Relational Algebra on Bags

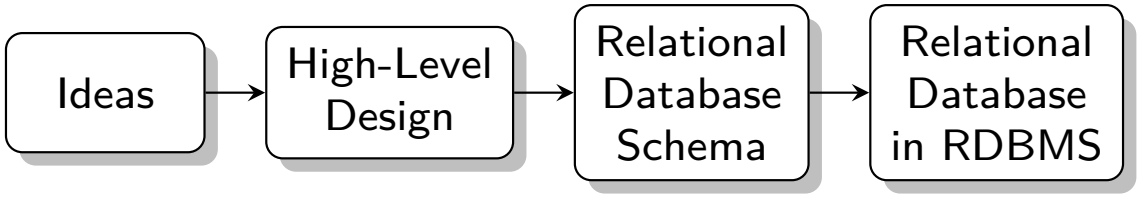
Hui Chen [a]

[a]CUNY Brooklyn College, Brooklyn, NY, USA

March 13, 2025

Outline

Overview


Ideas
High-Level Design
Relational Database Schema
Relational Database in RDBMS

Outline

Introduction to SQL

SQL (pronounced as "sequel") is the principal language used to describe and manipulate relational database, and has several aspects:

- Data definition language (DDL).
  - SQL includes commands to create database objects such as tables, indexes, and views, as well as commands to define access rights to those database objects.
  - Topics of this lecture: SQL commands to create database tables (relations)
- Data manipulation language (DML).
  - SQL includes commands to insert, update, delete, and retrieve data within the database tables.
- Transaction control language (TCL).
  - The DML commands in SQL are executed within the context of a transaction.
- Data control language (DCL).
  - Data control commands are used to control access to data objects.

To Discuss Subset of DML

- Relational Algebra
- Query Databases

Outline

Why Relational Algebra

- SQL, incorporates relational algebra at its center, and many SQL programs are really "syntactically sugared" relational algebra expressions
- When a DBMS processes queries, it first the first thing translate queries into relational algebra or a very similar internal representation.
- Help us understand queries.

What's Algebra?

A 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.

Relational Algebra

An algebra whose operands are relations or variables that represent relations.

- designed to do the most common tasks with relations in a relational database
- can be used as a query language for relations

Outline

Relational Algebra on Bags

Examined relational algebra on sets, now examine relational algebra on bags

- A bag (or multiset) is like a set, but an element may appear more than once.
- Examples, $\{1, 2, 1, 3\}$, $\{'a', 'a', 'a', 'b'\}$, $\{1, 2, 3\}$, and $'a', 'b'$

Why Bags

- SQL is actually a bag language.
- Some operations, like projection, are more efficient on bags than sets.

Bag Union

An element appears in the union of two bags the sum of the number of times it appears in each bag.

- Example: $\{1, 2, 1\} \cup \{1, 1, 2, 3, 1\} = \{1, 1, 1, 1, 1, 2, 2, 3\}$

Bag Intersection

An element appears in the intersection of two bags the minimum of the number of times it appears in either.

- Example: $\{1, 2, 1, 1\} \cap \{1, 2, 1, 3\} = \{1, 1, 2\}$

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; however, never less than 0 times.

- Example: $\{1, 2, 1, 1\} - \{1, 2, 3\} = \{1, 1\}$
- Example: $\{1, 2, 1, 1\} - \{1, 1, 1, 1, 3\} = \{2\}$

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.

Extended Operators

- Duplicate-elimination operator δ
- Aggregation operators, e.g., sum, average, min, max
- Grouping operator γ combines grouping and aggregation (see the aggregation operators above)
- Extended projection π – extending π with computation
- Sorting operator τ
- Outer-join operator $⟗$, $⟕$, and $⟖$

Duplicate Elimination

That is to convert a bag to a set. Use operator σ

$$\sigma(R)$$

Aggregation

These operators are used to "summarize" or "aggregate" the values in one column of a relation, e.g.,

- Apply to a column with numerical values
  - `SUM`. Produces the sum
  - `AVG`. Compute the average
  - `MIN` and `MAX`. Obtain the smallest or the largest value
- Apply to a column of any data type
  - `COUNT`. Produces the number of values in a column

Grouping

Divide tuples in a relation into groups.

$$\gamma_L(R)$$

where L is a list of elements, each of which is either

1. an attribute of the relation R to group the tuples;
2. an aggregation operator applied to an attribute of the relation.

Example:

$$\gamma_{\text{starName,MIN}(year)\rightarrow\text{minYear,COUNT}(title)\rightarrow\text{ctTitle}}(StarsIn)$$

Outer Joins

- Previously discussed the join are in fact "inner join". We have also outer join. This is motivated by the existence of the "dangling" value.
  - Semantics of"Dangling Value"? Don't have it vs. don't know it.
  - For outer joins, we consider: comparing to something you don't know is meaningless.
- Considering this, we have three types of outer joins:
  - Outer Join (a.k.a., full outer join). $R_1 \rtimes\!\ltimes R_2$: Include both dangling values from left and right relations (both R_1 and R_2).
  - Left Join (a.k.a., left outer join). $R_1 \rtimes R_2$: Include dangling values from left relation (R_1) only.
  - Right Join (a.k.a., right outer join) $R_1 \ltimes R_2$: Include dangling values from right relation (R_2) only.

Outline

Assignment

Let's work on an assignment using paper and pencil/pen ...