that attributes (table columns) be *atomic* and *single-valued*.
    - a) The former requirement can be easily met by allocating one table column to each component of a composite attribute.

Example: A person's address is not atomic. In the US, we normally consider it to consist of a street address, city, state, and ZIP code. Therefore, in setting up a relation scheme for a person, we will generally have separate attributes for each. (Sometimes even two attributes for street address, separating it into the street address proper and an apartment number or the like.)

- b) The latter requirement may necessitate creating a separate table to store the different values of a multivalued attribute - we will discuss this further when we discuss multivalued dependencies in connection with database design.

Example: Many people have more than one phone number. In a relational database, we would need to use a table to store phone numbers, rather than having a phone number attribute that stores an array of values.

- (1)Historically, the need to do this - which led to searching multiple tables - had performance implications which led to the earlier models persisting.

- (2)A lot of work has gone into indexing strategies to produce major improvements in the performance of relational systems, to the point where they now dominate the database world.

- (3)While modern relational database systems generally give very good performance, for extremely large datasets efficiency considerations have led to the rise of some non-relational models we will look at near the end of the course.

D. We looked at relational databases briefly in CPS221. Some of what we look at here will be review of concepts we saw then (just in case someone might have forgotten something - which would never happen, of course :-)) but much of what we will discuss here will be new.

### III. Basic Terminology

A. Because the relational model is grounded in the mathematical theory of relations, writers often use mathematical terms when discussing the relational model. However, since mathematical terminology can be intimidating to some, there is an alternate, non-mathematical set of terms that can also be used.

#### B. Terminology drawn from mathematical relations

1. A relational database is a collection of relations.

2. Formally, a relation is a set of ordered tuples of some arity.

a. Note that the term tuple is not as unfamiliar as it sounds. For example, if a woman gives birth to two children, we say she has had twins; but if she has five, we say she has had quintuplets; if six, sextuplets etc.

b. The arity of a tuple is the number of components in it - e.g.  
(aardvark, anthony) is a tuple of arity 2

c. Within any given relation, all the tuples have the same arity.

d. When we say that the tuples are ordered, we mean that the order of listing the components is important - e.g.

(aardvark, anthony) is not the same tuple as (anthony, aardvark)

e. Since a relation is a set, the order of the tuples themselves is immaterial - e.g. the following are the same relation

(aardvark, anthony)

(elephant, emily)

and

(elephant, emily)

(aardvark, anthony)

3. The components of a tuple are generally called attributes. Mathematically, each attribute of a tuple is given a number - e.g. in (aardvark, anthony), aardvark is attribute 1 and anthony is attribute 2. (Note that, by convention, 1-origin indexing is used.) However, in relational databases, we normally give the attributes names (perhaps last\_name and first\_name, in this case.)
4. Further, each attribute of a tuple is drawn from a specific domain, or set of possible values.

Example: the last\_name attribute above is drawn from the set of all possible last names.

- a. In many cases, actual relational DBMS software only allows the domain to be specified in terms of a basic data type - e.g. “integer” or “string of 30 characters”.
- b. Some systems do, however, allow the enforcement of domain constraints beyond basic data type.

In such a system, proper specification of domains could prevent semantic errors, such as mistakenly storing a ZIP+4 ZIP code (9 digits) into a social security number field!

5. The specification of the structure of tuples in a relation is called a relation scheme.

Example: suppose we had a relation in which the first attribute of each tuple was drawn from the set of ASCII characters and the second from the set of integers. (Perhaps this relation comprises an ASCII code table.)

We could describe the relation scheme for this relation as follows:

ASCII\_code\_table\_scheme = (character: char, code: integer)

(Note the analogy to a type in a programming language.)

A specific relation on some scheme is called a relation instance or an instance of that scheme

(Note the analogy to the value of a variable in a programming language)

6. A relation instance on some scheme is, in fact, a subset of the cartesian product of the domains of its attributes. (It can be any subset, including an improper subset or the empty subset)

Example: Suppose we had a relation on the following scheme:

(One\_digit\_integer: 0..9, flag: boolean)

The domain of the first attribute would be { 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 }  
and that of the second would be { false, true }

The Cartesian product of these domains is:

0	false
1	false
2	false
3	false
4	false
...	
8	true
9	true

Any relation on this scheme would be a subset of this cartesian product. E.g. the table which indicates whether a given integer is odd would be the subset:

0	false
1	true
2	false
3	true
4	false
...	
8	false
9	true

### C. Alternate (non-mathematical) terminology

1. We may also speak of a relational database as a collection of tables. (A table, then is another name for a relation.)

2. Each table, in turn, consists of some number of rows and columns. (A row, then, is another name for a tuple; and a column is another name for an attribute.)

