Relational Algebra – Week2

CIS3530, Introduction to Databases
Ritu Chaturvedi

Some slides are taken from DB courses I have taught previously at UTM and University of Windsor.

Recap

- Worksheet2 – due Saturday Sept 25th
- Lab1 - Sept 27th – due Friday Oct 1st
- Quiz1 – due Sunday Oct 3rd
- A1 – due Thursday Oct 14th

- Rest of the dates are on the course syllabus and in Week0 slides

Structure of classes each week (Week 2 onwards)

- Teach live on zoom Mondays
- Post recorded lecture on Wednesdays
- Teach live on Fridays (worksheets)

Request for volunteer notetaker

Student Accessibility Services (SAS)
Needs a Volunteer Note Taker
For This Course

WHY BE A NOTE TAKER?

- It's **EASY**. No more effort than uploading or scanning your notes
- Typically helps improve your grades
- Help other students on campus who really need it
- Earn a great reference letter

Register and find more information at
www.uoguelph.ca/sas

Make a difference today!

Textbook (6th edition) Readings

- 6.1, 6.2, 6.3, 6.5

Relational Algebra

- Formal query language for a relational model

- Has a set of operations for manipulating the data and constraints in a relational model.

RA vs. Elementary Algebra

- algebra in high school
  - $9y^2 + 12y - 3$
- Operands?
- Operators?

UNIVERSITY OF TORONTO

Relational Algebra

- Operands: relations

- Operators:
  - choose only the rows you want (select (σ))
  - choose only the columns you want (project (π))
  - combine tables (e.g. cartesian product (x), natural join)
  - and a few other things (e.g. division, set operators)

- Allows expressions to be nested, just as in arithmetic
  - Results are also relations!

A schema for our examples

Movies(mID, title, director, year, length)

Artists(aID, aName, nationality)

Roles(mID, aID, character)

Foreign key constraints: shown as arrows in red.

Movies:

mID	title	director	year	length
1	Shining	Kubrick	1980	146
2	Player	Altman	1992	146
3	Chinatown	Polaski	1974	131
4	Repulsion	Polaski	1965	143
5	Star Wars IV	Lucas	1977	126
6	American Graffiti	Lucas	1973	110
7	Full Metal Jacket	Kubrick	1987	156

Roles:

mID	aID	character
1	1	Jack Torrance
3	1	Jake ‘J.J.’ Gittes
1	3	Delbert Grady
5	2	Han Solo
6	2	Bob Falfa
5	4	Princess Leia Organa

Artists:

aID	aName	nationality
1	Nicholson	American
2	Ford	American
3	Stone	British
4	Fisher	American

Select: choose rows

- Notation: $\sigma_c(R)$
  - R is a table.
  - Condition c is a boolean expression.
  - It can use comparison operators and boolean operators
  - The operands are either constants or attributes of R.

$R1 = \sigma_{\text{<boolean expression >}}(R)$;

$\sigma_{\text{length} < 150}(\text{Movies})$

- The result is a relation
  - with the same schema as the operand
  - but with only the tuples that satisfy the condition

Degree of R1 = degree of R
Cardinality of R1 <= Cardinality of R

Exercise

Movies(<u>mID</u>, title, director, year, length)
Artists(<u>aID</u>, aName, nationality)
Roles(<u>mID, aID,</u> character)

- Write queries to find:
  - All British actors
  - All movies from 1970

- What if we only want the names of all British actors?
  We need a way to filter the columns.

Project: choose columns

- Notation: $\pi_L(R)$
  - R is a table.
  - L is a subset (not necessarily a proper subset) of the attributes of R.
- The result is a relation
  - with all the tuples from R
  - but with only the attributes in L, and in that order

Project

Movies(<u>mID</u>, title, director, year, length)
Artists(<u>aID</u>, aName, nationality)
Roles(<u>mID, aID,</u> character)

- Write a RA expression to find the names of all British actors?
- Answer:

- Exercise: Write an RA expression to find all directors of movies from the 1970s
- Answer:

Combine expressions

- Now, suppose you want the names of all characters in movies from the 1970s.

