HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

WEEK 5

BYOD

AGENDA

- Relational Model
  - based on “Database Systems - The Complete Book” (H. Garcia-Molina, J. D. Ullman, J. Widom)
- Sprint 3
  - Operators in Opossum
- Organization

MOTIVATION FOR THE RELATIONAL MODEL

- Previously, databases tightly coupled logical and physical layers which impeded maintainability
- No conceptual idea of which operators are required
- Ted Codd proposed the **relational model** in the 1970s
  - Abstraction model using simple data structures and high-level operators
  - Implementation and physical storage is up to vendor

RELATIONAL DATABASES

- Database - organized collection of data
- Database Management System (DBMS) - the program that manages the database
- Relational database is based on relational **data model**

1. Structure of the data
   - Physical model
   - Conceptual model
2. Operations on the data
   - Modifications - change the database
   - Queries - retrieve information
3. Constraints on the data

RELATIONAL MODEL – CONCEPTUAL DATA MODEL

- Data - two-dimensional table, called relation
  - Set or bag (multiset)
- Attribute - name of a column
- Schema - name of relation and set of attributes
- Tuple - row (except header) of a relation
- Further concepts:
  equality, relation instance, domain/data type, NULL

RELATIONAL MODEL – OPERATIONS

- Relational algebra is the basis for how the relational model is implemented in practice
  - Theoretical foundation for relational databases and SQL
- Operations
  - Take one or more relations as input(s) and output new relation
  - Can be chained to form more complex **queries**
- Classes of traditional operations:
  - Operations that remove parts of a relation: selection, and projection
  - Operations that combine tuples of two relations: cartesian product, and join
  - Renaming: relations and attributes
  - Set operations: union, intersection, and difference

RELATIONAL MODEL – OPERATIONS THAT REMOVE PARTS OF A RELATION

- Projection of R produces a new relation with a subset of R's columns
  - In the relational algebra of sets, duplicate tuples are eliminated
- Selections of R produces a new relation with a subset of R's tuples (those that satisfy a condition C)

RELATIONAL MODEL – OPERATIONS THAT COMBINE TUPLES OF TWO RELATIONS

- Cartesian product ((cross-)product) of R and S is the set of pairs formed by choosing the first element to be any element of R and the second any element of S
  - The schema of the new relation is the union of schemas for R and S (Exception: R and S have attribute A in common -> use new name R.A and S.A)
- Join of R and S pairs tuples that match in some way
  - **Dangling tuple**: tuple with no match
  - Natural join: match in common attributes of R and S
  - Theta join: match based on arbitrary condition C
    - Product of R and S, filtered by condition C
    - Schema of new relation: see cartesian product
  - Semi join of R and S is the set of tuples in R that match the join condition

RELATIONAL MODEL – SET OPERATIONS

- Union of R and S is the set of elements that are in R or S or both
- Intersection of R and S is the set of elements that are in both R and S
- Difference of R and S is the set of that are in R but not in S
  - R - S is different from S - R
- Conditions for R and S:
  - R and S must have schemas with identical attributes and domains

RELATIONAL MODEL – MINIMAL RELATIONAL ALGEBRA?

- Union, intersection, difference, projection, selection, cartesian product, natural join, theta join, semi join, renaming

RELATIONAL MODEL – MINIMAL RELATIONAL ALGEBRA

- Union, ~~intersection~~, difference, projection, selection, cartesian product, ~~natural join, theta join, semi join~~, renaming

RELATIONAL MODEL – WHAT IS MISSING

- Bag semantic (+ duplicate elimination)
- Aggregation (and grouping)
- Sort
- Extended projection
- Outer join

RELATIONAL MODEL – BAG SEMANTIC

- Bags are multi sets (allow duplicates)
  - Redefinition of set operations necessary
- Some relational operations are more efficient with the bag model (without duplicate elimination)
  - Union
  - Projection
- Duplicate-elimination operator turns bag into set by eliminating all but one copy of each tuple

RELATIONAL MODEL – AGGREGATION

- Aggregations summarize or “aggregate” the values in one column
  - Examples: SUM, AVG, MIN, MAX, COUNT
- Groupings allow aggregations of tuple groups that correspond to the value of one or multiple columns

RELATIONAL MODEL – SORT

- Turns unordered container, e.g., set, bag, into an ordered one, e.g., list
  - Only useful as last operator of a **relational query** (and its **logical query plan**), because following operators turn list into set or bag
  - Of importance for **physical query plans** (an operator implementation may require sorted inputs)

RELATIONAL MODEL – EXTENDED PROJECTION

- Besides renamings, extended projections allow arbitrary expressions
  - Constants
  - Arithmetic operators
  - String operators

RELATIONAL MODEL – OUTER JOIN

- Outer join is the union of the natural join and all dangling tuples from R and S; dangling tuples of R and S must be padded with NULLs for missing attributes
  - Full, left, and right outer join
  - Theta join versions of outer join operate analogous
  - Inner join is a synonym of “normal” join

SQL – THE DATABASE LANGUAGE

- Structured Query Language
  - Express queries of relational algebra (declaratively)
  - Statements for modifying the database
  - Declaring the database schema
  - Further concepts: constraints, views, indexes, …

OPOSSUM'S OPERATOR CONCEPT

- Opossum implements the relational algebra with operators
  - Queries are formulated as graphs by chaining operators
  - Usually, the first operator is the GetTable operator
    - Each operator takes up to two operators as input
    - After its execution, an operator's result table is set
- The hard part
  - We have to deal with multiple table representations
  - While doing that, we have to keep an eye on efficiency

ORGANISATION

- Deadline Sprint 3
  - 3 December 2017
- Instructions for Code Review of Sprint 2 will follow
- Next Week
  - Sprint 2 Feedback
  - NULL Values
  - Virtual Method Call Overhead

THAT'S IT.