The relation-scheme is often represented as headings on the columns - e.g.

last_name	first_name
aardvark	anthony
cat	charlene

#### D.Keys

1. Because a relation is a set, the tuples comprising it must be distinct (no two can be identical)

- a. A superkey for a relation is a set of attributes which serve to distinguish any tuple in the relation from all others - i.e. each tuple in the relation will have a different set of values for superkey attributes than any other tuple in the relation.

A superkey may be a single attribute, or it may be composite (two or more attributes). (if the superkey is composite, some attributes may have the same value in different tuples as long as at least one differs so that the exact same combination does not appear more than once.).

Because a relation is a set, all of the attributes, together, will necessarily be a superkey.

- b. A candidate key is a superkey which has no proper subset that is also a superkey.
  - c. A primary key is one particular candidate key chosen by the designer to serve as the basis for identifying tuples.
2. An essential characteristic of a database design is that every relation has a primary key (though perhaps this primary key will, in fact, be all of the attributes together.)

3. Another important notion in relational database design is the notion of a *foreign key*. When a column or group of columns in one table is the primary key of some other table, we call this a foreign key (the key of some other table is present in this table.) Foreign keys are, of course, the way we represent relationships in relational tables.
4. It is also possible to require a reference to attributes that are not the primary key - though this is less common.

#### E. Nulls

1. One interesting question that arises in database design is how are we to handle a situation where we don't have values available for all the attributes of an entity.  
Example: Suppose we need to add a person to a scheme that includes the phone number, but for some reason we don't know the person's phone number.
2. For cases like this, relational DBMS's provide a special value called NULL that can be stored in a field. When we print out the field, NULL will normally print as spaces; but storing NULL in a field is NOT the same as storing a string of spaces (i.e. the DBMS software distinguishes between NULL and " ").

#### F. The scheme for a relational database can be represented in diagrammatic form by a schema diagram

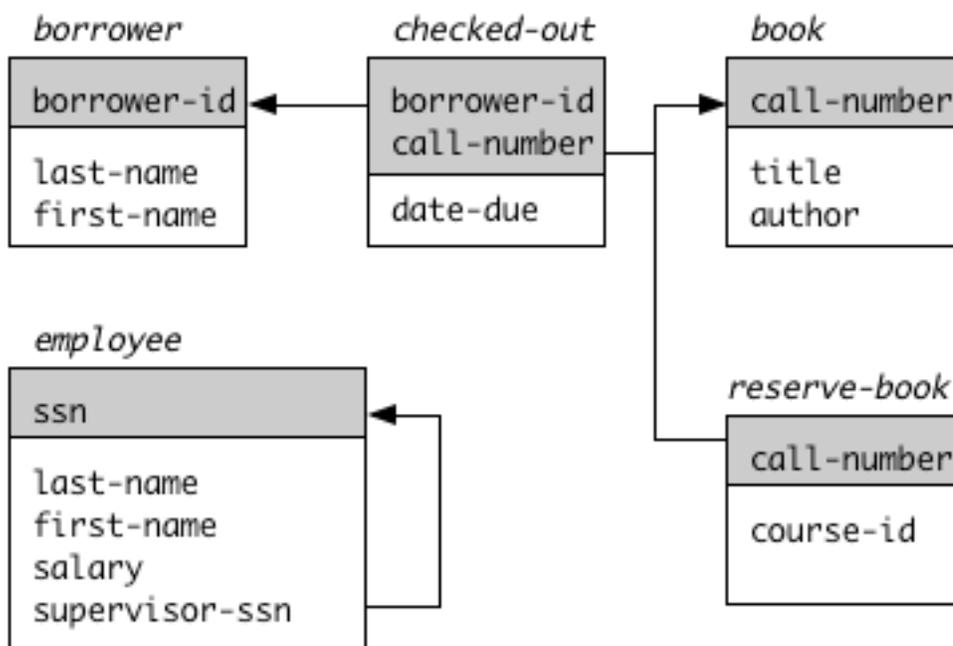
PROJECT Figure 2.9 from book, Note the explicit representation for foreign keys, and the use of a double-headed arrow to represent a reference to something other than the primary key. (Sometimes in a case like this the connecting line is drawn without an arrow at all instead of a double-headed arrow)

