CS430/630
Database Management Systems
Fall 2017

Midterm Review

Grading

- Final exam (150 points) – open book
- Midterm (100 points) – open book
  - Open book does NOT include electronic devices!

Approximately 6 homework assignments
- 10 points each, drop lowest score.
- Assignments for CS630 will have additional questions
- Assignments are individual – submit your own work only!
- No plagiarism! See student code of conduct

Course Overview

- Relational Data Model
- Relational Algebra
- Structured Query Language
  - The most important part of the course
- Conceptual design – the ER model
- Views
- After midterm Exam:
  - Database application development
    - Java/JDBC, PL/SQL
  - Design Theory
  - Database Security

The Relational Model

Relational Schema

- **Schema**: specifies name of relation, plus name and domain of each column
  - Students (sid: integer, name: string, login: string, age: integer, gpa: real )

- Each relation must have a schema
  - Similar to a data type in programming languages

- Relational database schema = collection of relations’ schemas

More about Relations

- **Relations are sets of tuples**
  - Sets are NOT ordered
  - Do NOT retrieve by row number, but by content!

- **Relation Instance**
  - Contents of a relation may change over time
    - Tuples are added/deleted/modified
    - E.g., Students join or leave the university
    - **Instance** represents set of tuples at a certain point in time

- Schemas may change too
  - Although this is not frequent in practice
  - Changing schema is very expensive
Keys

- A key of a relation is a set of fields \( K \) such that:
  1. No two distinct tuples in ANY relation instance have the same values in all key fields, and
  2. No subset of \( K \) is a key (otherwise \( K \) is a superkey)

- Key may not be unique
- Multiple candidate keys may exist
- One of the keys is chosen (by DBA) to be the primary key
- Keys are shown underlined in schema
- In the relational model, duplicate tuples do not exist!
- But most DBMS implementations do allow duplicates
- Key constraints must be set by DBA to avoid duplicates
- Rule of thumb: every table should have a primary key

Example of Keys

Students:
- sid is a key;
- \{sid, name\} is a superkey

In practice, it is not easy to know when there exists a unique attribute combination in the data (e.g., names)
- artificial keys are created: student ID, customer ID, etc.
- SSN is also often used for keys (but what if multinational…)

Relational Algebra (examples in Lecture 2)

- Basic operations:
  - Selection \( \sigma \) Selects a subset of rows from relation
  - Projection \( \pi \) Deletes unwanted columns from relation
  - Cross-product \( \times \) Allows us to combine several relations
  - Join \( \bowtie \) Combines several relations using conditions
  - Division \( \div \) A bit more complex, will cover later on
  - Set-difference \( \setminus \) Intersection \( \cap \)
  - Renaming \( \rho \) Helper operator, does not derive new result, just renames relations and fields

Sample Query 0

Find names of sailors who have ratings at least 8.

\( \pi_{\text{name}}(\sigma_{\text{rating} \geq 8}(\text{Sailors})) \)
Sample Query 1

\[
\pi_{sname}(\sigma_{\text{bid}=103}(\text{Reserves} \bowtie \text{Sailors}))
\]

Sample Query 2

\[
\pi_{sname}(\sigma_{\text{color}='\text{red'}}(\text{Boats} \bowtie \text{Reserves} \bowtie \text{Sailors}))
\]

Sample Query 3

\[
\rho(\text{Tempboats}, \pi_{\text{color}='\text{red'}} \lor \text{color}='\text{green'}(\text{Boats}))
\]

Sample Query 4

\[
\rho(\text{Tempred}, \pi_{\text{color}='\text{red'}}(\text{Boats} \bowtie \text{Reserves}))
\]

Sample Query 5

\[
\pi_{sname}(\text{Tempred} \cap \text{Tempothers} \bowtie \text{Sailors})
\]

Structured Query Language
Relational Query Language: SQL

- Supports simple, yet powerful querying of data.
- Precise semantics for relational queries.
- SQL (Data Manipulation Language)
- DDL (Data Definition Language)
- SQL developed by IBM (system R) in the 1970s

Standards:
- SQL-86
- SQL-89 (minor revision)
- SQL-92 (major revision): third and most important standard
  - Early enough to affect Oracle, DB2, other important commercial databases, so the real common ground.
- SQL-99 (major extensions, triggers, recursive queries)

