

Relational Algebra

Eugene Wu

Supplemental Materials

Helpful References

https://en.wikipedia.org/wiki/Relational_algebra

Relational algebra is the basis for the most popular **query language on earth**

What's a Query Language?

Allows manipulation and **retrieval of data** from a database.

Traditionally: QL != programming language

Doesn't need to be turing complete

Not designed for computation

Supports easy, efficient access to (very) large databases

Today

Scaling to large datasets is a reality

Powerful way to think about...

data algorithms that scale

asynchronous/parallel programming

2 Formal Relational Query Languages

Relational Algebra

Operational, used to represent execution plans

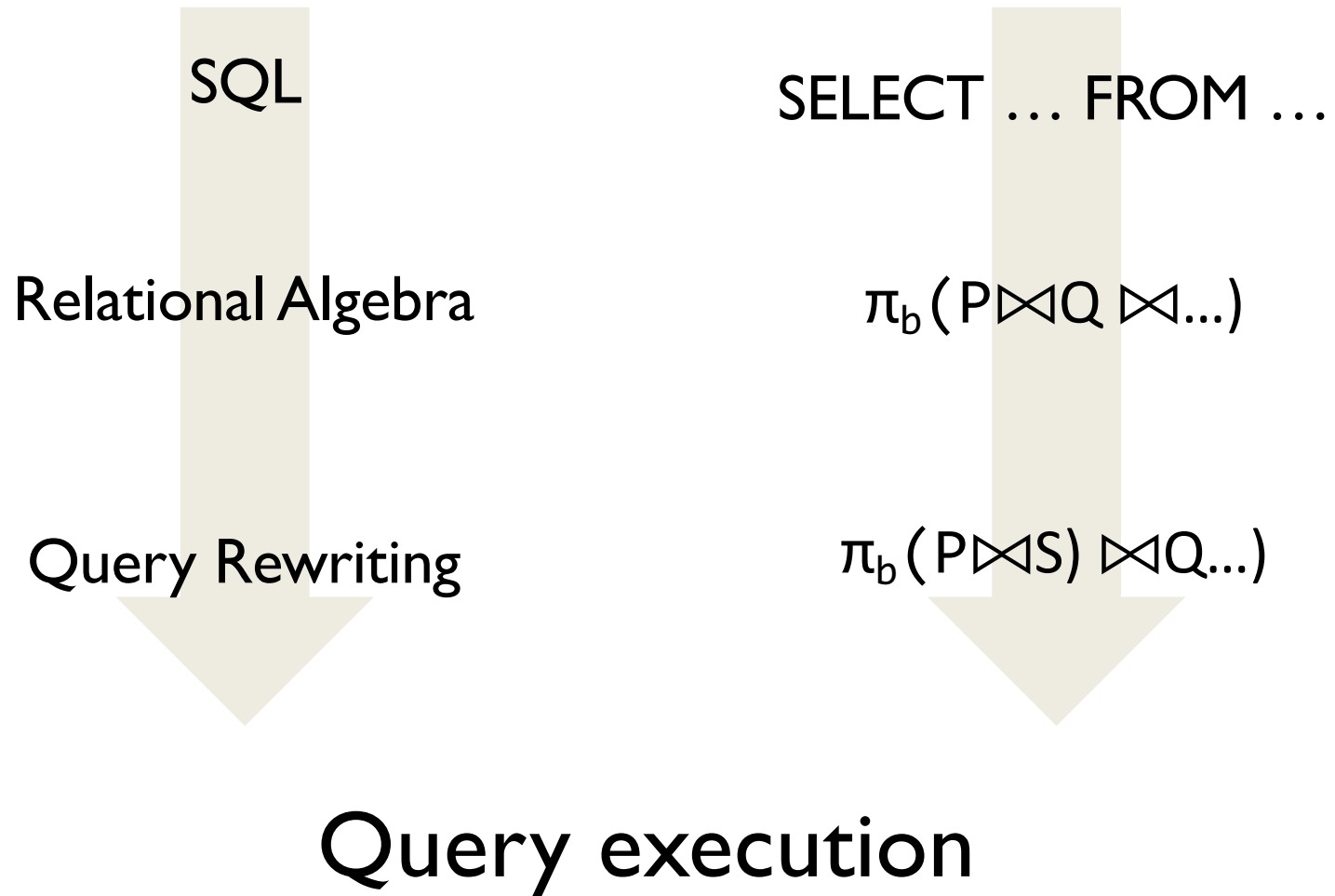
$\pi_{\text{name}}(\sigma_{\text{age} < 30}(\text{Sailors}))$ sailor names younger than 30

Relational Calculus

Logical, describes what data users want (declarative)

$\{ s: \text{name} \mid s \in \text{Sailors} \wedge s.\text{age} < 30 \}$
(this is shorthand)

Journey of a Query



Prelims

Query is a function over **relation instances**

$$Q(R_1, \dots, R_n) = R_{\text{result}}$$

Schemas of input and output relations are *fixed* and well defined by the query Q .

Positional vs Named field notation

- Position easier for formal defs

 - one-indexed (not 0-indexed!!!)

- Named is more readable

- Both used in SQL

Prelims

Relation (for this lecture)

Instance: **set** of tuples (important!)

Schema: list of field names and types (domains)

Students(sid int, name text, major text, gpa int)

How are relations different than generic sets (\mathbb{R})?

Can assume item structure due to schema

Some algebra operations (x) need to be modified

Will use this later

Relational Algebra Overview

Core 5 operations

PROJECT (π)

SELECT (σ)

UNION (\cup)

SET DIFFERENCE ($-$)

CROSSPRODUCT (\times)

Additional operations

RENAME (ρ)

INTERSECT (\cap)

JOIN (\bowtie)

DIVIDE ($/$) not on exam

Instances Used Today: Library

Students, Reservations

Use positional or named field notation

Fields in query results are inherited from input relations (unless specified)

RI

| sid | rid | day |
|-----|-----|-------|
| 1 | 101 | 10/10 |
| 2 | 102 | 11/11 |

S1

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

S2

| sid | name | gpa | age |
|-----|-------|-----|-----|
| 4 | aziz | 3.2 | 21 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |
| 5 | rusty | 3.5 | 21 |

Project

$$\pi_{\langle \text{attr1}, \dots \rangle}(A) = R_{\text{result}}$$

Pick out desired attributes (subset of columns)

Schema is subset of input schema in the projection list

$\pi_{\langle a, b, c \rangle}(A)$ has output schema (a, b, c) w/ types carried over