## IV. Query Languages

A. For the rest of this series of lectures, we will make use of the following very simple example database for a small library. (This is similar to the one we used in CPS221, but has two more tables and allows illustrating several more operations.)

The structure of this database has been contrived to allow us to illustrate most kinds of query operations we might want to perform. (Relations have been kept unrealistically small, both in terms of the number of rows and in terms of the number of columns).

Schema diagram:



Simplifying assumptions for this example:

- 1) author of a book is single-valued
- 2) there is only one copy of a book with a given call number
- 3) a given book can only be on reserve for a single course
- 4) course-id is presumably a foreign key in a table not shown

Example instance: (Note: primary key attributes are underlined)

borrower(borrower\_id, last\_name, first\_name)

12345	aardvark	anthony
20147	cat	charlene
89754	dog	donna
60984	fox	frederick
54872	zebra	zelda

book(call\_number, title, author)

QA76.093	Wenham Zoo Guide	elephant
RZ12.905	Fire Hydrants I Have Known	dog
LM925.04	21 Ways to Cook a Cat	dog
AB123.40	Karate	koala

checked\_out(borrower\_id, call\_number, date\_due)

89754	RZ12.905	11-10-02
89754	LM925.04	11-10-02
20147	AB123.40	11-15-02

reserve\_book(call\_number, course\_id)

QA76.093	BY123
AB123.40	PE075

employee(ssn, last\_name, first\_name, salary, supervisor\_ssn)

123-45-6789	aardvark	anthony	\$40000	null
567-89-1234	buffalo	boris	\$30000	123-45-6789
890-12-3456	elephant	emily	\$50000	123-45-6789
111-11-1111	fox	frederick	\$45000	567-89-1234

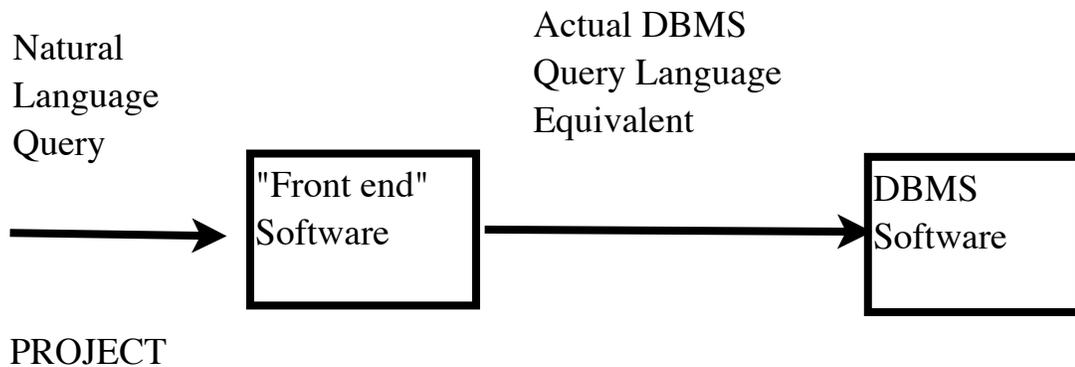
PROJECT - SHOWING CONTENT AND SCHEME AND REFERENCE  
COPY ON CANVAS

- B. One of the main reasons for having a database is to be able to answer queries, such as the following that might be posed against our example. (Some of these are ones used in CPS221, but others involve more advanced features than those discussed there)

1. Who is the borrower whose borrower id is 12345?
2. List the names of all borrowers
3. What is the title of the book whose call number is QA76.093?
4. List the titles of all books that are currently checked out.
5. List the names of all borrowers having one or more books overdue.
6. List the titles of all books, together with the borrower id of the person (if any) who has the book out.
7. List the names of employees together with their supervisor's name.
8. List the names of all employees who earn more than their supervisor.
9. List the names of all people connected with the library - whether borrowers, employees, or both.
10. List the names of all borrowers who are not employees.
11. List all books needed as course reserves that are currently checked out to someone.
12. List the call numbers of all overdue books, together with the number of days they are overdue.

