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

February 04, 2022

Exam 706.520 Data Integration and Large-Scale Analysis (WS21/22)

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 Message-oriented Middleware (5 points)

Assume a message-oriented middleware with a single FIFO message queue. Indicate, in the table below, true (✓) and false (×) properties of the following three message delivery guarantees.

	At Most Once	**At Least Once**	**Exactly Once**
Requires Message Persistence			
Requires Transactional Behavior			
Prevents Message Outrun			
Prevents Message Loss			
Prevents Message Double Delivery			

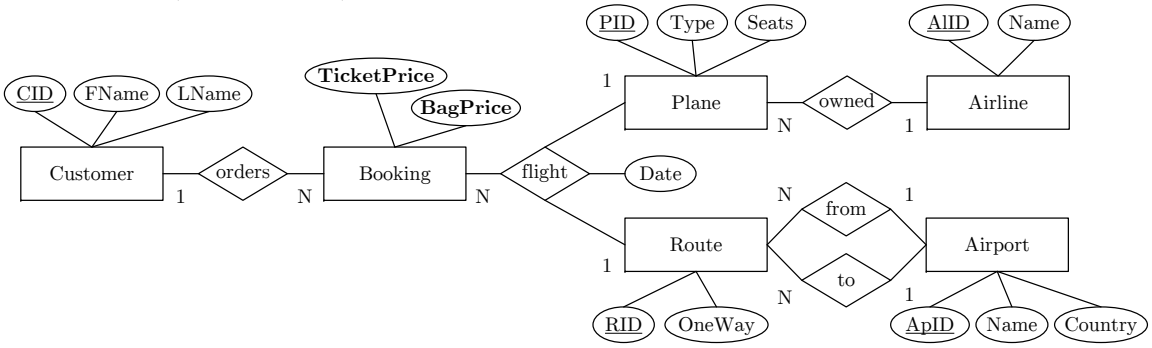
Task 2 Data Warehousing (25 points)

(a) Describe the overall system architecture of a *data warehouse*, name its components, and briefly describe the purpose of these components. (**6 points**)

(b) The central metaphor of multi-dimensional modeling is the data cube, described by dimensions and measures. Explain the following related concepts with examples. (**3 points**)

- Dimension Hierarchy:
- Fact:
- Measure:

(c) Given the entity relationship (ER) diagram below, create corresponding relational *star and snowflake schemas*. Data types can be ignored, but indicate primary and foreign key constraints. (**8+8 points**)



Star Schema:

Snowflake Schema:

Task 3 Entity Resolution (20 points)

(a) Explain the phases of a typical *entity resolution pipeline* (deduplication pipeline), and discuss example techniques for the individual phases. (**16 points**)

(b) Schema detection is a common preparation step for both schema matching/mapping and entity resolution. Name example techniques for detecting data types, primary-key/foreign-key relationships, as well as semantic types. (**4 points**)

Task 4 Data Cleaning (12 points)

In the context of missing value imputation, describe the the following types of missing data, name related techniques for *missing value imputation*, and provide imputed values for the missing values on the right and the different imputation techniques.

Name	Age	Salary
Red	45	4500
Orange	50	NULL
Yellow	20	2000
Green	40	4000
Blue	25	2500
Violet	35	NULL

- Missing Completely at Random (MCAR):

- Missing at Random (MAR):

- Not Missing at Random (NMAR):

Task 5 Data Provenance (5 points)

(a) Explain the general goal and concept of *data provenance*. (**2 points**)

(b) Given the tables R and S below (with tuples r_i and s_i, respectively), compute the query results, and provide the *provenance polynomials* for every result tuple. (**3 points**)

R

	A	B
r_1	X	1
r_2	Y	2
r_3	Z	1

S

	C	D
s_1	1	A
s_2	2	B
s_3	2	A
s_4	2	C

```
SELECT R.A, count(*)
  FROM R, S
  WHERE R.B = S.C
  GROUP BY R.A
```

Task 6 Cloud Computing (7 points)

(a) Explain the motivation of cloud computing in terms of overall goal, key drivers, and advantages. (**4 points**)

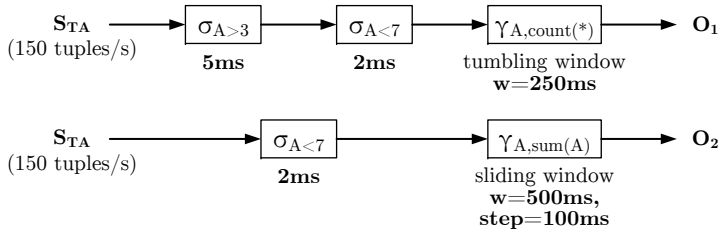
(b) Discuss the advantages and disadvantages of *task scheduling* for multiple workers with a single task queue, and per-worker task queues. (**3 points**)

- Single Task Queue:

- Per-worker Task Queues:

Task 7 Stream Processing (8 points)

Assume an input stream S_{TA} with schema $S(T, A)$—where T refers to event time and A is a positive integer column—as well as the two continuous queries below (filter on A, group-by A, return count/sum) with *stream window aggregation*.



(a) Draw an optimized continuous query that produces semantically equivalent output streams O_1 and O_2, but avoids unnecessary redundancy. (**4 points**)

(b) Compute the maximum attainable output stream rates (tuples/second) for both output streams O_1 and O_2. (**4 points**)

Task 8 Distributed, Data-Parallel Computation (18 points)

(a) Given the distributed dataset of three partitions below, describe a data-parallel—potentially multi-phase—approach for estimating the number of distinct items of Attr1, and Attr2, respectively. In detail, (a) explain an approach of estimating the number of distinct items, (b) describe or draw its data-parallel execution, and (c) discuss means for improving performance, and ensuring fault tolerance in case of task failures. (**4+8+6 points**)

Attr1	**Attr2**
X	3
X	4
X	1
Y	7

X	2
Y	3.7
X	1
X	2

Y	5
X	3.7
Z	8
X	4