SQL’s Most Important Data Types

- Character strings
  - CHAR(n), VARCHAR(n): fixed and variable-length strings
- Numerical:
  - INTEGER
    - Usually 32-bit, but not guaranteed by SQL-92.
  - REAL
    - Oracle: use BINARY_DOUBLE for IEEE format.
- Fixed precision:
  - DECIMAL(n,d)
    - 1234.56 is of type DECIMAL(6,2), precision 6, scale 2

Creating Relations in SQL

```
CREATE TABLE Students
  (sid CHAR(20),
   name CHAR(20),
   login CHAR(10),
   age INTEGER,
   gpa REAL);
```

```
CREATE TABLE Enrolled
  (sid CHAR(20),
   cid CHAR(20),
   grade CHAR(2));
```

Destroying and Altering Relations

```
DROP TABLE Students;
```

```
ALTER TABLE Students
  ADD firstYear INTEGER;
```

Structure of SQL SELECT Query in Lecture 4

```
SELECT [DISTINCT] target-list
FROM relation-list
WHERE qualification
```

- **relation-list** = list of relation names
  - possibly with a range-variable (AKA correlation name or table alias) after each name
- **target-list** = list of attributes of relations in relation-list
- **qualification** = conditions Attr op const or Attr1 op Attr2
  - op is one of <, >, =, >=, <=, <>, or string operators
  - Expressions connected using AND, OR and NOT
- **DISTINCT** = optional, eliminates duplicates
  - By default duplicates are NOT eliminated!

Conceptual Evaluation Strategy

- **Semantics of SQL query**
  1. Compute the cross-product of relation-list
  2. Discard resulting tuples if they fail qualifications
  3. Delete attributes that are not in target-list
  4. If DISTINCT is specified, eliminate duplicate rows

- This strategy is least efficient way to compute a query!
  - Optimizer finds efficient strategies to compute the same result
Expressions and Strings

- "Find rating and number of years to retirement for sailors whose names begin with 'd', end with 'n' and contain at least three characters"

```sql
SELECT S.rating, 60 - S.age AS Yr_to_retire
FROM Sailors S
WHERE S.sname LIKE 'd_%n'
```

- AS allows to (re)name fields in result.
- LIKE is used for string matching
  - _ stands for any one character
  - % stands for 0 or more arbitrary characters
- Mysql also has RLIKE, with regular expression matching

---

Expressions and Strings - Example

<table>
<thead>
<tr>
<th>sid</th>
<th>sname</th>
<th>rating</th>
<th>age</th>
<th>rating</th>
<th>Yr_to_retire</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>dustin</td>
<td>7</td>
<td>45.0</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>31</td>
<td>lubber</td>
<td>8</td>
<td>55.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>rusty</td>
<td>10</td>
<td>35.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Set Operations

- UNION
  - compute the union of any two union-compatible sets of tuples
- INTERSECT (not supported in mysql)
  - compute the intersection of any two union-compatible sets of tuples
- EXCEPT or MINUS (not supported in mysql)
  - Set difference of any two union-compatible sets of tuples
- Duplicates eliminated by default!
  - UNION ALL, INTERSECT ALL, EXCEPT ALL retain duplicates
- Contrast with non-set SQL operations

---

Connecting queries and subqueries (Lecture 6)

- Depends on what the subquery returns:
  - A scalar value (1x1 table) – such a subquery can appear in a query in the same place where a constant appears, in SQL-99-compliant DBs.
  - A relation
- Where can subqueries appear?
  - Most often in WHERE clause of parent query
  - Also used in FROM clause followed by range variable
  - … FROM Sailors, (SELECT bid FROM Boats) Bids …
  - In HAVING clauses
  - Will discuss later on
  - Also, if scalar, in the select-list. Not covered in text, and not in Entry-level SQL-92.

---

Integrity Constraints (ICs)

Basics in Lectures 4-5

- IC: condition that must hold for any instance of the database; e.g., domain constraints
  - Specified when schema is defined.
  - Checked when relations are modified.
- A legal instance satisfies all specified ICs
  - It is the DBMS's role to enforce IC
- ICs we study
  - Primary key constraints
  - Foreign key constraints: referential integrity
  - Unique constraints
  - NOT NULL constraints

---

Conditions involving relations

- Test that a relation satisfies some condition
  - … WHERE EXISTS (SELECT …) - TRUE if subquery result is not empty
  - … WHERE UNIQUE (SELECT …) - TRUE if subquery result has no duplicates?
```sql
SELECT S.sname
FROM Sailors S
WHERE EXISTS (SELECT * FROM Reserves R
WHERE R.bid=103 AND R.sid=S.sid)
```  
- Find names of sailors who've reserved boat #103

Subquery is CORRELATED with parent query  
  *Note UNIQUE (select…) is not in Entry SQL-92
Conditions involving relations and tuples

- Typically have some sort of set operations semantics
  - WHERE field IN (SELECT ...)
  - WHERE field op ANY (SELECT ...)
  - WHERE field op ALL (SELECT ...)

- Find names of sailors with maximum age

Subqueries in the FROM clause

- Find names of sailors who reserved 'interlake'

Rewriting INTERSECT Queries Using IN

- Find sid's of sailors who've reserved both a red and a green boat:

  SELECT S.sid
  FROM Sailors S, Boats B, Reserves R
  WHERE S.sid=R.sid AND B.bid=R.bid AND B.color='red'
  AND S.sid IN (SELECT S2.sid
                   FROM Sailors S2, Boats B2, Reserves R2
                   WHERE S2.sid=R2.sid AND B2.bid=R2.bid
                   AND B2.color='green')

- Similarly, EXCEPT queries re-written using NOT IN.

Nested Queries - Review

- Nested queries returning a constant
  - Typically constant is compared with other value in the WHERE clause
  - ...WHERE field = (SELECT bid FROM ...)
  - ...WHERE field IN (SELECT bid FROM ...)

- Nested queries returning a relation
  - in WHERE clause
  - ...WHERE EXISTS|UNIQUE (SELECT bid FROM ...)
  - ...WHERE field op ANY|ALL (SELECT bid FROM ...)
  - in FROM clause followed by range variable
  - ...FROM Sailors, (SELECT bid FROM Boats) Bids ...

Aggregate Operators (Lecture 7)

Significant extension of relational algebra

<table>
<thead>
<tr>
<th>COUNT(*)</th>
<th>COUNT ([DISTINCT] A)</th>
<th>SUM ([DISTINCT] A)</th>
<th>AVG ([DISTINCT] A)</th>
<th>MAX(A)</th>
<th>MIN(A)</th>
</tr>
</thead>
</table>

A is a single column

Result is single value obtained by applying aggregate over all qualifying tuples

- SELECT COUNT(*)
  FROM Sailors S

Queries With GROUP BY and HAVING

- SELECT [DISTINCT] target-list
  FROM relation-list
  WHERE qualification
  GROUP BY grouping-list
  HAVING group-qualification

- The target-list contains:
  1. (i) attribute names list (column names, possibly with corr. names)
  2. (ii) terms with aggregate operations (e.g., MIN(S.age))

- The attribute list (i) must be a subset of grouping-list
  - A group is a set of tuples that have the same value for all attributes in grouping-list
  - Each answer tuple corresponds to a group, so these attributes must have a single value per group.
GROUPBY Query Example

"Find age of the youngest sailor with age at least 18, for each rating with at least 2 such sailors"

```sql
SELECT S.rating, MIN(S.age) AS minage
FROM Sailors S
WHERE S.age >= 18
GROUP BY S.rating
HAVING COUNT(*) > 1
```

<table>
<thead>
<tr>
<th>sid</th>
<th>sname</th>
<th>rating</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>dustin</td>
<td>7</td>
<td>45.0</td>
</tr>
<tr>
<td>29</td>
<td>brutus</td>
<td>1</td>
<td>33.0</td>
</tr>
<tr>
<td>31</td>
<td>lubber</td>
<td>8</td>
<td>55.5</td>
</tr>
<tr>
<td>32</td>
<td>andy</td>
<td>8</td>
<td>25.5</td>
</tr>
<tr>
<td>58</td>
<td>rusty</td>
<td>10</td>
<td>35.0</td>
</tr>
<tr>
<td>64</td>
<td>horatio</td>
<td>7</td>
<td>35.0</td>
</tr>
<tr>
<td>71</td>
<td>zorba</td>
<td>10</td>
<td>16.0</td>
</tr>
<tr>
<td>74</td>
<td>horatio</td>
<td>9</td>
<td>35.0</td>
</tr>
<tr>
<td>85</td>
<td>art</td>
<td>3</td>
<td>25.5</td>
</tr>
<tr>
<td>95</td>
<td>bob</td>
<td>3</td>
<td>63.5</td>
</tr>
<tr>
<td>96</td>
<td>frodo</td>
<td>3</td>
<td>25.5</td>
</tr>
</tbody>
</table>

