Relational Algebra

Operators
Expression Trees

What is an “Algebra”

- Mathematical system consisting of:
  - **Operands** --- variables or values from which new values can be constructed.
  - **Operators** --- symbols denoting procedures that construct new values from given values.

What is Relational Algebra?

- An algebra whose operands are relations or variables that represent relations.
- Operators are designed to do the most common things that we need to do with relations in a database.
  - The result is an algebra that can be used as a *query language* for relations.

Roadmap

- There is a core relational algebra that has traditionally been thought of as *the* relational algebra.
- But there are several other operators we shall add to the core in order to model better the language SQL --- the principal language used in relational database systems.

Core Relational Algebra

- Union, intersection, and difference.
  - Usual set operations, but require both operands have the same relation schema.
- Selection: picking certain rows.
- Projection: picking certain columns.
- Products and joins: compositions of relations.
- Renaming of relations and attributes.

Selection

- \( R1 := \text{SELECT}_C(R2) \)
  - \( C \) is a condition (as in "if" statements) that refers to attributes of \( R2 \).
- \( R1 \) is all those tuples of \( R2 \) that satisfy \( C \).
Example

Relation Sells:

<table>
<thead>
<tr>
<th>bar</th>
<th>beer</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe's</td>
<td>Bud</td>
<td>2.50</td>
</tr>
<tr>
<td>Joe's</td>
<td>Miller</td>
<td>2.75</td>
</tr>
<tr>
<td>Sue's</td>
<td>Bud</td>
<td>2.50</td>
</tr>
<tr>
<td>Sue's</td>
<td>Miller</td>
<td>3.00</td>
</tr>
</tbody>
</table>

JoeMenu := SELECT_{bar="Joe’s"}(Sells):

<table>
<thead>
<tr>
<th>bar</th>
<th>beer</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe's</td>
<td>Bud</td>
<td>2.50</td>
</tr>
<tr>
<td>Joe's</td>
<td>Miller</td>
<td>2.75</td>
</tr>
</tbody>
</table>

Projection

R1 := PROJ_{L}(R2)
L is a list of attributes from the schema of R2.
R1 is constructed by looking at each tuple of R2, extracting the attributes on list L, in the order specified, and creating from those components a tuple for R1.
Eliminate duplicate tuples, if any.

Example

Relation Sells:

<table>
<thead>
<tr>
<th>bar</th>
<th>beer</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe's</td>
<td>Bud</td>
<td>2.50</td>
</tr>
<tr>
<td>Joe's</td>
<td>Miller</td>
<td>2.75</td>
</tr>
<tr>
<td>Sue's</td>
<td>Bud</td>
<td>2.50</td>
</tr>
<tr>
<td>Sue's</td>
<td>Miller</td>
<td>3.00</td>
</tr>
</tbody>
</table>

Prices := PROJ_{beer, price}(Sells):

<table>
<thead>
<tr>
<th>beer</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bud</td>
<td>2.50</td>
</tr>
<tr>
<td>Miller</td>
<td>2.75</td>
</tr>
<tr>
<td>Miller</td>
<td>3.00</td>
</tr>
</tbody>
</table>

Example: R3 := R1 * R2

R1( | R2( | R3( A B) | B C) | A B C R1  R2  C |)
| | | 1 2 | 5 | 6 | 1 2 5 | 6 | 1 2 5 | 6 |
| | | 1 2 | 7 | 8 | 1 2 7 | 8 | 1 2 7 | 8 |
| | | 3 4 | 5 | 6 | 3 4 5 | 6 | 3 4 5 | 6 |
| | | 3 4 | 7 | 8 | 3 4 7 | 8 | 3 4 7 | 8 |
| | | 3 4 | 9 | 10| 3 4 9 | 10| 3 4 9 | 10|

Product

R3 := R1 * R2
Pair each tuple t1 of R1 with each tuple t2 of R2.
Concatenation t1 t2 is a tuple of R3.
Schema of R3 is the attributes of R1 and R2, in order.
But beware attribute A of the same name in R1 and R2: use R1.A and R2.A.

Theta-Join

R3 := R1 JOIN_{C} R2
Take the product R1 * R2.
Then apply SELECT_{C} to the result.
As for SELECT, C can be any boolean-valued condition.
Historic versions of this operator allowed only A theta B, where theta was =, <, etc.; hence the name “theta-join.”
Example

Sells(bar, beer, price)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe's</td>
<td>Bud</td>
<td>2.50</td>
</tr>
<tr>
<td>Joe's</td>
<td>Miller</td>
<td>2.75</td>
</tr>
<tr>
<td>Sue's</td>
<td>Bud</td>
<td>2.50</td>
</tr>
<tr>
<td>Sue's</td>
<td>Coors</td>
<td>3.00</td>
</tr>
</tbody>
</table>

Bars(name, addr)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe's</td>
<td>Maple St.</td>
</tr>
<tr>
<td>Sue's</td>
<td>River Rd.</td>
</tr>
</tbody>
</table>

BarInfo := Sells JOIN Bars

BarInfo := Sells JOIN Bars

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe's</td>
<td>Bud</td>
<td>2.50</td>
<td>Joe's</td>
</tr>
<tr>
<td>Joe's</td>
<td>Miller</td>
<td>2.75</td>
<td>Joe's</td>
</tr>
<tr>
<td>Sue's</td>
<td>Bud</td>
<td>2.50</td>
<td>Sue's</td>
</tr>
<tr>
<td>Sue's</td>
<td>Coors</td>
<td>3.00</td>
<td>Sue's</td>
</tr>
</tbody>
</table>

Natural Join

◆ A frequent type of join connects two relations by:
  1. Equating attributes of the same name, and
  2. Projecting out one copy of each pair of equated attributes.
◆ Called *natural join.*
◆ Denoted $R_3 := R_1 \text{JOIN} R_2.$

Renaming

◆ The RENAME operator gives a new schema to a relation.
◆ $R_1 := \text{RENAME}_{A_1, ..., A_n}(R_2)$ makes $R_1$ be a relation with attributes $A_1, ..., A_n$ and the same tuples as $R_2.$
◆ Simplified notation: $R_1(A_1, ..., A_n) := R_2.$

Building Complex Expressions

◆ Algebras allow us to express sequences of operations in a natural way.
  1. Example: in arithmetic --- $(x + 4)(y - 3).$
  2. Relational algebra allows the same.
  3. Three notations, just as in arithmetic:
      1. Sequences of assignment statements.
      2. Expressions with several operators.
      3. Expression trees.
Sequences of Assignments

- Create temporary relation names.
- Renaming can be implied by giving relations a list of attributes.
- Example: \( R_3 := R_1 \JOIN_C R_2 \) can be written:
  - \( R_4 := R_1 \times R_2 \)
  - \( R_3 := \SELECT_C (R_4) \)

Expressions in a Single Assignment

- Example: the theta-join \( R_3 := R_1 \JOIN_C R_2 \) can be written: \( R_3 := \SELECT_C (R_1 \times R_2) \)
- Precedence of relational operators:
  1. Unary operators --- select, project, rename --- have highest precedence, bind first.
  2. Then some products and joins.
  3. Then intersection.
  4. Finally, union and set difference bind last.
- But you can always insert parentheses to force the order you desire.

Expression Trees

- Leaves are operands --- either variables standing for relations or particular, constant relations.
- Interior nodes are operators, applied to their child or children.

Example

- Using the relations \( \text{Bars}(\text{name}, \text{addr}) \) and \( \text{Sells}(\text{bar}, \text{beer}, \text{price}) \), find the names of all the bars that are either on Maple St. or sell Bud for less than $3.

As a Tree:

```
  UNION
 /     \                   
PROJECT_name RENAME_{R(name)}
 |         |                   
/ \       / \                 
SELECT \_addr = “Maple St.” PROJECT_bar
 |         |                   
/ \       / \                 
\_SELECT \_price<3 AND beer=“Bud” \_Sells
 |         |                   
\_Bars
```

Example

- Using \( \text{Sells}(\text{bar}, \text{beer}, \text{price}) \), find the bars that sell two different beers at the same price.
- Strategy: by renaming, define a copy of \( \text{Sells} \), called \( \text{S}(\text{bar}, \text{beer}1, \text{price}) \). The natural join of \( \text{Sells} \) and \( \text{S} \) consists of quadruples \( (\text{bar}, \text{beer}, \text{beer}1, \text{price}) \) such that the bar sells both beers at this price.
The Tree

```
PROJECT_{bar}
|
SELECT_{beer != beer1}
|
JOIN
|
RENAME_{(bar, beer1, price)}
```

Schemas for Interior Nodes

- An expression tree defines a schema for the relation associated with each interior node.
- Similarly, a sequence of assignments defines a schema for each relation on the left of the := sign.

Schema-Defining Rules 1

- For union, intersection, and difference, the schemas of the two operands must be the same, so use that schema for the result.
- Selection: schema of the result is the same as the schema of the operand.
- Projection: list of attributes tells us the schema.

Schema-Defining Rules 2

- Product: the schema is the attributes of both relations.
  - Use R.A, etc., to distinguish two attributes named A.
- Theta-join: same as product.
- Natural join: use attributes of both relations.
  - Shared attribute names are merged.
- Renaming: the operator tells the schema.