Combine tables

Movies:

mID	title	director	year	length
1	Shining	Kubrick	1980	146
2	Player	Altman	1992	146
3	Chinatown	Polaski	1974	131
4	Repulsion	Polaski	1965	143
5	Star Wars IV	Lucas	1977	126
6	American Graffiti	Lucas	1973	110
7	Full Metal Jacket	Kubrick	1987	156

Artists:

aID	aName	nationality
1	Nicholson	American
2	Ford	American
3	Stone	British
4	Fisher	American

Roles:

mID	aID	character
1	1	Jack Torrance
3	1	Jake ‘J.J.’ Gittes
1	3	Delbert Grady
5	2	Han Solo
6	2	Bob Falfa
5	4	Princess Leia Organa

Project and duplicates

- Projecting onto fewer attributes can remove what it was that made two tuples distinct.
- Wherever a project operation might “introduce” duplicates, only one copy of each is kept.
- Example:

People

name	age
Karim	20
Ruth	18
Minh	20
Sofia	19
Jennifer	19
Sasha	20

π_{age} People

age
20
18
19

Forming larger expressions

- Relational algebra is the same as Math:
  - So you can “compose” larger expressions out of smaller ones.
  - There are precedence rules.
  - We can use brackets to override the normal precedence of operators.

Cartesian Product

Now, suppose you want the names of all characters in movies from the 1970s.

- Notation: R1 x R2
- The result is a relation with
  - every combination of a tuple from R1 concatenated to a tuple from R2
- How many tuples are in R1 x R2 (cardinality)?
- What is the degree of R1 x R2?

R1

col1
a
b

R2

col1	col2
c	1
a	2
d	3

col1	col2	col3

Cartesian Product

- Its schema is every attribute from R followed by every attribute of S, in order
- If an attribute occurs in both relations, it occurs twice in the result (prefixed by relation name)
- Example: Movies x Roles

- Now, suppose you want the names of all characters in movies from the 1970s.

Movies:

mID	title	director	year	length
1	Shining	Kubrick	1980	146
2	Player	Altman	1992	146
3	Chinatown	Polaski	1974	131
4	Repulsion	Polaski	1965	143
5	Star Wars IV	Lucas	1977	126
6	American Graffiti	Lucas	1973	110
7	Full Metal Jacket	Kubrick	1987	156

Artists:

aID	aName	nationality
1	Nicholson	American
2	Ford	American
3	Stone	British
4	Fisher	American

Roles:

mID	aID	character
1	1	Jack Torrance
3	1	Jake ‘J.J.’ Gittes
1	3	Delbert Grady
5	2	Han Solo
6	2	Bob Falfa
5	4	Princess Leia Organa

Cartesian product can be inconvenient

- It can introduce nonsense tuples.
- You can get rid of them with selects.
  - Back to our query: how to we get all characters from movies from the 1970s?
- Answer:

- But this is so highly common, an operation was defined to make it easier: natural join.

Natural Join

- Notation: R $\bowtie$ S
- The result is formed by
  - taking the Cartesian product
  - selecting to ensure equality on attributes that are in both relations (as determined *by name*)
  - projecting to remove duplicate attributes.
- Example:

R1

col0	col1
a	1
b	2

R2

col1	col2
1	x
1	y
3	z

col0	col1	col2
a	1	x
a	1	y

UNIVERSITY OF TORONTO

Example of Cartesian product

profiles:

ID	Name
Oprah	Oprah Winfrey
ginab	Gina Bianchini

follows:

a	b
Oprah	ev
edyson	ginab
ginab	ev

profiles X follows:

ID	Name	a	b
Oprah	Oprah Winfrey	Oprah	ev
Oprah	Oprah Winfrey	edyson	ginab
Oprah	Oprah Winfrey	ginab	ev
ginab	Gina Bianchini	Oprah	ev
ginab	Gina Bianchini	edyson	ginab
ginab	Gina Bianchini	ginab	ev

Natural join with no common attributes

profiles:

ID	Name
Oprah	Oprah Winfrey
ginab	Gina Bianchini

follows:

a	b
Oprah	ev
edyson	ginab
ginab	ev

profiles $\bowtie$ follows

ID	Name	a	b
Oprah	Oprah Winfrey	Oprah	ev
Oprah	Oprah Winfrey	edyson	ginab
Oprah	Oprah Winfrey	ginab	ev
ginab	Gina Bianchini	Oprah	ev
ginab	Gina Bianchini	edyson	ginab
ginab	Gina Bianchini	ginab	ev

Natural join with one common attribute

profiles:

ID	Name
Oprah	Oprah Winfrey
ginab	Gina Bianchini

follows:

ID	b
Oprah	ev
edyson	ginab
ginab	ev

profiles ⋈ follows:

ID	Name	~~ID~~	b
Oprah	Oprah Winfrey	~~Oprah~~	ev
~~Oprah~~	~~Oprah Winfrey~~	~~edyson~~	~~ginab~~
~~Oprah~~	~~Oprah Winfrey~~	~~ginab~~	~~ev~~
~~ginab~~	~~Gina Bianchini~~	~~Oprah~~	~~ev~~
~~ginab~~	~~Gina Bianchini~~	~~edyson~~	~~ginab~~
ginab	Gina Bianchini	~~ginab~~	ev

(The redundant ID column is omitted in the result)

Natural join with a different common attribute

profiles:

ID	Name
Oprah	Oprah Winfrey
ginab	Gina Bianchini

follows:

a	ID
Oprah	ev
edyson	ginab
ginab	ev

profiles ⋈ follows:

ID	Name	a	ID
~~Oprah~~	~~Oprah Winfrey~~	~~Oprah~~	~~ev~~
~~Oprah~~	~~Oprah Winfrey~~	~~edyson~~	~~ginab~~
~~Oprah~~	~~Oprah Winfrey~~	~~ginab~~	~~ev~~
~~ginab~~	~~Gina Bianchini~~	~~Oprah~~	~~ev~~
ginab	Gina Bianchini	edyson	ginab
~~ginab~~	~~Gina Bianchini~~	~~ginab~~	~~ev~~

(The redundant ID column is omitted in the result)

Exactly the same attributes

R1

Artist	Name
1234	William Shatner
8762	Harrison Ford
5555	Patrick Stewart
1868	Angelina Jolie

R2

Artist	Name
1234	Brad Pitt
1868	Angelina Jolie
5555	Patrick Stewart

Properties of Natural Join

- Commutative:
  $R \bowtie S = S \bowtie R$
  (although attribute order may vary)
- Associative:
  $R \bowtie (S \bowtie T) = (R \bowtie S) \bowtie T$
- So when writing n-ary joins, brackets are irrelevant. We can just write:
  $R_1 \bowtie R_2 \bowtie \ldots \bowtie R_n$

- Degree of a natural join = (Sum of degrees of R and S) –(number of common attributes in R and S)

Natural join can “over-match”

- Natural join bases the matching on attribute names.
- What if two attributes have the same name, but we don’t want them to have to match?
- Example: if Artists used “name” for actors’ names and Movies used “name” for movies’ names.
  - Can rename one of them (we’ll see how).
  - Or use a Cartesian product.

Example – Natural join can "over- match"

- famous S-P-SP database by C. J. Date [Date, C. J. "An Introduction to Database System Volume 1." (1990)]

S

SNO	SNAME	STATUS	CITY
S1	SMITH	20	LONDON
S2	JONES	10	PARIS
S3	BLAKE	30	PARIS
S4	CLARK	20	LONDON
S5	ADAMS	30	ATHENS
S6	HARRY	25	WINDSOR

SP

SNO	PNO	QTY
S1	P1	200
S2	P3	400
S2	P5	100
S3	P3	200
S3	P4	500
S4	P6	300
S5	P2	200
S5	P5	500
S5	P6	200
S5	P1	100
S5	P3	200
S5	P4	800

P

PNO	PNAME	COLOR	WEIGHT	CITY
P1	NUT	RED	12	LONDON
P2	BOLT	GREEN	17	PARIS
P3	SCREW	BLUE	17	ROME
P4	SCREW	RED	14	LONDON
P5	CAN	BLUE	12	PARIS
P6	COG	RED	19	LONDON

Natural join can "under-match"

- What if two attributes don't have the same name and we *do* want them to match?
- Example: Suppose we want aName and director to match.
- Solution?

Movies:

mID	title	director	year	length
1	Shining	Kubrick	1980	146
2	Player	Altman	1992	146
3	Chinatown	Polaski	1974	131
4	Repulsion	Polaski	1965	143
5	Star Wars IV	Lucas	1977	126
6	American Graffiti	Lucas	1973	110
7	Full Metal Jacket	Kubrick	1987	156

Artists:

aID	aName	nationality
1	Nicholson	American
2	Ford	American
3	Stone	British
4	Fisher	American

Roles:

mID	aID	character
1	1	Jack Torrance
3	1	Jake ‘J.J.’ Gittes
1	3	Delbert Grady
5	2	Han Solo
6	2	Bob Falfa
5	4	Princess Leia Organa

Theta Join - Equi join

- It's common to use σ to check conditions after a Cartesian product.
- Theta Join makes this easier.
- Notation: $R \bowtie_{condition} S$
- The result is
  - the same as Cartesian product (not natural join!) followed by select.
- In other words, $R \bowtie_{condition} S = \sigma_{condition} (R \times S)$.

UNIVERSITY OF TORONTO

Outer Joins

Dangling tuples

- If a tuple in one relation has no match in the other, natural join leaves that tuple out.

- List all artists and the roles they play.

artists $\bowtie$ *roles*

```
   AID ANAME                    NATIONALITY                     MID CHARACTER
------ ------------------------ ------------------------ ---------- --------------------------
--------------------
     1 Nicholson                American                          1 Jack Torrance
     1 Nicholson                American                          3 Jake ?J.J.? Gittes
     2 Ford                     American                          6 Bob Falfa
     2 Ford                     American                          5 Han Solo
     3 Stone                    British                           1 Delbert Grady
     4 Fisher                   American                          5 Princess Leia Organa
```

Movies (Artists table UPDATED):

mID	title	director	year	length
1	Shining	Kubrick	1980	146
2	Player	Altman	1992	146
3	Chinatown	Polaski	1974	131
4	Repulsion	Polaski	1965	143
5	Star Wars IV	Lucas	1977	126
6	American Graffiti	Lucas	1973	110
7	Full Metal Jacket	Kubrick	1987	156

Artists:

aID	aName	nationality
1	Nicholson	American
2	Ford	American
3	Stone	British
4	Fisher	American
5	Bachchan	Indian

Note that this table now has an additional tuple

Roles:

mID	aID	character
1	1	Jack Torrance
3	1	Jake ‘J.J.’ Gittes
1	3	Delbert Grady
5	2	Han Solo
6	2	Bob Falfa
5	4	Princess Leia Organa

Outer Join

- List all artists and the roles they play, if they play any??

Outer Join

R1

col0	col1
a	1
b	2

R2

col1	col2
1	x
1	y
3	z

R1 ⋈ R2

col0	col1	col2
a	1	x
a	1	y

R1 ⋈ *left* *R2*

col0	col1	col2
a	1	x
a	1	y
b	2	null

R1 ⋈ *right* *R2*

col0	col1	col2
a	1	x
a	1	y
null	3	z

Outer Join

- List all artists and the roles they play, if they play any??

artists

aID	aName	nationality
1	Nicholson	American
2	Ford	American
3	Stone	British
4	Fisher	American
5	Bachchan	Indian

$\bowtie$ *left*

Roles

mID	aID	character
1	1	Jack Torrance
3	1	Jake ‘J.J.’ Gittes
1	3	Delbert Grady
5	2	Han Solo
6	2	Bob Falfa
5	4	Princess Leia Organa

```
 AID ANAME                     NATIONALITY                    MID CHARACTER
---- ------------------------- ------------------------- ---------- -------------------------
-------------------
   1 Nicholson                 American                          1 Jack Torrance
   1 Nicholson                 American                          3 Jake ?J.J.? Gittes
   2 Ford                      American                          6 Bob Falfa
   2 Ford                      American                          5 Han Solo
   3 Stone                     British                           1 Delbert Grady
   4 Fisher                    American                          5 Princess Leia Organa
   5 A Bachchan                Indian
```

Precedence

- Expressions can be composed recursively.
- Make sure attributes match as you wish.
  - It helps to annotate each subexpression, showing the attributes of its resulting relation.
- Parentheses and precedence rules define the order of evaluation.
- Precedence, from highest to lowest, is:

  σ, π

  $\times, \bowtie$

  $\cap$

  $\cup, -$
- Unless very sure, use brackets!

Assignment operator

- Notation:

  R := Expression

- Alternate notation:

  $R(A_1, \ldots, A_n)$:= Expression

  - Lets you name all the attributes of the new relation
  - Sometimes you don't want the name they would get from Expression.

- R must be a temporary variable, not one of the relations in the schema.
  I.e., you are not updating the content of a relation!

- Example:
  - Temp1 := $Q \times R$
  - Temp2 := $\sigma_{a=99}$ (Temp1) $\times$ S
  - Answer(part, price) := $\pi_{b,c}$ (Temp2)
- Whether / how small to break things down is up to you. It's all for readability.
- Assignment helps us break a problem down

Rename operation

- Notation: $\rho_{R1}(R2)$
- Note that these are equivalent:
  $R1 := R2$
  $\rho_{R1}(R2)$
- ρ is useful if you want to rename *within* an expression.

