

L4

Relational Algebra

Eugene Wu

Fall 2015

Overview

Last time, learned about
pre-relational models
an informal introduction to relational model
an introduction to the SQL query language.

Learn about formal relational query languages
Relational Algebra (algebra: perform operations)
Relational Calculus (logic: are statements true?)

Keys to understanding SQL and query processing

Who Cares?

Query Semantics

Concise *domain specific language (DSL)*

Rich program analysis

Helps/enables optimization

Opens up rich set of topics

Materialized views

Data lineage/provenance

Query by example

...

Reading

Ramakrishnan

Optional

Helpful References

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

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 large databases

Recent Years

- Scaling to large datasets is a reality

- Query languages are a powerful way to

 - think about data algorithms scale

 - think about asynchronous/parallel programming

Formal Relational Query Languages

Formal basis for real languages e.g., SQL

Relational Algebra

Operational, used to represent execution plans

Relational Calculus

Logical, describes what data users want
(not operational, fully declarative)

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 more readable

- Both used in SQL

Prelims

Relation (for this lecture)

Instance is a set of tuples

Schema defines field names and types (domains)

Students(sid, name, bday, major, gpa)

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 ($/$)

Instances Used Today: Library

Students, Reservations,
Books

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

SI

sid	name	gpa	age
1	eugene	4	20
2	barak	3	21
3	trump	2	88

S2

sid	name	gpa	age
4	aziz	3.2	21
2	barak	3	21
3	trump	2	88
5	rusty	3.5	21

Project

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

Pick out desired attributes (subset of columns)

Schema is subset of input schema in the projection list

$\pi_{\langle a, b, c \rangle}(R_{\text{in}})$ has output schema (a,b,c) with types carried over

Project

S2

sid	name	gpa	age
4	aziz	3.2	21
2	barak	3	21
3	trump	2	88
5	rusty	3.5	21

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

name	age
aziz	21
barak	21
trump	88
rusty	21

Project

S2

sid	name	gpa	age
4	aziz	3.2	21
2	barak	3	21
3	trump	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. Why?

Select

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

Select subset of rows that satisfy condition p

Won't have duplicates in result. Why?

Result schema same as input

Select

S1

sid	name	gpa	age
1	eugene	4	20
2	barak	3	21
3	trump	2	88

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

sid	name	gpa	age
1	eugene	4	20
2	barak	3	21

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

name
eugene
barak

Union, Set-Difference

$$R_1 \text{ op } R_2 = R_{\text{result}}$$

R_1, R_2 must be *union-compatible*

Same number of fields

Field i in each schema have same type

Result Schema borrowed from first arg (R_1)

$R_1(\text{big int, poppa int}) \cup R_2(\text{thug int, life int}) = ?$

Union, Intersect, Set-Difference

S1

sid	name	gpa	age
1	eugene	4	20
2	barak	3	21
3	trump	2	88

S2

sid	name	gpa	age
4	aziz	3.2	21
2	barak	3	21
3	trump	2	88
5	rusty	3.5	21

$S1 \cup S2 =$

sid	name	gpa	age
1	eugene	4	20
4	aziz	3.2	21
2	barak	3	21
3	trump	2	88
5	rusty	3.5	21

Union, Intersect, Set-Difference

S1

sid	name	gpa	age
1	eugene	4	20
2	barak	3	21
3	trump	2	88

S2

sid	name	gpa	age
4	aziz	3.2	21
2	barak	3	21
3	trump	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 $S1 \supseteq S2 \rightarrow Q(S1, T) \supseteq Q(S2, T)$
can compute *incrementally*

Set Difference is *not monotonic*

e.g., $T - S1$

if $S1 \supseteq S2 \rightarrow T - S1 \subseteq T - S2$

Set difference is *blocking*:

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

Cross-Product

$$R_1(a_1, \dots, a_n) \times R_2(a_{n+1}, \dots, a_m) = R_{\text{result}}(a_1, \dots, a_m)$$

Each row of S1 paired with each row of R1

Result schema concatenates S1 and R1's fields, inherit if possible

Conflict: S1 and R1 have *sid* field

Different than mathematical “X” by flattening results:

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

$$\text{e.g., } \{1, 2\} \times \{3, 4\} = \{ (1, 3), (1, 4), (2, 3), (2, 4) \}$$

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

Cross-Product

SI

sid	name	gpa	age
1	eugene	4	20
2	barak	3	21
3	trump	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	barak	3	21	1	101	10/10
3	trump	2	88	1	101	10/10
1	eugene	4	20	2	102	11/11
2	barak	3	21	2	102	11/11
3	trump	2	88	2	102	11/11

Rename

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

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	barak	3	21	1	101	10/10
3	trump	2	88	1	101	10/10
1	eugene	4	20	2	102	11/11
2	barak	3	21	2	102	11/11
3	trump	2	88	2	102	11/11

Compound/Convenience Operators

Convenience operations

INTERSECT (\cap)

JOIN (\bowtie)

DIVIDE ($/$)

Intersect

$$R_1 \cap R_2 = R_{\text{result}}$$

R_1, R_2 must be *union-compatible*

Intersect

S1

sid	name	gpa	age
1	eugene	4	20
2	barak	3	21
3	trump	2	88

S2

sid	name	gpa	age
4	aziz	3.2	21
2	barak	3	21
3	trump	2	88
5	rusty	3.5	21

$S1 \cap S2 =$

sid	name	gpa	age
2	barak	3	21
3	trump	2	88

Intersect

$$R_1 \cap R_2 = R_{\text{result}}$$

R_1, R_2 must be *union-compatible*

Can we express using core operators?

$$S \cap T = ?$$

Intersect

$$R_1 \cap R_2 = R_{\text{result}}$$

R_1, R_2 must be *union-compatible*

Can we express using core operators?