(When we get to SQL we will look at some other examples)

Queries such as these are termed natural language queries. Of course, most commercial DBMS's cannot accept queries like this, though natural language query capability is an area of work in AI. Instead, most DBMS's use a more formal query language. Even those with natural language capabilities typically consist of a regular query language plus a natural language front end - e.g.



C. Actual queries against a database are expressed in a query language, which is of one of two kinds:

1. A formal query languages - using mathematical notation - is used for theoretical work. Typically, a query written in a formal query language is not readily understandable to a typical user.
2. A commercial query language is a query language used by actual DBMS's.
3. Though commercial languages are what is used in actual DBMS applications, formal languages are worthy of some study:
  - a. All commercial languages build on underlying formal language concepts.
  - b. Formal languages provide a vehicle for research (especially the proving of various theorems) that generates results useful in the development of commercial languages.

D. Both formal and commercial query languages are of two general types: procedural and non-procedural. (The 7th ed of the book uses different terminology - functional and declarative - but the idea is the same and I will stick with the older terms for clarity)

1. A procedural language requires the user to spell out the series of steps needed to satisfy his query.
2. A non-procedural language allows the user to simply specify what he/she wants, with the DBMS figuring out the steps needed to get it.

Example (from a different realm): Suppose one wanted to find the roots of the following equation:

$$x^2 + 3x + 2 = 0$$

a) In a non-procedural language, one would simply specify the equation and indicate in some way that an appropriate value of  $x$  is wanted.  
 (“Find the value of  $x$  such that  $x^2 + 3x + 2 = 0$ )

b) In a procedural language, one would specify steps like:

$$\text{Compute discriminant} = 3^2 - 4 * 1 * 2$$

$$\text{Compute root1} = (-3 + \text{sqrt}(\text{discriminant})) / 2$$

$$\text{Compute root2} = (-3 - \text{sqrt}(\text{discriminant})) / 2$$

Note that our sample English queries were all of a non-procedural sort. A major part of the challenge in developing a natural language front-end for a DBMS is the process of converting a non-procedural query to a correct series of procedural steps if the actual DBMS query language is procedural.

E. In our study of relational databases, we will look at four query languages - one of each possible combination:

	<b>Formal</b>	<b>Commercial</b>
<b>Procedural</b>	Relational Algebra	SQL
<b>Non-procedural</b>	Relational Calculus	QBE

## PROJECT

1. We will look at relational algebra today, and SQL as the next major topic in the course.
2. We will look much more briefly at relational calculus and QBE later in the course.

## IV. Query Operations in the Relational Algebra

Queries in the relational algebra are formulated in terms of six basic operations on relations. Three operate on a single relation to give a new relation; and three operate on pairs of relations to give a new relation.

Relational algebra uses symbols for expressing these basic operations. When this course was offered previously, students often wrote RA expressions by hand when doing homework - but since homework is now being submitted online, it will need to be word-processed. I have placed a handout on Canvas with cut and paste versions of these symbols.

Show "Cut and Paste" handout on Canvas

- A. The operation of SELECTION (denoted by the Greek letter sigma -  $\sigma$ ) selects rows of a table that meet certain criteria. The result is a new table with (generally) fewer rows than the original table, but with the same columns.

Example: What is the relational algebra expression corresponding to our first English query: "Who is the borrower whose id is 12345?"

ASK

$\sigma$  borrower  
borrower\_id = 12345

What is the result of this query?

ASK

12345      aardvark      anthony

Note, then, that selection reduces the table by squeezing out rows that do not satisfy the selection condition.

PROJECT RA VERSION AND RESULTS OF THIS AN SUBSEQUENT OPERATIONS

B. The operation of PROJECTION (denoted by the uppercase Greek letter pi -  $\Pi$ ) chooses only specific columns from all the rows of a table

Example: What is the relational algebra expression corresponding to our second query: "List the names of all borrowers"?

ASK

$\Pi$  borrower  
last\_name  
first\_name

What is the result of this query?

ASK

aardvark	anthony
cat	charlene
dog	donna
fox	frederick
zebra	zelda

Example: our third query ("What is the title of the book whose call number is QA76.093?") could be satisfied by a selection from book to get the row for the book in question, followed by a projection on the book-title attribute. (I.e. *composing* two relational algebra operations) - written as

$\Pi$   $\sigma$  book  
title call\_number = QA76.093

What is the result of this query?

ASK

Wenham Zoo Guide

Note, then, that projection reduces the table by squeezing out unwanted columns.

1. One possible problem with a projection operation is that it could produce a result with duplicate rows, by projecting out the attribute that

distinguishes them.

Example: Project book on the author attribute. We now have two books with the same author.

This is a problem. Why? ASK

A relation is a set, and a set cannot have the same element appear twice. Thus definition of projection specifies that the result of a projection has any such duplications eliminated. That is, the projection operation can also result in squeezing rows out of the result.

2. However, the SQL language (which is based on relational algebra) does not do this automatically, because of the time involved - though there is a SQL query language feature to specify that duplicates are to be eliminated.

D. The operation CARTESIAN PRODUCT or a special form of it called JOIN is used when a query involves information contained in more than one table.

Example: Our next two queries would require a join.

“List the titles of all books that are currently checked out” requires information from the checked\_out table (to find out the call number(s) of book(s) currently checked out) and from the book table (to get the titles)

“List the names of all borrowers having one or more books overdue” requires information from the checked\_out table (to find out the ids of borrowers having books overdue) and from the borrower table (to get their names)

1. There are several variants of this operation.
2. The cartesian product - denoted by X - generates all possible pairings of the rows from the two tables.

Example: we could use a cartesian product as part of the performing the for the query "list the titles of all books that are currently checked out":

checked\_out X book

a. This produces large result

borrower id	call number	date-due	call number	title	author
89754	RZ12.905	11-10-02	QA76.093	Wenham Zoo Guide	elephant
89754	RZ12.905	11-10-02	RZ12.905	Fire Hydrants ...	dog
89754	RZ12.905	11-10-02	LM925.04	21 Ways to Cook ..	dog
89754	RZ12.905	11-10-02	AB123.40	Karate koala	
89754	LM925.04	11-10-02	QA76.093	Wenham Zoo Guide	elephant
89754	LM925.04	11-10-02	RZ12.905	Fire Hydrants ...	dog
89754	LM925.04	11-10-02	LM925.04	21 Ways to Cook ..	dog
89754	LM925.04	11-10-02	AB123.40	Karate koala	
20147	AB123.40	11-15-02	QA76.093	Wenham Zoo Guide	elephant
20147	AB123.40	11-15-02	RZ12.905	Fire Hydrants ...	dog
20147	AB123.40	11-15-02	LM925.04	21 Ways to Cook ..	dog
20147	AB123.40	11-15-02	AB123.40	Karate koala	

### *PROJECT*

- i. Because the checked\_out table contains 3 rows, and the book table 4 rows, the result of the join contains  $3 \times 4 = 12$  rows.
- ii. Because the checked\_out table has three columns and the book table has three columns, the result of the join has  $3 + 3 = 6$  columns.

This will always be true for a cartesian product: the result has as many rows as the product of the numbers of rows of the two tables, and as many columns as the sum of the numbers of columns of the two tables. (For the former reason, simple join is sometimes called product or times.)

Question: For the second query - requiring a cartesian product of checked\_out and borrower) how many rows would the result table have? How many columns?

*ASK*

3 x 5 = 15 rows  
3 + 3 = 6 columns

- b. However this result is not really useful, Instead, when we combine two tables generally only certain rows are meaningful.

Example: in the above, the call number appears twice - once from each table - but only the rows in which the two values of call number are the same tell us anything worth knowing - i.e. the fact that book RZ12.905 is checked out means that the title of RZ12.905 is of interest. On the other hand, the row that pairs RZ12.905 with QA76.093 means nothing. Thus, the cartesian product must almost always be immediately followed by a select.

1. One way to do this is to do a select after the join - e.g.

```
||      σ  checked_out X      book
      title checked_out.call_number =
          book.call_number checked_out
```

2. Since relational algebra is a procedural language, this implies that we do the cartesian product first and then go through every row of the result keeping only those that satisfy the condition.
3. A much more efficient way to do this would be to examine each row of the cartesian product when it is generated and only keep rows that satisfy the condition, discarding the rest.

This yields an operation called JOIN , which has the following symbol (where the theta stands for the selection condition)

$\bowtie_{\theta}$

Often the join condition specified by the theta will be requiring equality of values for some columns between the two tables - as is the case here. Such a join is sometimes called an equijoin, And

sometimes a join in which the join condition is something other than a match of values is called a theta join.

- c. In formulating the condition, of course, one must use attribute names from both relations in the join. Since there is no rule that each relation must have unique attribute names (and, in fact, there are good reasons for having common names between relations), one must often qualify attribute names by the name of the relation they come from.

If we now project this on title, we get the answer to our query about title of books checked out:

```
Π      checked_out |X|θ      book
  title      checked_out.call_number =
              book.call_number
```

(Note that we do not need to qualify the title attribute, since it appears in only one relation.)

What is the result of this query?

*ASK*

Fire Hydrants I Have Known  
21 Ways to Cook a Cat  
Karate

- d. In similar fashion, our fifth query (“List the names of all borrowers having one or more books overdue.”) can be answered by using a join with project as above, but with an additional join conditions (or, less efficiently, a select after the fact)

```
Π      checked_out |X|θ borrower
  last_name,      checked_out.borrower_id =
  first_name      borrower.borrower_id ^
                  date_due < today
```

4. When our join is an equijoin based on columns having the same name (as was true in this case), we typically don't want the common value to appear twice. Relational algebra defines a special variant of the join for it called the NATURAL JOIN (denoted by the join symbol with bars on each side but without an explicit join condition.  $\bowtie$ ).

a. Example: the query "list the titles of all books that are currently checked out" could be done using a natural join as follows:

$\Pi$  (checked\_out  $\bowtie$  book)  
title

This eliminates the need for explicitly specifying the join condition, though projection is still needed to eliminate columns other than the title.

b. However, the query "List the names of all borrowers having one or more books overdue", if done with a natural join, also requires a selection following the natural join to select only the rows for overdue books.

c. The natural join is defined to require equality of values in the columns in the two joined tables having the same name and the definition also specifies that only one copy of the common column(s) is included in the result.

d. It could be realized by using a join combined with projection (or cartesian join plus both selection and projection.) However, if the natural join is available it is simpler to use because one need not explicitly specify the join conditions.

e. Note that natural join is defined in terms of attribute names. This presupposes that when columns in two different tables have the same name, they refer to the same thing. If the different names mean different things, the result could be undesirable - e.g. suppose our employee table included a "title" attribute which was the employee's job title - then

book  $\bowtie$  employee

would give a table listing employees whose job title is the same as the title of some book! (The possibility that something like this might happen accidentally when a column is added to a table is a reason why some experts discourage use of natural join and some systems (e.g. the one we will use, IBM Db2) do not support it.)

5. Another variant of join is the OUTER JOIN, which can either be a variant of any equijoin. (We will use the natural join for example here - but remember it can also be used with any type of equijoin.)

a. An ordinary natural join operation will omit from the result any row from either table that does not join with some row from the other table.

Example:

List the titles of all books, together with the borrower id of the person (if any) who has the book out. As a first attempt, we could use the following approach:

```
Π          book |X| checked_out
title
borrower_id
```

What is the result of this query?

ASK

Fire Hydrants I Have Known	89754
21 Ways to Cook a Cat	89754
Karate	20147

b. Note that there is no row in this table for Wenham Zoo Guide. Why?