Relational Algebra on Bags

- A bag is like a set, but an element may appear more than once.
  - Multiset is another name for “bag.”
- Example: \{1,2,1,3\} is a bag. \{1,2,3\} is also a bag that happens to be a set.
- Bags also resemble lists, but order in a bag is unimportant.
  - Example: \{1,2,1\} = \{1,1,2\} as bags, but \[1,2,1\] != \[1,1,2\] as lists.

Why Bags?

- SQL, the most important query language for relational databases is actually a bag language.
  - SQL will eliminate duplicates, but usually only if you ask it to do so explicitly.
- Some operations, like projection, are much more efficient on bags than sets.
Operations on Bags

- Selection applies to each tuple, so its effect on bags is like its effect on sets.
- Projection also applies to each tuple, but as a bag operator, we do not eliminate duplicates.
- Products and joins are done on each pair of tuples, so duplicates in bags have no effect on how we operate.

Example: Bag Selection

\[
\begin{array}{ccc}
R( & A & B \\
1 & 2 \\
5 & 6 \\
1 & 2 \\
\end{array} \\
S( & B & C \\
3 & 4 \\
7 & 8 \\
1 & 2 \\
\end{array}
\]

\[
\text{SELECT}_{A \neq B} (R) = \begin{array}{ccc}
A & B \\
1 \\
5 \\
1 \\
\end{array}
\]

Example: Bag Projection

\[
\begin{array}{ccc}
R( & A & B \\
1 & 2 \\
5 & 6 \\
1 & 2 \\
\end{array} \\
S( & B & C \\
3 & 4 \\
7 & 8 \\
1 & 2 \\
\end{array}
\]

\[
\text{PROJECT}_{A} (R) = \begin{array}{c}
A \\
1 \\
5 \\
1 \\
\end{array}
\]

Example: Bag Product

\[
\begin{array}{ccc}
R( & A & B \\
1 & 2 \\
5 & 6 \\
1 & 2 \\
\end{array} \\
S( & B & C \\
3 & 4 \\
7 & 8 \\
1 & 2 \\
\end{array}
\]

\[
R \times S = \begin{array}{cccc}
A & R.B & S.B & C \\
1 & 2 & 3 & 4 \\
1 & 2 & 7 & 8 \\
5 & 6 & 3 & 4 \\
5 & 6 & 7 & 8 \\
1 & 2 & 3 & 4 \\
1 & 2 & 7 & 8 \\
\end{array}
\]

Example: Bag Theta-Join

\[
\begin{array}{ccc}
R( & A & B \\
1 & 2 \\
5 & 6 \\
1 & 2 \\
\end{array} \\
S( & B & C \\
3 & 4 \\
7 & 8 \\
\end{array}
\]

\[
\text{R JOIN}_{A = B} S = \begin{array}{cccc}
A & R.B & S.B & C \\
1 & 2 & 7 & 8 \\
5 & 6 & 7 & 8 \\
1 & 2 & 3 & 4 \\
1 & 2 & 7 & 8 \\
\end{array}
\]

Bag Union

- Union, intersection, and difference need new definitions for bags.
- An element appears in the union of two bags the sum of the number of times it appears in each bag.
- Example: \{1,2,1\} UNION \{1,1,2,3,1\} = \{1,1,1,1,1,2,2,3\}
Bag Intersection

◆ An element appears in the intersection of two bags the minimum of the number of times it appears in either.
◆ Example: \{1,2,1\} \text{ INTER } \{1,2,3\} = \{1,2\}.

Bag Difference

◆ An element appears in the difference \( A – B \) of bags as many times as it appears in \( A \), minus the number of times it appears in \( B \).
◆ But never less than 0 times.
◆ Example: \{1,2,1\} – \{1,2,3\} = \{1\}.

Beware: Bag Laws != Set Laws

◆ Not all algebraic laws that hold for sets also hold for bags.
◆ For one example, the commutative law for union \((R \text{ UNION } S = S \text{ UNION } R)\) does hold for bags.
  ◆ Since addition is commutative, adding the number of times \( x \) appears in \( R \) and \( S \) doesn’t depend on the order of \( R \) and \( S \).

An Example of Inequivalence

◆ Set union is\textit{idempotent}, meaning that \( S \text{ UNION } S = S \).
◆ However, for bags, if \( x \) appears \( n \) times in \( S \), then it appears \( 2n \) times in \( S \text{ UNION } S \).
◆ Thus \( S \text{ UNION } S \neq S \) in general.

