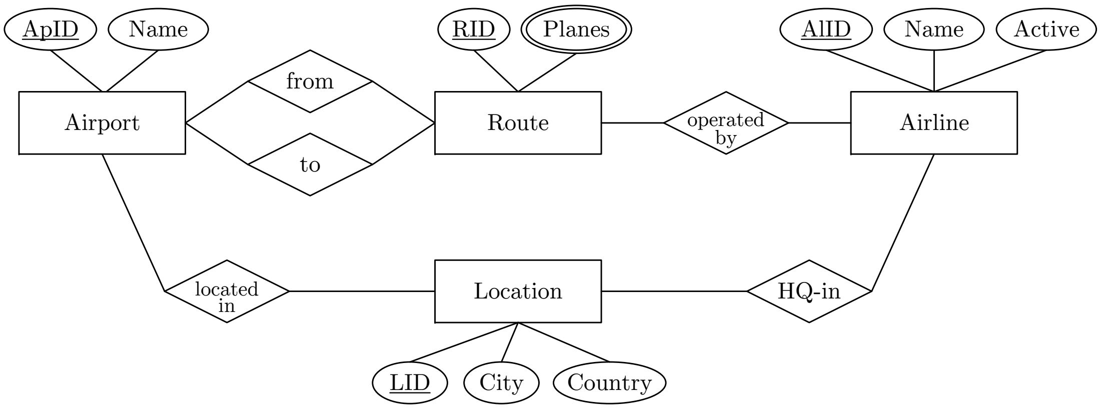
Univ.-Prof. Dr.-Ing. Matthias Boehm
Graz University of Technology
Computer Science and Biomedical Engineering
Institute of Interactive Systems and Data Science
BMVIT endowed chair for Data Management

June 22, 2020

Exam INF.01017UF Data Management (Summer 2020, V1a)

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. (**9 points**)

- An airport can be the source (from) and target (to) of at least 1 but potentially many routes, and every route has exactly one source and one target airport.
- A route is operated by exactly one airline and airlines can operate between 1 and 1024 routes. The active status of an airline is a boolean indicator.
- Airports and the headquarters (HQ) of airlines are both located in exactly one location (city), and locations might have up to 16 airports, and host an arbitrary number of airline HQs. There might be cities with the same name in different countries.
- The planes attribute of a route is multi-valued, containing a list of plane types (e.g., Boeing 747, Airbus A380) used on a route. A plane type might be used on arbitrary many routes, and between 1 and 32 plane types can be used on a single route.

(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 route has an associated source, target, and operating airline. (**12 points**)

(c) Assume a relation $R(a, b, c, d)$ where both a and b are unique and defined (not null) attributes. Indicate in the table below, which attribute sets are valid super keys and candidate keys (✓for valid; otherwise blank), and then pick a valid primary key. (**4 points**)

Attributes	**Super Key**	**Candidate Key**	**Primary Key**
$\{a, b, c, d\}$			
$\{a, b, c\}$			
$\{a, b\}$			
$\{b, c\}$			
$\{c, d\}$			
$\{a\}$			
$\{b\}$			
$\{c\}$			

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	A
4	ML Pipeline	C
5	UX Design	C

(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**)

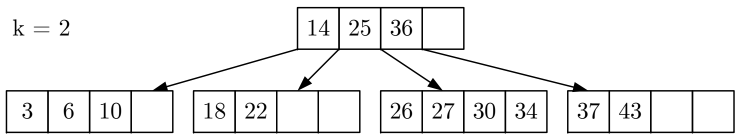
Expression 1	Expression 2	Equivalent? Why Not?
$\sigma_{c=3}(\sigma_{b=7}(R))$	$\sigma_{c=3 \wedge b=7}(R)$	
$(\sigma_{c=3}(R)) \cap (\sigma_{b=7}(R))$	$\sigma_{c=3 \vee b=7}(R)$	
$R \bowtie_{a=e} S$	$\sigma_{a=e}(R \times S)$	
$\pi_{b,d}(R \bowtie_{a=e} S)$	$(\pi_{a,b}(R)) \bowtie_{a=e} (\pi_{d,e}(S))$	
$R - (\sigma_{a<b \wedge b<c \wedge a=c}(R))$	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 table scan 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 with k=2 below, delete the key 14, then insert key 31, and draw the resulting final B-tree. (**7 points**)



(b) Describe the concept of horizontal table partitioning by providing relational algebra expressions for (1) partitioning, and (2) reconstructing a relation R. (**3 points**)

Task 4 Transaction Processing (10 points)

(a) Briefly describe the ACID transaction properties Atomicity, Consistency, Isolation, and Durability. (**4 points**)

(b) 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**)

Task 5 Distributed Data Analysis (5 points)

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