ASK

Wenham Zoo Guide is not checked out, so there is no row in the checked\_out table for the row in book to join with.

c. If we want to be sure all books are included, even if not checked out, we can use an outer join:

```

||      book  ⋈| checked_out
   title
   borrower_id

```

which will pair any tuple in book that does not join with any row in checked\_out with a “manufactured” row of all nulls

d. The outer join comes in three versions

- i. The left outer join - denoted  $\bowtie_{\theta}^{\leftarrow}$  or  $\bowtie_{\leftarrow}$  - which we used above - joins any tuple in the left relation that does not join with anything in the right relation with a dummy tuple of all nulls. Any tuple in the right relation that does not match a tuple in the left is absent from the result.
- ii. The right outer join - denoted  $\bowtie_{\theta}^{\rightarrow}$  or  $\bowtie_{\rightarrow}$  - joins any tuple in the right relation that does not join with anything in the left relation with a dummy tuple of all nulls. Any tuple in the left relation that does not match a tuple in the right is absent from the result.
- iii. The full outer join - denoted  $\bowtie_{\theta}^{\leftrightarrow}$  or  $\bowtie_{\leftrightarrow}$  - combines the above - hence no tuple from either table is omitted.

E. The operation RENAME takes a given table and gives it a new name, possibly also renaming its attributes as well. Rename is denoted by the Greek letter rho -  $\rho$ .

1. This operation is often used in doing a join between a table and itself.
2. Example: Our query "List the names of employees together with their supervisor's name" requires joining the employee table with itself on the condition that the employee's supervisor\_ssn matches the supervisor's ssn."

This requires giving the table a second name so that we can refer to it under two names, as follows:

```

Π      employee | X | θ      (ρ employee)
employee.last_name      employee.      supervisor
employee.first_name      supervisor_ssn =
supervisor.last_name      supervisor.ssn
supervisor.first_name

```

a. The rename operation here causes the second copy of employee to be called supervisor for the purposes of qualifying names in the select and project.

b. However, this fails to produce a row for anthony aardvark.

ASK WHY? (Hint - look closely at the employee table)

anthony aardvark has no supervisor!

c. How would we fix this?

ASK

Use a left outer join

```

Π      employee ⋈ X | θ      (ρ employee)
employee.last_name      employee.      supervisor
employee.first_name      supervisor_ssn =
supervisor.last_name      supervisor.ssn
supervisor.first_name

```

3. Another example: “List the names of all employees who earn more than their supervisor”. We can do this by

```

Π      σ      (ρ employee | X | θ      ρ employee)
e.last_name  e.salary > e      e.supervisor_ssn = s
e.first_name s.salary      s.ssn

```

or folding the salary comparison into the join condition (and also renaming both tables)

```
Π (ρ employee |X| ρ employee)
  e.last_name e      e.supervisor_ssn = s
  e.first_name      s.ssn ^
                    e.salary > s.salary
```

What is the result of this query?

*ASK*

```
elephant  emily
fox       frederick
```

4. Actually, as the example above suggests, when we need to qualify column names, it is often convenient to rename the tables in a join to simplify the names involved in the selection clause and other operations
  5. The rename operation can also be used to rename the columns of a table - we will see an example of this shortly
- F. The operation UNION takes two relations on the same scheme and combines them into one, eliminating duplicate tuples.
1. Example: Our query “List the names of all people connected with the library - whether borrowers, employees, or both.” requires combining information from the borrower table and the employee table. A join is not what we want - in no sense would pairing a borrower and an employee make sense - we want to know the names of people who are either or both. We need a union in this case.
  2. One requirement for the union is that the schemes of the two tables be the same - that is, the same number and domains of the columns. Borrowers don't have a salary or a supervisor, and borrower\_ids and ssns - though serving similar roles, are actually drawn from separate domains. We therefore use projection to project out all but the thing the two tables have in common: last and first names

$$(\Pi_{\text{borrower}}) \cup (\Pi_{\text{employee}})$$

last_name,	last_name,
first_name	first_name

3. The symbol is the familiar set algebra symbol for union:  $\cup$
4. Because relations are sets, duplicate rows that would result from a union are discarded - e.g., in the above, suppose an someone were both a borrower and an employee - then their name would appear only once, even though it is in both tables. (Though for efficiency reason, many relational DBMS's allow you to avoid this operation and accept duplicates, since discarding duplicates takes additional processing.)

Example: in the above, there is only one row for aardvark and fox

5. Sometimes, we may have a case where tables being combined by union have similar attributes with different names. In this case, we may need to rename individual attributes.