Project

S2

| sid | name | gpa | age |
|-----|-------|-----|-----|
| 4 | aziz | 3.2 | 21 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |
| 5 | rusty | 3.5 | 21 |

$\pi_{\text{name,age}}(S2) =$

| name | age |
|-------|-----|
| aziz | 21 |
| barb | 21 |
| tanya | 88 |
| rusty | 21 |

Project

S2

| sid | name | gpa | age |
|-----|-------|-----|-----|
| 4 | aziz | 3.2 | 21 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |
| 5 | rusty | 3.5 | 21 |

$\pi_{\text{name,name,age}}(S2) =$

| name | name | age |
|-------|-------|-----|
| aziz | aziz | 21 |
| barb | barb | 21 |
| tanya | tanya | 88 |
| rusty | rusty | 21 |

Project

S2

| sid | name | gpa | age |
|-----|-------|-----|-----|
| 4 | aziz | 3.2 | 21 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |
| 5 | rusty | 3.5 | 21 |

$\pi_{\text{age}}(S2) =$

| age |
|-----|
| 21 |
| 88 |

Where did all the rows go?

Real systems typically don't remove duplicates by default. Why?

Select

$$\sigma_{\langle p \rangle}(A) = R_{\text{result}}$$

Select subset of rows that satisfy condition p

p : Boolean expr over constants and attributes in A

Won't have duplicates in result. Why?

Result schema same as input

Select

S1

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

$\sigma_{\text{age} < 30} (S1) =$

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |

$\pi_{\text{name}}(\sigma_{\text{age} < 30} (S1)) =$

| name |
|--------|
| eugene |
| barb |

Commutative Operations

$$A + B = B + A$$

$$A * B = B * A$$

$$A + (B * C) = (B * C) + A$$

Associative Operations

$$A + (B + C) = (A + B) + C$$

$$A + (B * C) = (A + B) * C$$

Commutative Operations

$$A + B = B + A$$

$$A * B = B * A$$

$$A + (B * C) = (B * C) + A$$

Associative Operations

$$A + (B + C) = (A + B) + C$$

~~$$A + (B * C) = (A + B) * C$$~~

Commutatively

$$\pi_{\text{age}}(\sigma_{\text{age} < 30} (SI))$$

$\sigma_{\text{age} < 30}$

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |

Commutatively

$$\pi_{\text{age}}(\sigma_{\text{age} < 30}(SI))$$

$\sigma_{\text{age} < 30}$

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

π_{age}

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |

=

| age |
|-----|
| 20 |
| 21 |

Commutatively

$$\sigma_{\text{age} < 30}(\pi_{\text{age}}(S I))$$

π_{age}

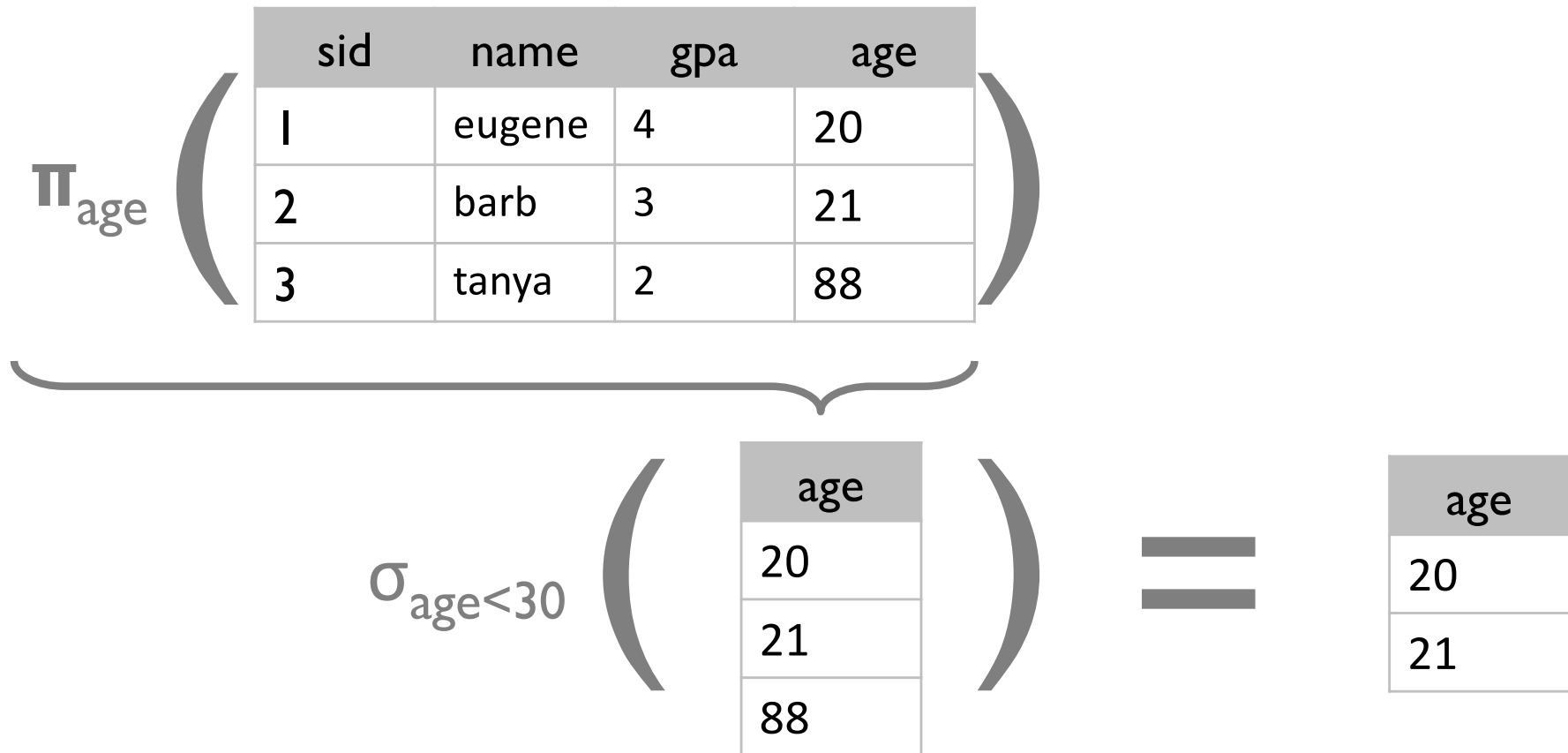
| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

$\sigma_{\text{age} < 30}$

| age |
|-----|
| 20 |
| 21 |
| 88 |

Commutatively

$$\sigma_{\text{age} < 30}(\pi_{\text{age}}(S))$$



Commutatively

Does Project and Select **always** commute?