The Extended Algebra

1. \text{ DELTA } = \text{ eliminate duplicates from bags.}
2. \text{ TAU } = \text{ sort tuples.}
3. \textit{Extended projection} : arithmetic, duplication of columns.
4. \text{ GAMMA } = \text{ grouping and aggregation.}
5. \text{ OUTERJOIN: avoids “dangling tuples” = tuples that do not join with anything.}

Duplicate Elimination

◆ \text{ R1 := DELTA(R2).}
◆ \text{ R1 consists of one copy of each tuple that appears in R2 one or more times.}
Example: Duplicate Elimination

\[
R = \begin{pmatrix}
1 & 2 \\
3 & 4 \\
1 & 2
\end{pmatrix}
\]

\[
\text{DELTA}(R) = \begin{pmatrix}
1 & 2 \\
3 & 4
\end{pmatrix}
\]

Example: Sorting

\[
R = \begin{pmatrix}
1 & 2 \\
3 & 4 \\
5 & 2
\end{pmatrix}
\]

\[
\text{TAU}_B(R) = [(5,2), (1,2), (3,4)]
\]

Example: Extended Projection

\[
R = \begin{pmatrix}
1 & 2 \\
3 & 4
\end{pmatrix}
\]

\[
\text{PROJ}_{A+B,A1,A2}(R) = \begin{pmatrix}
3 & 1 & 1 \\
7 & 3 & 3
\end{pmatrix}
\]

Sorting

\[\text{TAU}_L(R2).\]  
\[L\] is a list of some of the attributes of R2.

\[\text{R1 is the list of tuples of R2 sorted first on the value of the first attribute on } L, \text{ then on the second attribute of } L, \text{ and so on.}\]


\[\text{Break ties arbitrarily.}\]

\[\text{TAU is the only operator whose result is neither a set nor a bag.}\]

Extended Projection

\[\text{Using the same PROJ}_{L} \text{ operator, we allow the list } L \text{ to contain arbitrary expressions involving attributes, for example:}\]

1. Arithmetic on attributes, e.g., \(A + B\).
2. Duplicate occurrences of the same attribute.

Aggregation Operators

\[\text{Aggregation operators are not operators of relational algebra.}\]

\[\text{Rather, they apply to entire columns of a table and produce a single result.}\]

\[\text{The most important examples: SUM, AVG, COUNT, MIN, and MAX.}\]
Example: Aggregation

\[
R = \begin{array}{cc}
A & B \\
1 & 3 \\
3 & 4 \\
3 & 2 \\
\end{array}
\]

- \text{SUM}(A) = 7
- \text{COUNT}(A) = 3
- \text{MAX}(B) = 4
- \text{AVG}(B) = 3

Grouping Operator

\[ R_1 := \text{GAMMA}_L(R_2). \]
- \( L \) is a list of elements that are either:
  1. Individual (grouping) attributes.
  2. \( \text{AGG}(A) \), where \( \text{AGG} \) is one of the aggregation operators and \( A \) is an attribute.

Applying \( \text{GAMMA}_L(R) \)

- Group \( R \) according to all the grouping attributes on list \( L \).
- That is, form one group for each distinct list of values for those attributes in \( R \).
- Within each group, compute \( \text{AGG}(A) \) for each aggregation on list \( L \).
- Result has grouping attributes and aggregations as attributes. One tuple for each list of values for the grouping attributes and their group’s aggregations.

Example: Grouping/Aggregation

\[
R = \begin{array}{ccc}
A & B & C \\
1 & 2 & 3 \\
4 & 5 & 6 \\
1 & 2 & 5 \\
\end{array}
\]

\[
\text{GAMMA}_{A,B\text{AVG}(C)}(R) = ??
\]

First, group \( R \):

\[
\begin{array}{ccc}
A & B & C \\
1 & 2 & 3 \\
1 & 2 & 5 \\
1 & 5 & 6 \\
\end{array}
\]

Then, average \( C \) within groups:

\[
\begin{array}{ccc}
A & B & \text{AVG}(C) \\
1 & 2 & 4 \\
4 & 5 & 6 \\
\end{array}
\]

Outerjoin

- Suppose we join \( R \ JOIN C S \).
- A tuple of \( R \) that has no tuple of \( S \) with which it joins is said to be dangling.
- Similarly for a tuple of \( S \).
- Outerjoin preserves dangling tuples by padding them with a special NULL symbol in the result.

Example: Outerjoin

\[
R = \begin{array}{cc}
A & B \\
1 & 2 \\
4 & 5 \\
\end{array}
\]

\[
S = \begin{array}{cc}
B & C \\
2 & 3 \\
6 & 7 \\
\end{array}
\]

(1,2) joins with (2,3), but the other two tuples are dangling.

\[
R \text{ OUTERJOIN } S = \begin{array}{ccc}
A & B & C \\
1 & 2 & 3 \\
4 & 5 & \text{NULL} \\
\text{NULL} & 6 & 7 \\
\end{array}
\]