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

April 29, 2021

Exam 706.520 Data Integration and Large-Scale Analysis (WS20/21)

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 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) Assume two input publication datasets A and B that need deduplication. Explain the following two categories of schema matching techniques for pre-processing. (**4 points**)

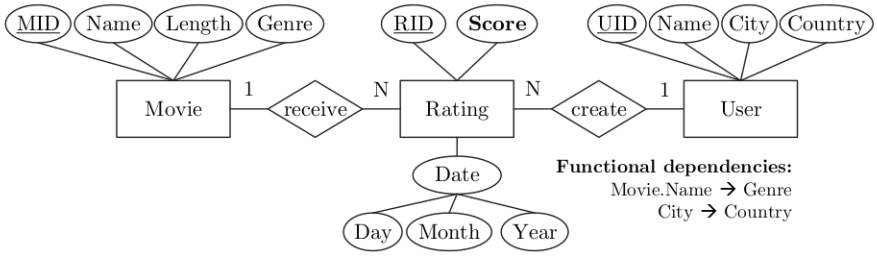
- Schema-based Matching:

- Instance-based Matching:

Task 2 Data Warehousing (15 points)

(a) Describe the overall system architecture of a *data warehouse* (not data center), name its components, and briefly describe the purpose of these components. (**5 points**)

(b) Given below entity relationship (ER) diagram, create the corresponding *star and snowflake schemas*. Data types can be ignored, but indicate primary and foreign key constraints. (**5+5 points**)



Task 3 Data Cleaning (20 points)

(a) In the context of missing value imputation, describe the the following types of missing data (**9 points**)

- Missing Completely at Random (MCAR):

- Missing at Random (MAR):

- Not Missing at Random (NMAR):

(b) Given the data below, name two techniques for *missing value imputation* (one for MCAR and one for MAR), and provide the imputed values. (**5 points**)

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

(c) Explain the difference between Outlier Detection and Anomaly Detection, with at least one example strategy for each. (**6 points**)

- Outlier Detection:

- Anomaly Detection:

Task 4 Data Provenance (8 points)

(a) Explain the general goal and concept of *data provenance*, and distinguish *why-provenance and how-provenance*. (**5 points**)

- Data Provenance:

- Why-Provenance:

- How-Provenance:

(b) Given below tables R and S (with tuples r_i and s_i, respectively), query Q and the results O, specify the *provenance polynomials* for every tuple in O. (**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 DISTINCT S.D
  FROM R, S
  WHERE R.B=S.C
```

O	Provenance Polynomials?
A	
B	
C	

Task 5 Cloud Computing (7 points)

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

(b) Explain the concept of *resource allocation* for multiple resources such as CPU and memory (e.g., dominant resource calculation in YARN). (**3 points**)

Task 6 Distributed, Data-Parallel Computation (20 points)

(a) Given the distributed dataset of three partitions below (left), describe a data-parallel, potentially multi-phase, approach of imputing the missing values (NULL) of Attr1 with its mode, and the missing values of Attr2 with its mean. Describe strategies for improving the performance of the entire (multi-phase) computation. Finally, fill in the concrete imputed values in the result below (right). (**12+5+3 points**)

Attr1	**Attr2**
X	3
X	4
NULL	1
Y	7
X	2
Y	NULL
X	1
X	2
Y	5
NULL	NULL
Z	8
NULL	4

Imputed

Attr1	**Attr2**
X	3
X	4
	1
Y	7
X	2
Y	
X	1
X	2
Y	5
Z	8
	4

Task 7 Stream Processing (10 points)

(a) Assume an input stream S with schema $S(A, T)$—where T refers to event time and A is an integer column—as well as the continuous query Q (filter $3 < A < 7$, group-by A, return count) with *stream window aggregation* below. Compute the maximum output stream rate (tuples/second) for the following two window definitions w. (**4 points**)

- Tumbling Window (size 200 ms):

- Sliding Window (size 500 ms, step 100 ms):

(b) Explain the following three techniques for *handling overload* situations in stream processing engines. (**6 points**)

- Back Pressure:

- Load Shedding:

- Distributed Stream Processing: