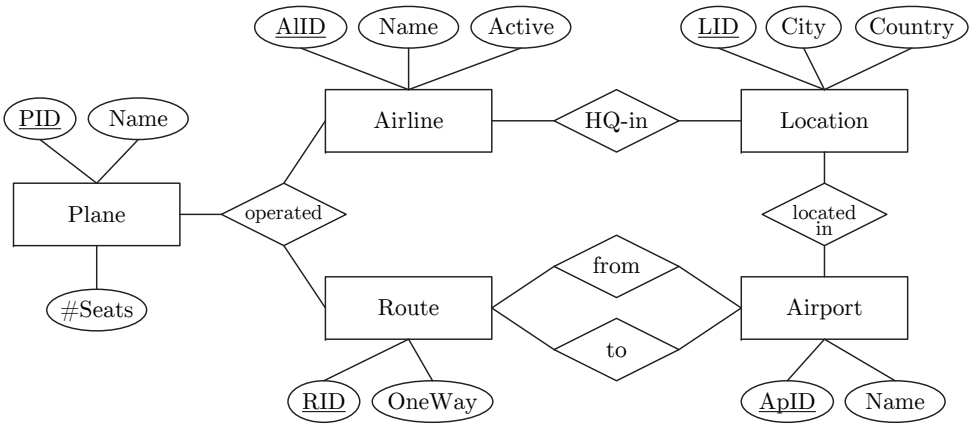
Univ.-Prof. Dr.-Ing. Matthias Boehm
Graz University of Technology
Computer Science and Biomedical Engineering
Institute of Interactive Systems and Data Science
BMK endowed chair for Data Management

June 30, 2021

Exam INF.01017UF Data Management (Summer 2021, 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 the first page of the task description, and each additional piece of paper. You may give the answers in English or German, written directly into the task description.

Task 1 Data Modeling (25 points)

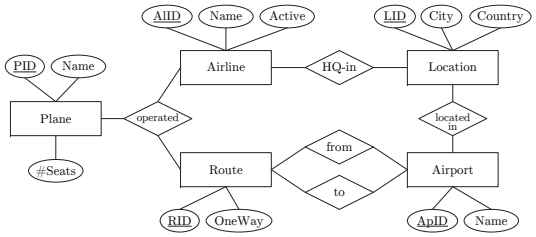


(a) Given the above Entity-Relationship diagram, specify the cardinalities in Modified Chen notation based on the following information. (**10 points**)

- 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.
- 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.
- An airline operates a route with at least 1 but potentially many planes; a single plane on a particular route is operated by exactly one airline; an airline might use the same plane to operate an arbitrary number of routes. The active status of an airline and the one-way flag of a route are boolean indicators.

(b) Briefly describe the first (1NF), second (2NF), and third (3NF) normal form. (**3 points**)

(c) 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 (from) and target (to) airport. (**12 points**)



Task 2 Structured Query Language (30 points)

Customers

CID	Name	Country
1	Red	AT
2	Orange	CH
3	Yellow	DE
4	Green	DE
5	Blue	AT
6	Violet	DE

Orders

CID	PID	Date	Qty
5	1	2021-06-29	4
6	3	2021-06-30	1
5	3	2021-06-30	2
2	3	2021-06-30	1
5	2	2021-06-30	10
3	1	2021-06-30	3
6	5	2021-06-30	1

Products

PID	Name	Price
1	Monitor	400
2	SSD	200
3	Laptop	2500
4	Tablet	600
5	Headphones	150

(a) Given the Customers, Orders, and Products tables above, compute the results for the following three queries: (**15 points**)

```
Q1: SELECT DISTINCT O.Date, P.Name
      FROM Orders O, Products P
      WHERE O.PID = P.PID
        AND P.Price > 200
```

```
Q2: SELECT C.Country, count(*)
      FROM Customers C, Orders O
      WHERE C.CID = O.CID
      GROUP BY C.Country
      ORDER BY count(*) DESC, C.Country
```

```
Q3: SELECT C.Name, C.Country, O.Date
      FROM Customers C LEFT JOIN Orders O
        ON C.CID = O.CID
      WHERE C.Country IN('AT','CH')
```

(b) Given the Customers, Orders, and Products tables above, write SQL queries to answer the following questions (in a way that is independent of the shown tuples): (**15 points**)

- Q4: Which products where ordered by customers from Germany, i.e., Country=DE? (return the product name and price)?

- Q5: Compute the revenue (i.e., `sum(Qty*Price)`) per product (return the product name, revenue, sorted ascending by name)?

- Q6: Which customers did not place any orders (return all customer attributes)?

Task 3 Query Processing (17 points)

(a) Given a relation $R(x, y, z)$ with four tuples (a, b, c), (d, e, f), (a, b, d), and (d, e, g), indicate in the table below for each relational algebra expression (row) the number of output tuples in set and bag (multiset) semantics, respectively. (**6 points**)

RA Expression	Set Semantics	Bag Semantics
$\pi_{x,y}(R)$		
$\sigma_{x=a \ \vee \ z=f}(R)$		
$R \cup R$		

(b) Draw a logical query tree for the following query. (**5 points**)

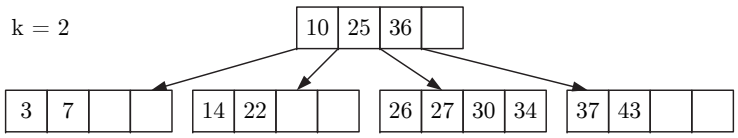
```
(SELECT Name
   FROM Products
   WHERE Price > 750)
UNION ALL
(SELECT P.Name
   FROM Orders O, Products P
   WHERE O.PID = P.PID
   GROUP BY P.Name
   HAVING sum(O.Qty) >= 5)
```

(c) Describe the conceptual ideas of a nested-loop join, and a hash join. Furthermore, assume $R \bowtie S$ with cardinalities $N = |R|$ and $M = |S|$, and enter the space and time complexity of these operators (in the open-next-close iterator model) in the table below. (**6 points**)

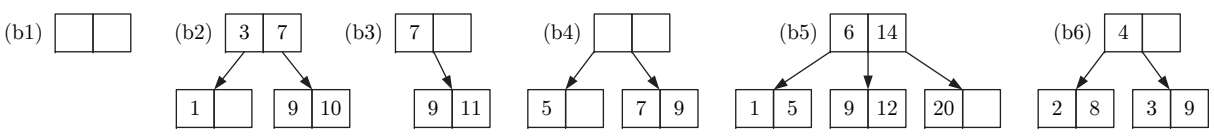
Operator	Time Complexity	Space Complexity
Nested Loop Join		
Hash Join		

Task 4 Physical Design (13 points)

(a) Given the B-tree with k=2 below, delete the key 10, then insert key 31, and draw the resulting final B-tree (or both trees individually). (**7 points**)



(b) Which of the following trees are valid—i.e., satisfy the constraints of—B-trees with k=1. Mark each tree as valid (✓), or invalid (×) and name the violations. (**6 points**)



Task 5 Transaction Processing

Explain the concept of a database transaction log, the recovery process, and how this approach ensures Atomicity and Durability of changes made by uncommitted and committed transactions in failure scenarios. (**8 points**)

Task 6 Distributed Data Analysis

Explain the vertex-centric graph processing model by example of computing the connected components of a graph. In this context, also elaborate how data-parallel computation frameworks like Apache Spark can be used for scale-out in a distributed environment. (**7 points**)