$S \cap T = S - ?$ (think venn diagram)

Intersect

$$R_1 \cap R_2 = R_{\text{result}}$$

R_1, R_2 must be *union-compatible*

Can we express using core operators?

$$S \cap T = S - (S - T)$$

theta (θ) Join

$$R \bowtie_c S = \sigma_c(R \times S)$$

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	barak	3	21
3	trump	2	88

RI

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

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

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

Equi-Join

$$R \bowtie_{\text{attr}} S = R \bowtie_{R.\text{attr} = S.\text{attr}} S$$

Special case where the condition is attribute equality

Result schema only keeps one copy of equality fields

Natural Join ($R \bowtie S$):

Equijoin on *all* shared fields

Equi-Join

SI

sid	name	gpa	age
1	eugene	4	20
2	barak	3	21
3	trump	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	barak	3	21	102	11/11

Division

Let us have relations $A(x, y)$, $B(y)$

$$A/B = \{ \langle x \rangle \mid \exists \langle x, y \rangle \in A \ \forall \langle y \rangle \in B \}$$

Good to ponder, not supported in most systems (why?)

Find all students that have reserved all books

$A/B =$ all x (students) s.t. for every y (reservation), $\langle x, y \rangle \in A$

Generalization

y can be a list of fields in B

$x \cup y$ is fields in A

Examples

A

sid	rid
1	1
1	2
1	3
1	4
2	1
2	2
3	2
4	2
4	4

R1

rid
2

sid

1
2
3
4

A/R1

R2

rid
2
4

sid

1
4

A/R2

R3

rid
1
2
4

sid

1

A/R3

Is A/B a Fundamental Operation?

No. Shorthand like Joins

joins so common, it's natively supported

Hint: Find all x s not 'disqualified' by some y in B .

x value is *disqualified* if by attaching y value from B , we obtain an xy tuple that is not in A .

A

sid	rid
1	1
1	2
1	3
1	4
2	1
2	2
3	2
4	2
4	4

B

rid
2
4

Disqualified =

A/B =

A

sid	rid
1	1
1	2
1	3
1	4
2	1
2	2
3	2
4	2
4	4

B

rid
2
4

 $\pi_x(A) \times B$

sid	rid
1	2
1	4
2	2
2	4
3	2
3	4
4	2
4	4

Disqualified =
A/B =

$(\pi_x(A) \times B)$

A

sid	rid
1	1
1	2
1	3
1	4
2	1
2	2
3	2
4	2
4	4

B

rid
2
4

 $\pi_x(A) \times B$

sid	rid
1	2
1	4
2	2
2	4
3	2
3	4
4	2
4	4

 $\pi_x(A) \times B) - A$

sid	rid
2	4
3	4

Disqualified = $((\pi_x(A) \times B) - A)$
 A/B =

A

sid	rid
1	1
1	2
1	3
1	4
2	1
2	2
3	2
4	2
4	4

B

rid
2
4

 $\pi_x(A) \times B$

sid	rid
1	2
1	4
2	2
2	4
3	2
3	4
4	2
4	4

 $\pi_x(A) \times B) - A$

sid	rid
2	4
3	4

sid
1
4

A/B

Disqualified = $\pi_x((\pi_x(A) \times B) - A)$

A/B = $\pi_x(A) - \text{Disqualified}$

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(bid, type) Reserve(sid, bid) Student(sid)

Need to join with Instructors to access role

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

More efficient query

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

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

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

Alternatives

define tmp using UNION (how?)

what if we replaced \vee with \wedge in the query?

Students that reserved a DB and HCI book

Does previous approach work?

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

NO

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

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

Is the intersection always allowed?

What if it projected book name?

Students that reserved all books

Use division

Be careful with schemas of inputs to / !

$$\rho(\text{tmp}, (\pi_{\text{sid,bid}} \text{Reserves}) / (\pi_{\text{bid}} \text{Books})) \\ \pi_{\text{name}}(\text{tmp} \bowtie \text{Student})$$

What if want students that reserved all horror books?

$$\rho(\text{tmp}, (\pi_{\text{sid,bid}} \text{Reserves}) / (\pi_{\text{bid}} (\sigma_{\text{type}='horror'} \text{Book})))$$

Summary

Relational Algebra (RA) operators

Operators are closed (inputs & outputs are relations)

Multiple Relational Algebra queries can have same semantics (always return same results)

Forms basis for optimizations

Next Time

Relational Calculus