Null Values (Lecture 8)

- Field values in a tuple may sometimes be unknown: e.g., a rating has not been assigned, or a new column is added to the table.
- inapplicable: e.g., CEO has no manager, single person has no spouse.
- SQL provides a special value NULL for such situations.
- Special operators IS NULL, IS NOT NULL.
- Note: NULL must not be used as constant in expressions!
- A field can be declared as NOT NULL, means NULL values are not allowed (by default, PK fields are NOT NULL).

Dealing with Null Values

- The presence of NULL complicates some issues.
  - NULL op value has as result NULL (op is +,*,/)
  - What does rating>8 evaluate to if rating is equal to NULL?
    - Answer: unknown
  - 3-valued logic: true, false and unknown
    - Recall that WHERE eliminates rows that don't evaluate to true
    - What about AND, OR and NOT connectives?
      - unknown AND true = unknown
      - unknown OR false = unknown
      - NOT unknown = unknown
    - Also, <NULL_value> = <NULL_value> is unknown!

Division in SQL: Solution 1 (Q9, pg. 150)

"Find sailors who've reserved all boats."

```sql
SELECT sname
FROM Sailors S
WHERE NOT EXISTS
  (SELECT bid FROM Boats
   WHERE sid = S.sid)
```

RA Division Example (Lecture 9)

"Find sailors who've reserved all red boats"

```sql
ρ(Temp.ππ_{sid,bid}(Reserves)\left/\pi\sigma_{bid}(\sigma_{color='red'}Boats))\right.π_{sname}(Temp->a\Sailors)
```

Join Expressions

- SQL shorthands for expressions we already saw.
- Cross Product:
  - Sailors CROSS JOIN Reserves
- Condition Join:
  - Sailors JOIN Reserves ON <condition>
- Natural Join: uses same-named columns for join columns
  - Sailors NATURAL JOIN Reserves
- Usage Example:
  - SELECT *
    FROM Sailors JOIN Reserves ON Sailors.sid = Reserves.sid
Outer Joins: Lecture 10

The explicit syntax for an outer join

```
SELECT select_list
FROM table_1
{LEFT|RIGHT|FULL} [OUTER] JOIN table_2
ON join_condition_1
{LEFT|RIGHT|FULL} [OUTER] JOIN table_3
ON join_condition_2)...```

What outer joins do

<table>
<thead>
<tr>
<th>Join</th>
<th>Keeps unmatched rows from</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>The left table</td>
</tr>
<tr>
<td>Right</td>
<td>The right table</td>
</tr>
<tr>
<td>Full</td>
<td>Both tables</td>
</tr>
</tbody>
</table>

A left outer join

```
SELECT department_name AS dept_name,
       d.department_number AS dept_no,
       last_name
FROM departments d
LEFT JOIN employees e
ON d.department_number = e.department_number
ORDER BY department_name```

- This join preserves departments even if they have no employees
- It would still be useful even if we had a non-null FK in place.