$$\pi_{\text{age}}(\sigma_{\text{age} < 30}(SI)) = \sigma_{\text{age} < 30}(\pi_{\text{age}}(SI))$$

What about

$$\pi_{\text{name}}(\sigma_{\text{age} < 30}(SI))?$$

Commutatively

Does Project and Select commute?

$$\pi_{\text{age}}(\sigma_{\text{age} < 30}(SI)) = \sigma_{\text{age} < 30}(\pi_{\text{age}}(SI))$$

What about

$$\pi_{\text{name}}(\sigma_{\text{age} < 30}(SI)) \neq \sigma_{\text{age} < 30}(\pi_{\text{name}}(SI))$$

Commutatively

Does Project and Select commute?

$$\pi_{\text{age}}(\sigma_{\text{age} < 30}(SI)) = \sigma_{\text{age} < 30}(\pi_{\text{age}}(SI))$$

What about

$$\pi_{\text{name}}(\sigma_{\text{age} < 30}(SI)) \neq \sigma_{\text{age} < 30}(\pi_{\text{name, age}}(SI))$$

Commutatively

Does Project and Select commute?

$$\pi_{\text{age}}(\sigma_{\text{age} < 30}(SI)) = \sigma_{\text{age} < 30}(\pi_{\text{age}}(SI))$$

What about

$$\pi_{\text{name}}(\sigma_{\text{age} < 30}(SI)) = \pi_{\text{name}}(\sigma_{\text{age} < 30}(\pi_{\text{name, age}}(SI)))$$

OK!

Union, Set-Difference

$$A \text{ op } B = R_{\text{result}}$$

A, B must be *union-compatible*

Same number of fields

Field i in each schema have same type

Result Schema taken from first relation (A)

A(id int, imgid int) \cup B(blah int, gloop int) = ?

Union, Set-Difference

$$A \text{ op } B = R_{\text{result}}$$

A, B must be *union-compatible*

Same number of fields

Field i in each schema have same type

Result Schema taken from first relation (A)

$$A(\text{id int, imgid int}) \cup B(\text{blah int, gloop int}) = \\ R_{\text{result}}(\text{id int, imgid int})$$

Union, Intersect, Set-Difference

S1

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

S2

| sid | name | gpa | age |
|-----|-------|-----|-----|
| 4 | aziz | 3.2 | 21 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |
| 5 | rusty | 3.5 | 21 |

$S1 \cup S2 =$

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 4 | aziz | 3.2 | 21 |
| 5 | rusty | 3.5 | 21 |
| 3 | tanya | 2 | 88 |
| 2 | barb | 3 | 21 |

Union, Intersect, Set-Difference

S1

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

S2

| sid | name | gpa | age |
|-----|-------|-----|-----|
| 4 | aziz | 3.2 | 21 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |
| 5 | rusty | 3.5 | 21 |

$S1 - S2 =$

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |

Note on Set Difference & Performance

Notice that most operators are monotonic
increasing size of inputs \rightarrow outputs grow
if $A \supseteq B \rightarrow Q(A, T) \supseteq Q(B, T)$
can compute *incrementally*

Set Difference is *not monotonic*

if $A \supseteq B \rightarrow T - A \subseteq T - B$
e.g., $5 > 1 \rightarrow 9 - 5 < 9 - 1$

Set difference is *blocking*:

For $T - S$, must wait for all S tuples before any results

Cross-Product

$$A(a_1, \dots, a_n) \times B(a_{n+1}, \dots, a_m) = R_{\text{result}}(a_1, \dots, a_m)$$

Each row of A paired with each row of B

Result schema **concat**s A and B's fields, inherit if possible

Names of fields found in both A and B are undefined in result
(some DBMSes set a default)

Not same as mathematical “X”, which returns **nested** results:

$$\text{math } A \times B = \{ (a, b) \mid a \in A \wedge b \in B \}$$

$$\{1, 2\} \times \{3, 4\} = \{ (1, 3), (1, 4), (2, 3), (2, 4) \}$$

what is $\{1, 2\} \times \{3, 4\} \times \{5, 6\}$?

$$\{ (1, 3), (1, 4), (2, 3), (2, 4) \} \times \{5, 6\} = \{ ((1, 3), 5), \dots \}$$

Cross-Product

SI

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

RI

| sid | rid | day |
|-----|-----|-------|
| 1 | 101 | 10/10 |
| 2 | 102 | 11/11 |

SI x RI =

| (sid) | name | gpa | age | (sid) | rid | day |
|-------|--------|-----|-----|-------|-----|-------|
| 1 | eugene | 4 | 20 | 1 | 101 | 10/10 |
| 2 | barb | 3 | 21 | 1 | 101 | 10/10 |
| 3 | tanya | 2 | 88 | 1 | 101 | 10/10 |
| 1 | eugene | 4 | 20 | 2 | 102 | 11/11 |
| 2 | barb | 3 | 21 | 2 | 102 | 11/11 |
| 3 | tanya | 2 | 88 | 2 | 102 | 11/11 |

Rename

$p(<\text{newRelationName}>(<\text{mappings}>), Q)$

Explicitly defines/changes field names of schema

$p(C(I \rightarrow \text{sid1}, 5 \rightarrow \text{sid2}), S1 \times R1)$

C =

| sid1 | name | gpa | age | sid2 | rid | day |
|------|--------|-----|-----|------|-----|-------|
| 1 | eugene | 4 | 20 | 1 | 101 | 10/10 |
| 2 | barb | 3 | 21 | 1 | 101 | 10/10 |
| 3 | tanya | 2 | 88 | 1 | 101 | 10/10 |
| 1 | eugene | 4 | 20 | 2 | 102 | 11/11 |
| 2 | barb | 3 | 21 | 2 | 102 | 11/11 |
| 3 | tanya | 2 | 88 | 2 | 102 | 11/11 |

Project



Select



Cross product



Difference



Union



Intersect



Compound/Convenience Operators

INTERSECT (\cap)

JOIN (\bowtie)

DIVIDE ($/$)

Intersect

$$A \cap B = R_{\text{result}}$$

A, B must be *union-compatible*

Intersect

S1

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

S2

| sid | name | gpa | age |
|-----|-------|-----|-----|
| 4 | aziz | 3.2 | 21 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |
| 5 | rusty | 3.5 | 21 |

$S1 \cap S2 =$

| sid | name | gpa | age |
|-----|-------|-----|-----|
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

Intersect

$$A \cap B = R_{\text{result}}$$

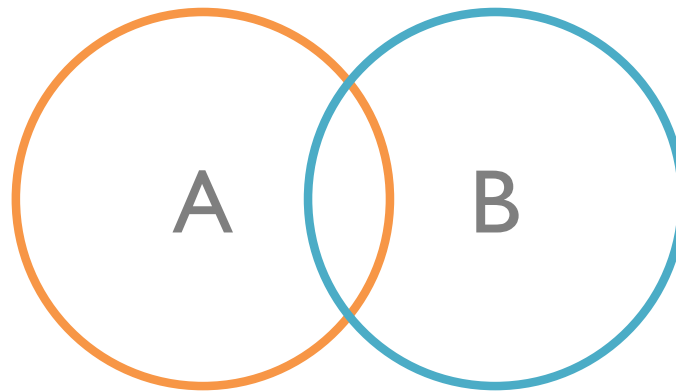
A, B must be *union-compatible*

Can we express using core operators?

$$A \cap B = ?$$

Intersect

$$A \cap B = R_{\text{result}}$$

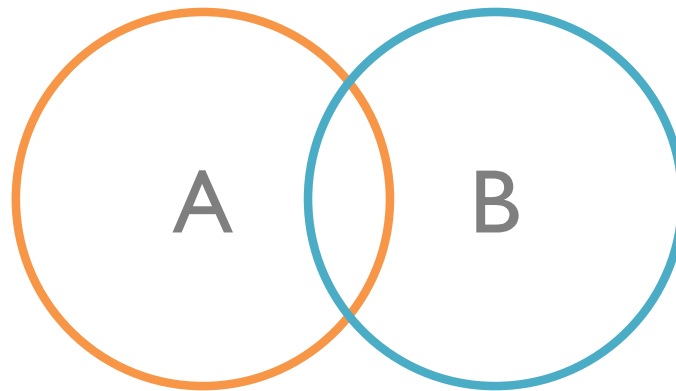


Can we express using core operators?

$A \cap B = A - ?$ (think venn diagram)

Intersect

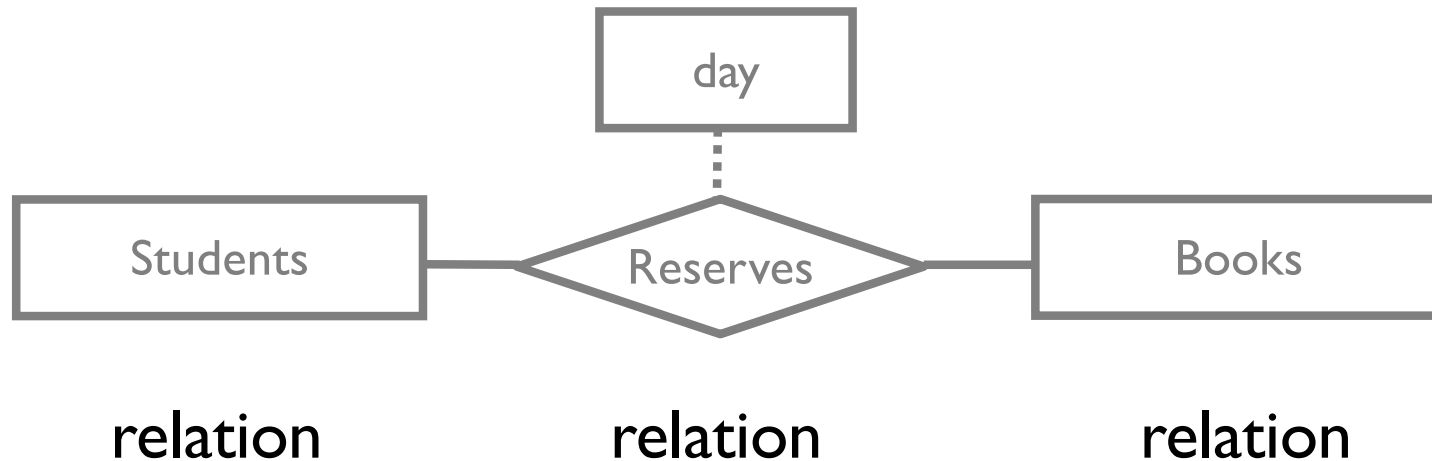
$$A \cap B = R_{\text{result}}$$



Can we express using core operators?

$$A \cap B = A - (A - B)$$

Joins (high level)

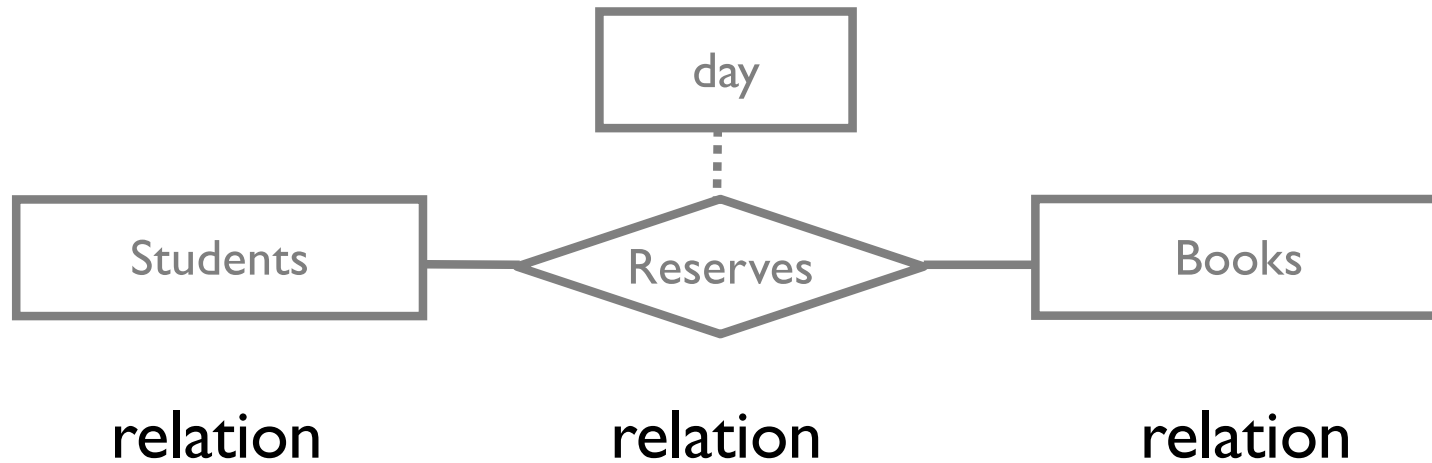


What if you want to query across all three tables?
e.g., all names of students that reserved “The Purple Crayon”

Need to combine these tables

Cross product? But that ignores foreign key references

Joins (high level)



SI

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

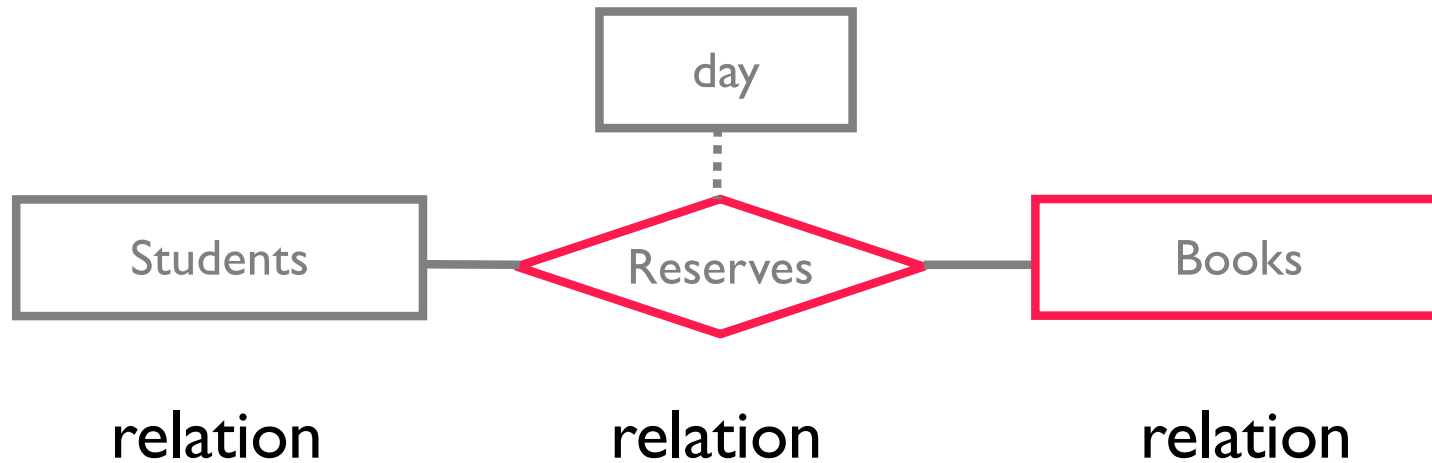
RI

| sid | rid | day |
|-----|-----|-------|
| 1 | 101 | 10/10 |
| 2 | 102 | 11/11 |

BI

| rid | name |
|-----|-------------------|
| 101 | The Purple Crayon |
| 102 | 1984 |

Joins (high level)



SI

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

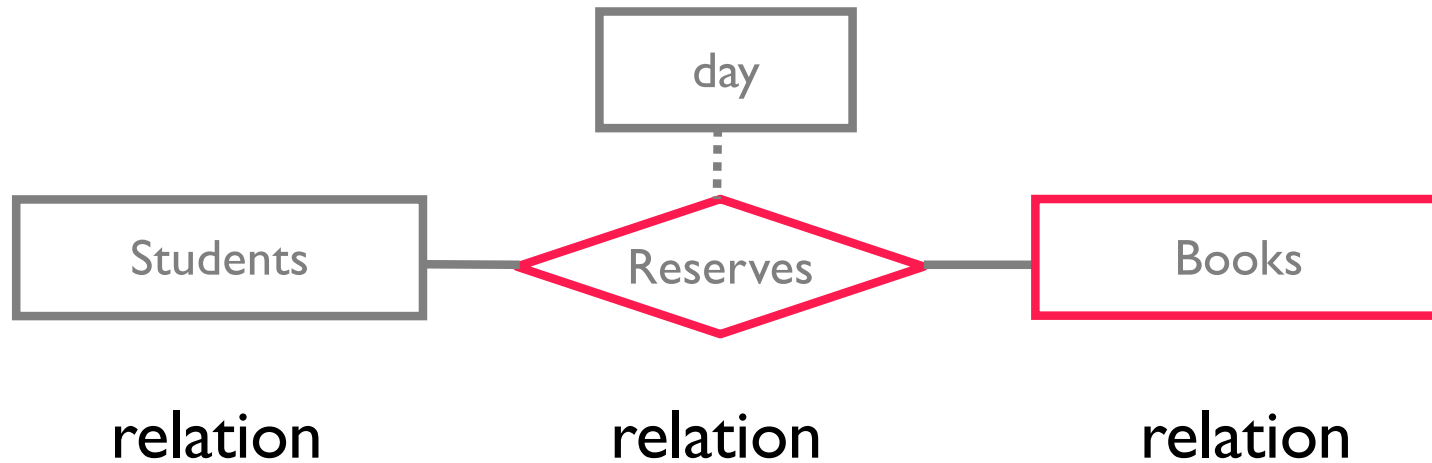
RI

| sid | rid | day |
|-----|-----|-------|
| 1 | 101 | 10/10 |
| 2 | 102 | 11/11 |

BI

| rid | name |
|-----|-------------------|
| 101 | The Purple Crayon |
| 102 | 1984 |

Joins (high level)