Example

- Find pairs of names of all artists with the same nationality.

aID	aName	nationality
1	Nicholson	American
2	Ford	American
3	Stone	British
4	Fisher	American

ROW	aID	aName	nationality	aID	aName	nationality
1	1	Nicholson	American	1	Nicholson	American
2	1	Nicholson	American	2	Ford	American
3	1	Nicholson	American	3	Stone	British
4	1	Nicholson	American	4	Fisher	American
5	2	Ford	American	1	Nicholson	American
6	2	Ford	American	2	Ford	American
7	2	Ford	American	3	Stone	British
8	2	Ford	American	4	Fisher	American
9	3	Stone	British	1	Nicholson	American
10	3	Stone	British	2	Ford	American
11	3	Stone	British	3	Stone	British
12	3	Stone	British	4	Fisher	American
13	4	Fisher	American	1	Nicholson	American
14	4	Fisher	American	2	Ford	American
15	4	Fisher	American	3	Stone	British
16	4	Fisher	American	4	Fisher	American

Example

- Find the names of all artists with the same nationality.

- Answer:

$$\rho_{a1}(artists)$$

$$\rho_{a2}(artists)$$

$$\pi_{a1.aName,a2.aName}(\sigma_{a1.nationality=a2.nationality\ AND\ a1.aID<a2.aID}(a1 X a2))$$

Movies:

mID	title	director	year	length
1	Shining	Kubrick	1980	146
3	Chinatown	Polaski	1974	131
5	Star Wars IV	Lucas	1977	126
6	American Graffiti	Lucas	1973	110

Note that I have changed the instance of table movies to demo the next operator we learn.

List all actors who play a role in every movie listed in the database.

Artists:

aID	aName	nationality
1	Nicholson	American
2	Ford	American
3	Stone	British
4	Fisher	American
5	Bachchan	Indian

Roles:

mID	aID	character
1	1	Jack Torrance
3	1	Jake ‘J.J.’ Gittes
1	3	Delbert Grady
5	1	Han Solo
6	1	Bob Falfa
5	4	Princess Leia Organa

Division

Note that to divide R1 by R2, every attribute in R2 must be an attribute in R1

- Example

R1

Col1	Col2
A	1
A	2
B	1
B	2
C	1

R2

Col2
1
2

R1 ÷ R2

Col1
A
B

Divide By

- List all actors who play a role in every movie listed in the database.

$$
\begin{aligned}
R1 &:= \pi_{mid,aid}(roles) \\
R2 &:= \pi_{mid}(movies) \\
R3 &:= R1 \div R2 \\
Ans &:= \pi_{aName}(R3 \infty artists)
\end{aligned}
$$

Set operations

- Because relations are sets, we can use set intersection, union and difference.
- Relations must be Union Compatible: Two relations R(A1,....An) and S(B1,...Bn) are said to be *Union Compatible* if they have the same degree n and Domain(Ai) = Domain(Bi) where $1 \leq i \leq n$
- You can make 2 relations union compatible by projecting (∏) the desired attributes, before applying any set operator.

Set operators

R ∪S: is a relation that includes all tuples that are either in R or in S or in both R and S, duplicate tuples being eliminated.

R ∩S: is a relation that includes all tuples that are in both R and in S.

R – S: is a relation that includes all tuples that are in R but not in S.

Set operators

Are Student and Instructor Union Compatible??

Student

STUDENT_ID	FIRST_NAME
101	HARRY
102	ARTHUR
103	YALE
104	ALISHA
105	SMITH
107	Nina
108	Rick
106	Anita

Instructor

FIRST_NAME	LAST_NAME
Fernand	Hanks
Tom	Wojick
Nina	Schorin
Gary	Pertez
Anita	Morris
Todd	Smythe
Marilyn	Frantzen
Charles	Lowry
Rick	Chow
Irene	Willig

- List all instructors that have the same name as a student's first name.
- List all instructors that do not have the same name as that of a student.
- List all students who do not have the same name as that of an instructor.

Relational algebra wrap-up

- Approaching the problem:
  - Ask yourself which relations need to be involved.
  - If an instance is given, use it to get results – that tells

    you a lot – the relations involved, operators required

    to get results etc.
  - Every time you combine relations, confirm that
    - attributes with same name will be made to match, unless otherwise stated in the design
      - Mostly, natural join will suffice when you join relations
      - If a self join is required, use cartesian product
      - Look for 'for all' or 'every' clause in the query – that should give you a hint that you might need to use divideBy (÷)

Relational Calculus

- Another abstract query language for the relational model.
- Based on first-order logic.
- RC is "declarative": the query describes what you want, but not how to get it.
- Queries look like this:
  $\{ t \mid t \in Movies \wedge t[director] = \text{“Scott”} \}$
- Expressive power (when limited to queries that generate finite results) is the same as RA. It is "relationally complete."

We will focus ONLY on RA for 3530 this term