A SELECT statement with an outer and inner join

```
SELECT department_name AS dept_name,
       last_name, project_number
FROM departments dpt
JOIN employees emp
ON dpt.department_number = emp.department_number
LEFT JOIN projects prj
ON emp.employee_id = prj.employee_id
ORDER BY department_name```

The result set

Here we're saying we are only interested in employees in known departments, but want to include employees with no projects.

(7 rows selected)

Referential Integrity in SQL

- SQL/92 and SQL:1999 support all options on deletes and updates.
  - Default is NO ACTION (delete/update is rejected)
  - CASCADE (delete/update all tuples that refer to deleted/updated tuple)
  - SET NULL/SET DEFAULT (sets foreign key value of referencing tuple)

CREATE TABLE Enrolled
    (sid CHAR(20) default '00',
     cid CHAR(20),
     grade CHAR(2),
     PRIMARY KEY (sid,cid),
     FOREIGN KEY (sid)
     REFERENCES Students (sid)
     ON DELETE SET DEFAULT
     ON UPDATE CASCADE)

Conceptual Design.
The Entity-Relationship (ER) Model

Database Design Overview

- Conceptual design
  - The Entity-Relationship (ER) Model, UML
  - High-level, close to human thinking
  - Semantic model, intuitive, rich constructs
  - Not directly implementable

- Logical Design
  - The relational data model
  - Machine-implementable, fewer and more basic constructs
  - Logical design translates ER into relational model (SQL)

- Physical Design
  - Storage and indexing details

CS430/630  
Lectures 10a, 11, 12

Slides based on "Database Management Systems" 3rd ed. Ramakrishnan and Gehrke
Relationship Set Representation

Representation Convention:
- Entity sets: rectangles
- Attributes: ovals, with key attributes underlined
- Relationship sets: diamonds
- Edges connect relationship sets to entity sets, and relationship sets to relationship set attributes
- Other notations are also in use: Murach uses "crow's foot" notation

Example 2

- Manages relationship: each dept has at most one manager
  - one-to-many from Employees to Departments, or
  - many-to-one from Departments to Employees

Participation Constraints

- Total vs Partial Participation
  - Total: every department must have a manager
    - "Departments" entity set has total participation in relationship
    - Represented as thickened line (there is a key constraint as well, represented by the arrow)
  - Partial: not every employee is a manager
    - "Employees" entity set has partial participation, thin line

“Each department has at most one manager” key constraint on Manages

- This means Manages is many-to-one (Departments to Employees) and deserves an arrow from Departments to Manages:

  - It’s not a thick arrow because a department may have no manager at all (this is assumed unless explicitly stated otherwise)

Variant 1 for “Each department has at most one manager” key constraint on Manages

- Map relationship to its own table:
  - Note that did is the key now!
  - That means we can’t have two rows here for one department, pointing to two managers for that department.
  - Note this allows a department without a manager: it would simply be missing from this table.

CREATE TABLE Manages(
  ssn CHAR(11),
  did INTEGER,
  since DATE,
  PRIMARY KEY (ssn, did),
  FOREIGN KEY (ssn) REFERENCES Employees,
  FOREIGN KEY (did) REFERENCES Departments)
Variant 2 for “Each department has at most one manager” key constraint on Manages

- This is the “thin arrow” case in a binary relationship
- Since each department has at most one manager, we could instead combine Manages and Departments.
- And use a nullable FK on ssn.
- A department without a manager would have a null ssn

```sql
CREATE TABLE Dept_Mgr(
  did INTEGER,
  dname CHAR(20),
  budget INTEGER,
  ssn CHAR(11), -- note nullable
  PRIMARY KEY (did),
  FOREIGN KEY (ssn) REFERENCES Employees
)
```

Variant 2 for “Each department has exactly one manager” key constraint on Manages

- This is the “thick arrow” case in a binary relationship
- We can again combine Manages and Departments.
- And use a not-null FK on ssn.
- Variant 1 can’t implement the total participation constraint, so this is the preferred schema

```sql
CREATE TABLE Dept_Mgr(
  did INTEGER,
  dname CHAR(20),
  budget INTEGER,
  ssn CHAR(11) not null, -- note “not null” here
  PRIMARY KEY (did),
  FOREIGN KEY (ssn) REFERENCES Employees
)
```

Creating a view

```sql
CREATE VIEW RegionalSales AS
SELECT P.category, S.sales, L.state, S.id
FROM Products P, Sales S, Locations L
WHERE P.pid = S.pid AND S.locid = L.locid
```

Views as subqueries

```sql
SELECT P.category, R.state, SUM(R.sales)
FROM RegionalSales R
GROUP BY R.category, R.state
```

Querying views

- Views are queried just like regular tables
- A view is just another relation (albeit a virtual one)
- Queries can involve both views and base tables
- Helps to think of views in terms of analogy with window on data

```sql
SELECT R.category, R.state, SUM(R.sales)
FROM RegionalSales R
GROUP BY R.category, R.state
```

Equivalent Query (without views)

```sql
SELECT P.category, R.state, SUM(R.sales)
FROM Products P, Sales S, Locations L
WHERE P.pid = S.pid AND S.locid = L.locid
GROUP BY P.category, R.state
```