Example: Suppose we wanted our list to include an "id" attribute in both cases - choosing either the borrower\_id or the ssn as appropriate. This could be done by:

$$(\rho_{t1}(\Pi_{\text{borrower}}) \cup (\rho_{t2}(\Pi_{\text{employee}}))$$

id,	borrower_id	id,	ssn
last_name,	last_name	last_name	last_name
first_name)	first_name	first_name)	first_name

(Note that we use  $\rho$  to rename the "id" attributes to both be called id in the result.)

(Note that, in this case, a person who was both a borrower and an employee (aardvark and fox) definitely appears twice in the result - one with each id!)

G. The operation of DIFFERENCE takes two relations on the same scheme and returns a relation having only those tuples which appear in the first but the not second.

Example: Consider the query “List the names of all borrowers who are not employees.” How would this be accomplished?

*ASK*

$$\begin{array}{cc} (\Pi \text{ borrower}) - (\Pi \text{ employee}) \\ \text{last\_name,} & \text{last\_name,} \\ \text{first\_name} & \text{first\_name} \end{array}$$

1. The symbol is the familiar set algebra symbol for difference: -
2. As with union, the difference operation only makes sense if the tables involved have the same arity, and corresponding attributes come from the same domain.

H. One other operation is not strictly speaking, not necessary, since it can be expressed in terms of the six primitives (selection, projection, join, rename, union, and difference). But it is useful enough to include it as an operation in its own right anyway.

The operation of INTERSECTION is similar to union and difference, in that it combines two relations having the scheme. But where union produces all tuples that occur in either table, and difference gives those tuples in the first relation but not the second, intersection gives only those tuples that occur in BOTH tables.

Example: How would we answer the question “List all books needed as course reserves that are currently checked out to someone”?

*ASK*

$$\begin{array}{cc} (\Pi \text{ reserve\_book}) \cap (\Pi \text{ checked\_out}) \\ \text{call\_number} & \text{call\_number} \end{array}$$

(Joining could now be used to get additional information such as the Title of the book and/or the borrower who has it.)

1. The symbol for intersection is the familiar set algebra symbol  $\cap$
2. If intersection is not available in a particular commercial query language, it can be implemented from the other primitives as follows:

to form the intersection of relations R1 and R2, compute

$$R1 - (R1 - R2)$$

- a.  $R1 - R2$  computes all tuples in R1 but not in R2
  - b. therefore  $R1 - (R1 - R2)$  leaves only those tuples from R1 which also occur in R2 - i.e. all tuples occurring in BOTH R1 and R2.
3. As with union and difference, intersection only makes sense if both tables have the same arity and corresponding attributes come from the same domains. (Note how we used projection in the example to facilitate this.)

- I. An additional operations that extends the basic relational algebra, is generalized projection - allow a projection expression to contain computations based on column values, not just column values themselves.

Example:

List the call numbers of all overdue books, together with the number of days they are overdue.

ASK

```

Π      σ  checked_out
call_number      date_due < today
today - date_due

```

## J. How do the various operations handle the special value NULL?

A key property of NULL is that it is never treated as any value during a query - the principle being that, since NULL means we don't know a value, it should never participate in any query. Thus, if we compare an attribute whose value is NULL to some value, the result of the comparison is always false, regardless of which comparison ( $=$ ,  $<.$ ,  $>.$ ,  $<>$ ,  $>=$  or  $<=$ ) we use. In fact, it is even the case that  $NULL = NULL$  is false, and so is  $NULL <> NULL$ . Further, a NULL value never participates in any aggregate operations such as average.

1. Example: suppose we are doing a natural join between two relations which each contain a phone-number attribute. If a tuple in the first relation contains a NULL phone-number, it will not join with any tuple in the second relation - even one that also has a NULL phone-number. (Since  $NULL = NULL$  is false!)
2. Example: suppose we are calculating the average salary for all employees in a company. If the salary for some employee is stored as NULL, then - for the purposes of taking the average - it will be as if that employee were not present. (Notice that this gives a different result than if we took the salary to be 0.)
3. In designing a database, it will sometimes be necessary to specify that certain fields CANNOT ever contain a NULL value. This will certainly be true of any field that is part of the primary key, and may be true of other fields as well. Most DBMS's allow the designer to specify that a given field cannot be NULL