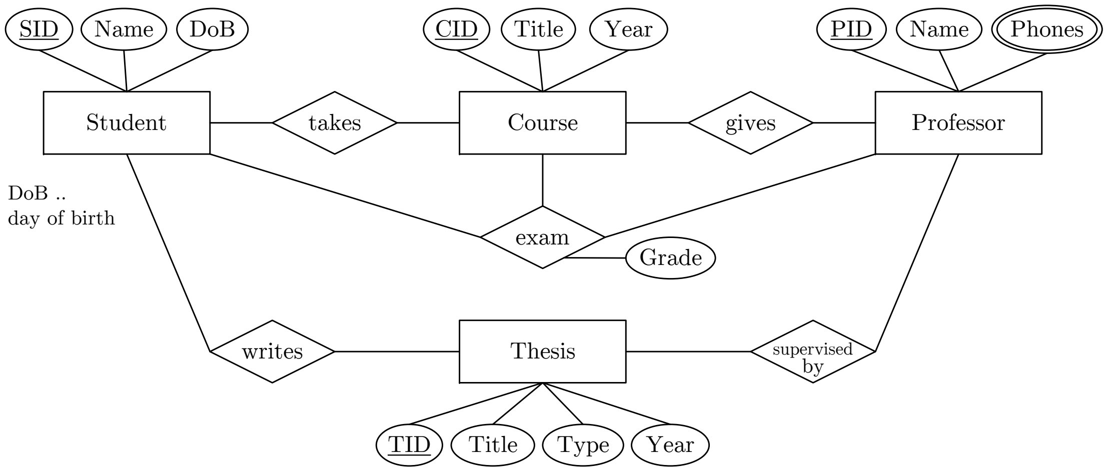
Univ.-Prof. Dr.-Ing. Matthias Boehm
Graz University of Technology
Computer Science and Biomedical Engineering
Institute of Interactive Systems and Data Science
BMK endowed chair for Data Management

July 01, 2020

Exam INF.01017UF Data Management (Summer 2020, V2a)

Important notes: The working time is 90min, and lecture materials or any kind of mobile devices are not allowed. Please, make sure to put your name and matriculation number on the top right of each piece of paper. You may give the answers in English or German, as well as directly write into the task description.

Task 1 Data Modeling (25 points)



(a) Given the above Entity-Relationship diagram, specify the cardinalities in Modified Chen notation based on the following information. (**10 points**)

- A student can take up to 64 courses, and a single course can be taken by up to 1000 students.
- A professor can give an arbitrary number of courses (including none), but every course it taught by exactly one professor. The phones attribute of a professor is a multi-valued attribute containing a list of phone numbers (e.g., {111-222-3333, office}; {444-555-6666 mobile}; {777-888-9999 private}).
- A student might have written multiple theses (e.g., types BS, MS, PhD), and every thesis is written by exactly one student. A thesis is also supervised by exactly one professor, and professors can supervise an arbitrary number of theses.
- A professor can take exams for many (an arbitrary number of) students of a course, and for many courses of a student, but the exam of a specific student-course combination is taken by exactly one professor.

(b) Map the given Entity-Relationship diagram into a relational schema in third normal form, including data types, primary keys, and foreign keys. Your schema should also ensure that each course has an associated professor, and each thesis has a student (author) and professor (supervisor). (**13 points**)

(c) Assume a relation Student(SID, Name, DoB), where SID is a unique and defined (not null) attribute. List all valid super keys and candidate keys (as attribute sets). (**2 points**)

- Super keys:
- Candidate keys:

Task 2 Structured Query Language (30 points)

Employees

EID	FName	LName	Age	Country	PID
4	Isabella	Brown	30	AT	2
2	Olivia	Johnson	30	FR	1
1	Emma	Smith	35	DE	3
3	Ava	Williams	20	DK	1
5	Sophie	Jones	35	AT	2
6	Taylor	Miller	55	DE	5
7	Charlotte	Davis	40	DE	2

Projects

PID	Name	Customer
1	UX Design	B
2	App Backend	B
3	Data Storage	C
4	ML Pipeline	A
5	UX Design	A
6	HW Accelerator	D

(a) Given the Employees and Projects tables above, and compute the results for the following three queries: (**15 points**)

```
Q1: SELECT DISTINCT P.Customer, P.Name
      FROM Employees E, Projects P
      WHERE E.PID = P.PID
        AND E.LName IN('Williams','Jones','Miller')
```

```
Q2: SELECT FName, LName
      FROM Employees WHERE Country = 'DE'
    UNION DISTINCT
    SELECT FName, LName
      FROM Employees WHERE Age >= 35
```

```
Q3: SELECT P.Name, round(avg(E.Age)) --avg=sum/count
      FROM Employees E, Projects P
      WHERE E.PID = P.PID
      GROUP BY P.Name
```

(b) Given the Employees and Projects table schemas above, write SQL queries to answer the following questions (in a way that is independent of the shown tuples): (**15 points**)

- Q4: Which employees work on projects for customer B (return the FName and LName, sorted in ascending order of LName)?

- Q5: Which customers have more than one project (return the Customer, and number of projects per Customer)?

- Q6: Which projects are not worked on by any employee (return the project PID, Name, and Customer)?

Task 3 Query Processing (20 points)

(a) Assume relations $R(a, b, c)$ and $S(d, e)$, and indicate in the table below whether or not the two relational algebra expressions per row are equivalent in bag semantics (✓for equivalent, × for non-equivalent). For non-equivalent expressions explain why. (**5 points**)

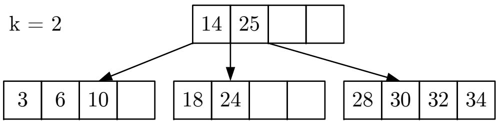
Expression 1	**Expression 2**	**Equivalent? Why Not?**
$\sigma_{c=3}(\sigma_{b=7}(R))$	$\sigma_{c=3\wedge b=7}(R)$	
$(\sigma_{c=3}(R)) \cap (\sigma_{b=7}(R))$	$\sigma_{c=3\vee b=7}(R)$	
$R \bowtie_{a=e} S$	$\sigma_{a=e}(R \times S)$	
$\pi_{b,d}(R \bowtie_{a=e} S)$	$(\pi_{a,b}(R)) \bowtie_{a=e} (\pi_{d,e}(S))$	
$R - (\sigma_{a<b\wedge b<c\wedge a=c}(R))$	R	

(b) Draw a logical query tree for each of the queries Q2 and Q3 from Task 2a. (**6 points**)

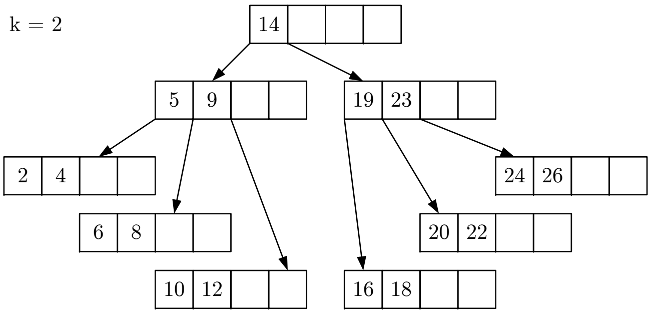
(c) Describe the volcano (open-next-close) iterator model by example of a selection operator $\sigma_{b \geq 7}(R)$ in both forms of a generic selection and an index scan, where the latter can utilize an index $\text{IX}_{R.b,\text{ASC}}$ for point lookups and ordered scans. For both operators, also state the time and space complexity assuming $N = |R|$, $M = |\sigma_{b \geq 7}(R)|$, and $N \gg M$. (**9 points**)

Task 6 Physical Design (10 points)

(a) Given the B-tree (k=2) below, insert key 26, and draw the resulting B-tree. (**4 points**)



(b) Given the B-tree (k=2) below, delete key 5, and draw the resulting B-tree. (**6 points**)



Task 4 Transaction Processing (10 points)

(a) Explain the concept of a database transaction log, and how it helps to ensure Atomicity and Durability of changes made by uncommitted and committed transactions in failure scenarios. (**6 points**)

(b) Indicate in the table below, which operation schedules are equivalent (✓for equivalent, × for non-equivalent). The notation $r_1(a)$ and $w_2(b)$ refers to the read of object a by transaction T_1 and the write of object b by T_2. (**4 points**)

Schedule 1	**Schedule 2**	**Equivalent?**
$\{r_1(a), w_1(a), r_2(b), w_2(b)\}$	$\{r_1(a), r_2(b), w_1(a), w_2(b)\}$	
$\{r_1(c), w_1(c), r_2(c), r_2(d), w_2(d)\}$	$\{r_1(c), r_2(c), r_2(d), w_1(c), w_2(d)\}$	
$\{r_1(e), w_1(e), w_2(e), w_2(f)\}$	$\{r_1(e), w_2(e), w_1(e), w_2(f)\}$	
$\{r_1(g), r_1(h), r_2(g), r_2(h), w_2(h)\}$	$\{r_2(g), r_2(h), r_1(h), r_1(g), w_2(h)\}$	

Task 5 Distributed Data Analysis (5 points)

Explain Apache Spark's abstraction of Resilient Distributed Datasets (RDDs), and how it facilitates data-parallel computation in distributed environments.