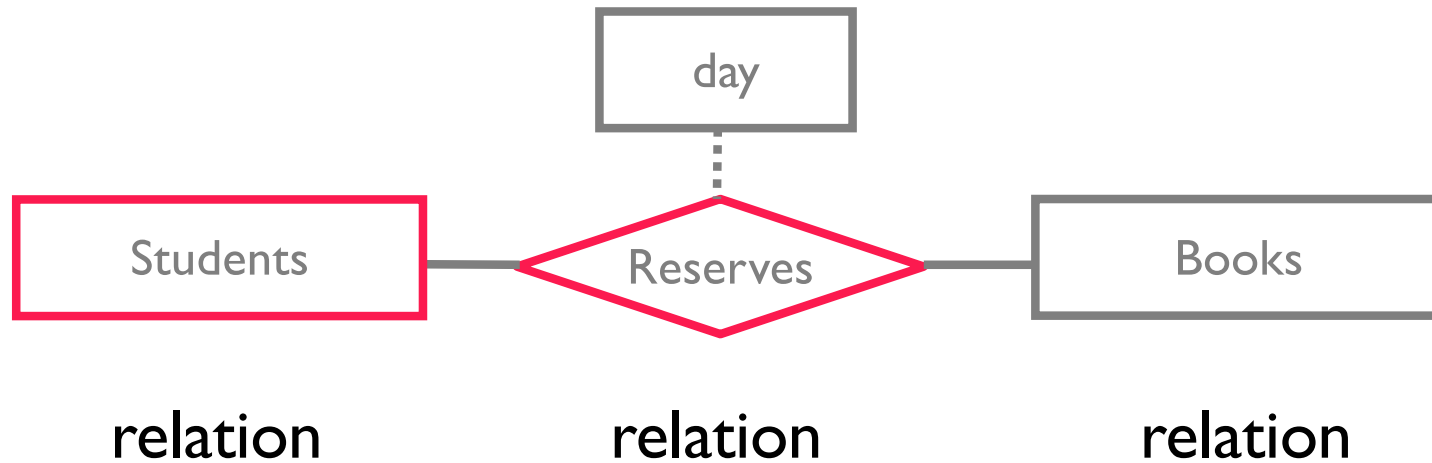
SI

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

RBI

| sid | (rid) | day | (rid) | name |
|-----|-------|-------|-------|-------------------|
| 1 | 101 | 10/10 | 101 | The Purple Crayon |
| 2 | 102 | 11/11 | 102 | 1984 |

Joins (high level)



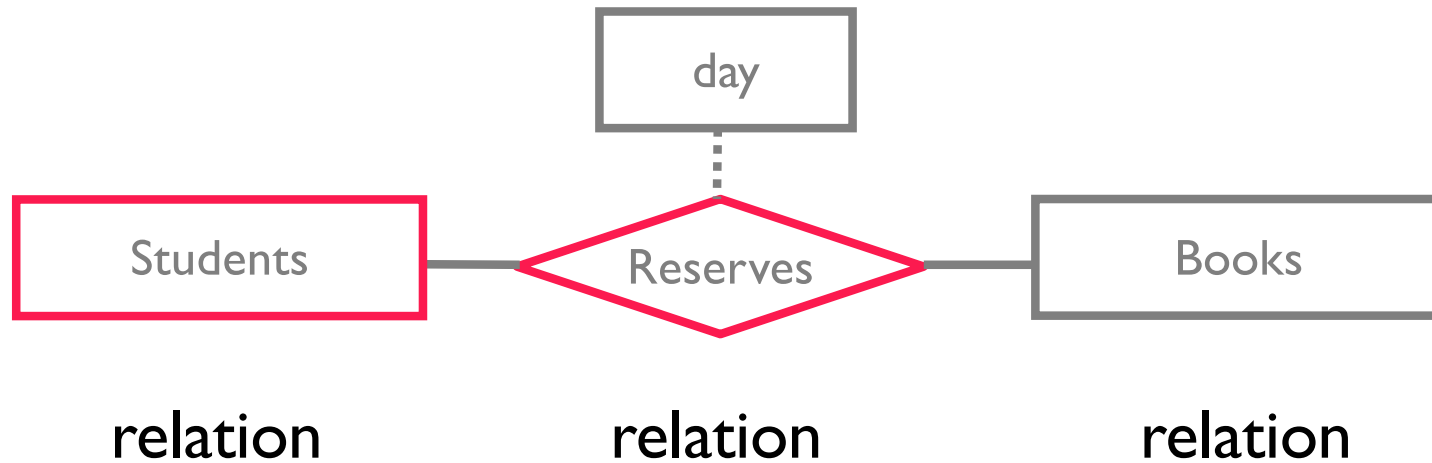
SI

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

RBI

| sid | (rid) | day | (rid) | name |
|-----|-------|-------|-------|-------------------|
| 1 | 101 | 10/10 | 101 | The Purple Crayon |
| 2 | 102 | 11/11 | 102 | 1984 |

Joins (high level)



SRBI

| (sid) | (name) | gpa | age | (sid) | (rid) | day | (rid) | (name) |
|-------|--------|-----|-----|-------|-------|-------|-------|-------------------|
| 1 | eugene | 4 | 20 | 1 | 101 | 10/10 | 101 | The Purple Crayon |
| 2 | barb | 3 | 21 | 2 | 102 | 11/11 | 102 | 1984 |

Joins

Theta (θ) Join

Equi-join

theta (θ) Join

$$A \bowtie_c B = \sigma_c(A \times B)$$

Most general form

Result schema same as cross product

Often *far* more efficient to compute than cross product

Commutative

$$(A \bowtie_c B) \bowtie_c C = A \bowtie_c (B \bowtie_c C)$$

theta (θ) Join

SI

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

RI

| sid | rid | day |
|-----|-----|-------|
| 1 | 101 | 10/10 |
| 2 | 102 | 11/11 |

$$SI \bowtie_{SI.sid \leq RI.sid} RI =$$

$$\sigma_{SI.sid \leq RI.sid}(SI \times RI) =$$

| (sid) | name | gpa | age | (sid) | rid | day |
|-------|--------|-----|-----|-------|-----|-------|
| 1 | eugene | 4 | 20 | 1 | 101 | 10/10 |
| 2 | barb | 3 | 21 | 1 | 101 | 10/10 |
| 3 | tanya | 2 | 88 | 1 | 101 | 10/10 |
| 1 | eugene | 4 | 20 | 2 | 102 | 11/11 |
| 2 | barb | 3 | 21 | 2 | 102 | 11/11 |
| 3 | tanya | 2 | 88 | 2 | 102 | 11/11 |

theta (θ) Join

SI

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

RI

| sid | rid | day |
|-----|-----|-------|
| 1 | 101 | 10/10 |
| 2 | 102 | 11/11 |

$$SI \bowtie_{SI.sid \leq RI.sid} RI =$$

$$\sigma_{SI.sid \leq RI.sid}(SI \times RI) =$$

| (sid) | name | gpa | age | (sid) | rid | day |
|-------|--------|-----|-----|-------|-----|-------|
| 1 | eugene | 4 | 20 | 1 | 101 | 10/10 |
| 2 | barb | 3 | 21 | 1 | 101 | 10/10 |
| 3 | tanya | 2 | 88 | 1 | 101 | 10/10 |
| 1 | eugene | 4 | 20 | 2 | 102 | 11/11 |
| 2 | barb | 3 | 21 | 2 | 102 | 11/11 |
| 3 | tanya | 2 | 88 | 2 | 102 | 11/11 |

theta (θ) Join

SI

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

RI

| sid | rid | day |
|-----|-----|-------|
| 1 | 101 | 10/10 |
| 2 | 102 | 11/11 |

$$SI \bowtie_{SI.sid \leq RI.sid} RI =$$

$$\sigma_{SI.sid \leq RI.sid}(SI \times RI) =$$

| (sid) | name | gpa | age | (sid) | rid | day |
|-------|--------|-----|-----|-------|-----|-------|
| 1 | eugene | 4 | 20 | 1 | 101 | 10/10 |
| 1 | eugene | 4 | 20 | 2 | 102 | 11/11 |
| 2 | barb | 3 | 21 | 2 | 102 | 11/11 |

Equi-Join

$$A \bowtie_{\text{attr}} B = \pi_{\text{all attrs except B.attr}}(A \bowtie_{A.\text{attr} = B.\text{attr}} B)$$

List the attributes that the two relations will be joined on explicitly.

$A \bowtie_{x,y} B$ is an equijoin on attributes x and y

Special case where the condition is attribute equality

Result schema only keeps *one copy* of equality fields

Natural Join ($A \bowtie B$):

Equijoin on *all* shared fields (fields w/ same name)

Not recommended since query results can unexpectedly change if someone changes the schemas (renames an attr)

Equi-Join

SI

| sid | name | gpa | age |
|-----|--------|-----|-----|
| 1 | eugene | 4 | 20 |
| 2 | barb | 3 | 21 |
| 3 | tanya | 2 | 88 |

RI

| sid | rid | day |
|-----|-----|-------|
| 1 | 101 | 10/10 |
| 2 | 102 | 11/11 |

$SI \bowtie_{sid} RI$

| sid | name | gpa | age | rid | day |
|-----|--------|-----|-----|-----|-------|
| 1 | eugene | 4 | 20 | 101 | 10/10 |
| 2 | barb | 3 | 21 | 102 | 11/11 |

Equi-Join

SI

| sid | name | day | gpa | age |
|-----|--------|-------|-----|-----|
| 1 | eugene | 10/10 | 4 | 20 |
| 2 | barb | 12/12 | 3 | 21 |
| 3 | tanya | 3/3 | 2 | 88 |

RI

| sid | rid | day |
|-----|-----|-------|
| 1 | 101 | 10/10 |
| 2 | 102 | 11/11 |

$SI \bowtie_{sid, name} RI = \text{INVALID! name not in RI}$

$SI \bowtie_{sid, day} RI = \pi_{SI.*, RI.rid} SI \bowtie_{SI.sid=RI.sid \wedge SI.day=RI.day} RI$

| sid | name | day | gpa | age | rid |
|-----|--------|-------|-----|-----|-----|
| 1 | eugene | 10/10 | 4 | 20 | 101 |

$SI \bowtie RI = SI \bowtie_{sid, day} RI$

Different Plans, Same Results

Semantic equivalence:
results are *always* the same

Note that it is independent
of the database instance!

Names of students that reserved book 2

$$\pi_{\text{name}}(\sigma_{\text{rid}=2} (R1) \bowtie SI)$$

Equivalent Queries

$$\begin{aligned} & p(\text{tmp1}, \sigma_{\text{rid}=2} (R1)) \\ & p(\text{tmp2}, \text{tmp1} \bowtie SI) \\ & \pi_{\text{name}}(\text{tmp2}) \end{aligned}$$
$$\pi_{\text{name}}(\sigma_{\text{rid}=2}(R1 \bowtie SI))$$

Names of students that reserved db books

Book(rid, type) Reserve(sid, rid) Student(sid, name)

Need to join DB books with reserve and students

$\sigma_{\text{type}='db'}$ (Book)

Names of students that reserved db books

Book(rid, type) Reserve(sid, rid) Student(sid, name)

Need to join DB books with reserve and students

$\sigma_{\text{type}='db'} (\text{Book}) \bowtie \text{Reserve}$

Names of students that reserved db books

Book(rid, type) Reserve(sid, rid) Student(sid, name)

Need to join DB books with reserve and students

$\sigma_{\text{type}='db'} (\text{Book}) \bowtie \text{Reserve} \bowtie \text{Student}$

Names of students that reserved db books

Book(rid, type) Reserve(sid, rid) Student(sid, name)

Need to join DB books with reserve and students

$\pi_{\text{name}}(\sigma_{\text{type}='db'}(\text{Book}) \bowtie \text{Reserve} \bowtie \text{Student})$

Names of students that reserved db books

Book(rid, type) Reserve(sid, rid) Student(sid, name)

Need to join DB books with reserve and students

$\pi_{\text{name}}(\sigma_{\text{type}='db'}(\text{Book}) \bowtie \text{Reserve} \bowtie \text{Student})$

More efficient query

$\pi_{\text{name}}(\pi_{\text{sid}}((\pi_{\text{rid}} \sigma_{\text{type}='db'}(\text{Book})) \bowtie \text{Reserve}) \bowtie \text{Student})$

Query optimizer can find the more efficient query!

Names of students that reserved db books

Book(rid, type) Reserve(sid, rid) Student(sid, name)

Need to join DB books with reserve and students

$\pi_{\text{name}}(\sigma_{\text{type}='db'}(\text{Book}) \bowtie \text{Reserve} \bowtie \text{Student})$

More efficient query

$\pi_{\text{name}}(\pi_{\text{sid}}((\pi_{\text{rid}} \sigma_{\text{type}='db'}(\text{Book})) \bowtie \text{Reserve}) \bowtie \text{Student})$

Query optimizer can find the more efficient query!

Names of students that reserved db books

Book(rid, type) Reserve(sid, rid) Student(sid, name)

Need to join DB books with reserve and students

$\pi_{\text{name}}(\sigma_{\text{type}='db'}(\text{Book}) \bowtie \text{Reserve} \bowtie \text{Student})$

More efficient query

$\pi_{\text{name}}(\pi_{\text{sid}}((\pi_{\text{rid}} \sigma_{\text{type}='db'}(\text{Book})) \bowtie \text{Reserve}) \bowtie \text{Student})$

Query optimizer can find the more efficient query!

Names of students that reserved db books

Book(rid, type) Reserve(sid, rid) Student(sid, name)

Need to join DB books with reserve and students

$\pi_{\text{name}}(\sigma_{\text{type}='db'}(\text{Book}) \bowtie \text{Reserve} \bowtie \text{Student})$

More efficient query

$\pi_{\text{name}}(\pi_{\text{sid}}((\pi_{\text{rid}} \sigma_{\text{type}='db'}(\text{Book})) \bowtie \text{Reserve}) \bowtie \text{Student})$

Query optimizer can find the more efficient query!

Students that reserved DB or HCI book

1. Find all DB or HCI books
2. Find students that reserved one of those books

$\rho(\text{tmp}, (\sigma_{\text{type}='DB' \vee \text{type}='HCI'} (\text{Book})))$
 $\pi_{\text{name}}(\text{tmp} \bowtie \text{Reserve} \bowtie \text{Student})$

” \vee ” means logical OR

Alternatives

define tmp using UNION (how?)

Using UNION

$p(tmp1, (\sigma_{type='DB'} (Book)))$

$p(dbnames, \pi_{name}(tmp1 \bowtie Reserve \bowtie Student))$

$p(tmp2, (\sigma_{type='HCI'} (Book)))$

$p(hcinames, \pi_{name}(tmp2 \bowtie Reserve \bowtie Student))$

$dbnames \cup hcinames$

Students that reserved a DB and HCI book

Can we change \vee into \wedge (AND)?

$\rho(\text{tmp}, (\sigma_{\text{type}='DB' \wedge \text{type}='HCI'}(\text{Book})))$
 $\pi_{\text{name}}(\text{tmp} \bowtie \text{Reserve} \bowtie \text{Student})$

NO

Why?

$\rho(\text{tmp}, (\sigma_{\text{type}='DB' \wedge \text{type}='HCI'} (\text{Book})))$
 $\pi_{\text{name}}(\text{tmp} \bowtie \text{Reserve} \bowtie \text{Student})$

for b in Book:

if b.type = 'DB' and b.type = 'HCI': // resolves to FALSE

for r in Reserve:

for s in Student:

if r.sid = s.sid and r.bid = b.bid:

yield b.name

Students that reserved a DB and HCI book

Does previous approach work?

1. Find students that reserved DB books
2. Find students that reserved HCI books
3. Intersection

$$\begin{aligned} & \rho(\text{tmpDB}, \pi_{\text{sid}}(\sigma_{\text{type}=\text{'DB'}} \text{Book}) \bowtie \text{Reserve}) \\ & \rho(\text{tmpHCI}, \pi_{\text{sid}}(\sigma_{\text{type}=\text{'HCI'}} \text{Book}) \bowtie \text{Reserve}) \\ & \pi_{\text{name}}((\text{tmpDB} \cap \text{tmpHCI}) \bowtie \text{Student}) \end{aligned}$$

Students that reserved all books

Students where, **for all books**, the student reserved the book
no concept of “for all” in relational algebra...

Students – Students that didn't reserve all books

Students that reserved all books

Students where, **for all books**, the student reserved the book
no concept of “for all” in relational algebra...

Students – Students where there is a book that they did not reserve

Students that reserved all books

Students where, **for all books**, the student reserved the book
no concept of “for all” in relational algebra...

Students – (Students s where (Books – Books s reserved)) (say s is bob)
$$p(s_reserved, \pi_{bid} \sigma_{sid=bob}(Reserve))$$

Students – (Students s where (Books – Books s reserved))
$$p(s_not_reserved, \pi_{bid}(Books) - s_reserved)$$

Students – (Students s where (Books – Books s reserved)) (for each student)
$$\pi_{sid,bid}(Students \times Books)$$

Students that reserved all books

Students where, **for all books**, the student reserved the book
no concept of “for all” in relational algebra...

Students – (Students s where (Books – Books s reserved))
$$p(s_reserved, \pi_{bid} \sigma_{sid=bob}(Reserve))$$

Students – (Students s where (Books – Books s reserved))
$$p(s_not_reserved, \pi_{bid}(Books) - s_reserved)$$

Students – (Students s where (Books – Books s reserved)) (for each student)
$$\pi_{sid,bid}(Students \times Books) - \pi_{sid,bid}(Reserve)$$

Students that reserved all books

Students where, **for all books**, the student reserved the book
no concept of “for all” in relational algebra...

Students – (Students s where (Books – Books s reserved))
 $\rho(s_reserved, \pi_{bid} \sigma_{sid=bob}(Reserve))$

Students – (Students s where (Books – Books s reserved))
 $\rho(s_not_reserved, \pi_{bid}(Books) - s_reserved)$

Students – (Students s where (Books – Books s reserved)) (for each student)
 $\rho(del_sids, \pi_{sid}(\pi_{sid,bid}(Students \times Books) - \pi_{sid,bid}(Reserve)))$

Students that reserved all books

Students where, **for all books**, the student reserved the book
no concept of “for all” in relational algebra...

Students – (Students s where (Books – Books s reserved))
 $\rho(s_reserved, \pi_{bid} \sigma_{sid=bob}(Reserve))$

Students – (Students s where (Books – Books s reserved))
 $\rho(s_not_reserved, \pi_{bid}(Books) - s_reserved)$

Students – (Students s where (Books – Books s reserved)) (for each student)
 $\rho(del_sids, \pi_{sid}(\pi_{sid,bid}(Students \times Books) - \pi_{sid,bid}(Reserve)))$

$\pi_{sid}(Students) - del_sids$

Let's step back

Relational algebra is expressiveness benchmark

A language that can express relational algebra is
“relationally complete”

Limitations

nulls

aggregation

recursion

duplicates

can't really type on keyboard...

Equi-Joins are everywhere

Matching of two sets based on shared attributes

Yelp: Join between your location and restaurants

Market: Join between consumers and suppliers

High five: Join between two hands on time and space

Comm.: Join between minds on ideas/concepts

PLANES



ARE JOINS BETWEEN CITIES

imgflip.com

Who Cares about Relational Alg?

Clean query semantics & rich program analysis

Helps/enables optimization

Opens up rich set of topics

- Materialized views

- Data lineage/provenance

- Query by example

- Distributed query execution

- ...

You see its fingerprints **EVERYWHERE!**

What can we do with RA?

Query(DB instance) \rightarrow Relation instance

What can we do with RA?

Query(**DB instance**) = **Relation instance**

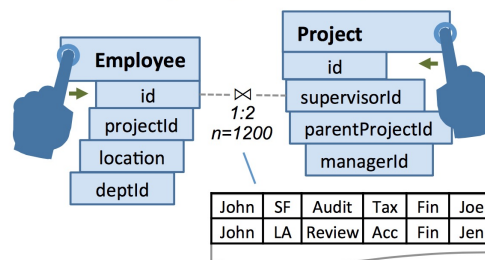
Query by example

Here's **DB instance** and **result**, generate the **query**

Data Generation:

Here's **query** and **result**, generate a **DB instance**

Novel relationally
complete interfaces



GestureDB. Nandi et al.

Summary

Relational Algebra (RA) operators

Operators are closed

inputs & outputs are relations

Multiple Relational Algebra queries can be equivalent

It is operational

Same semantics but different performance

Forms basis for optimizations