Univ.-Prof. Dr.-Ing. Matthias Boehm
Technische Universität Berlin
Faculty IV - Electrical Engineering and Computer Science
Berlin Institute for the Foundations of Learning and Data (BIFOLD)
Big Data Engineering (DAMS Lab) Group

February 06, 2025

Exam Data Integration and Large-Scale Analysis (WiSe 24/25)

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 your own paper. You may give the answers in English or German, written directly into the task description.

Task 1 Data Warehousing (25 points)

(a) Describe the overall system architecture of a *data warehouse* (not a data center), name its components, and briefly describe the purpose of these components. (**6 points**)

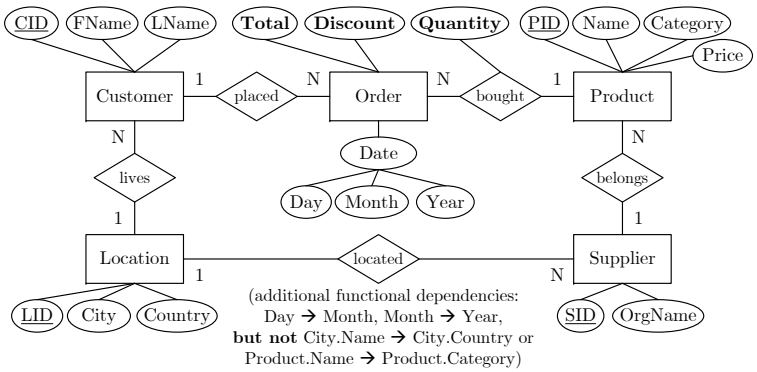
(b) The SQL standard defines the multi-grouping extensions `GROUPING SETS`, `ROLLUP`, and `CUBE`. Given the table `Sales` and SQL query below, compute and list the results. (**5 points**)

Sales

Year	Quarter	Revenue
2023	Q4	25
2024	Q1	20
2024	Q2	15
2024	Q3	25
2024	Q4	30
2025	Q1	20

```
SELECT Year, Quarter,
       SUM(Revenue) AS Rev,
       GROUPING(Year) AS Agg
  FROM Sales
  WHERE Year > 2023
  GROUP BY ROLLUP(Year, Quarter)
```

(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. (**7+7 points**)



Star Schema:

Snowflake Schema:

Task 2 Message-oriented Middleware (5 points)

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

	At Most Once	**At Least Once**	**Exactly Once**
Requires Message Persistence			
Requires Transaction Mechanism			
Prevents Message Outrun			
Prevents Message Loss			
Prevents Message Double Delivery			

Task 3 Schema Matching and Mapping (6 points)

Characterize the concepts of schema matching and schema mapping by indicating in the table below true (✓) characteristics.

	Schema Matching	**Schema Mapping**
Produces Schema Correspondences		
Consumes Schema Correspondences		
Applies Similarity Functions		
Analyzes Available Data		
Utilizes Schema Constraints		
Produces Transformations Programs		

Task 4 Entity Resolution (16 points)

Explain the phases of a typical *entity resolution pipeline* (deduplication pipeline), and name two example techniques for each individual phase.

Task 5 Data Cleaning (8 points)

In the context of data validation, what are useful statistics or metrics for evaluating data quality? Name three of such metrics, describe how they can be used for data validation and cleaning, and compute them for each column of the table on the right.

Country	Population	Region
Germany	84.7 M	Central
Denmark	5.9 M	North
Poland	37.5 M	Central
Czechia	NULL	Central
Austria	9.1 M	Central
Switzerland	8.9 M	Central
France	68.5 M	West
Luxembourg	NULL	Central
Belgium	NULL	West
Netherlands	18.0 M	West

Task 6 Data Provenance (8 points)

(a) Explain the general goal and concept of *data provenance* in a broad sense. (**3 points**)

(b) Given the tables R and S below (with tuples r_i and s_i, respectively), provide the *provenance polynomials* for every result tuple in the table on the right. (**5 points**)

```
SELECT DISTINCT S.D
  FROM R, S
  WHERE R.B = S.C
```

R

	A	B
r_1	B	1
r_2	C	2
r_3	D	4
r_4	E	2

S

	C	D
s_1	1	X
s_2	2	Y
s_3	4	X
s_4	5	Z
s_5	1	X

Output	Provenance Polynomials
D	
X	
Y	

Task 7 Cloud Computing (4 points)

In the context of resource management and scheduling, discuss the advantages and disadvantages of a single task queue versus per-worker task queues.

Task 8 Distributed, Data-Parallel Computation (14 points)

(a) Given the distributed dataset D of three partitions below, describe the data-parallel (MapReduce-like) computation of $Q : \gamma_{A,\max(B)}(D)$ (group-by A, return max(B) per group) including how shuffling works and the *actual example intermediates and results*. (**7 points**)

A	B
X	3
X	4
X	1
Y	7

X	2
Y	3
X	1
X	2

Y	5
X	3
Z	7
X	4

(b) Briefly name three conceptual techniques for improving the runtime performance of the data-parallel computation above. (**3 points**)

-
-
-

(c) Name and describe techniques for ensuring fault-tolerance in distributed computation as well as distributed storage. (**2+2 points**)

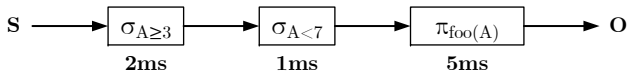
Task 9 Stream Processing (8 points)

(a) Assume an input stream S with schema $S(A, T)$—where T is the event time (the smaller the older, start at zero)—and a query $Q : \gamma_{A,count()}(S)$ (group-by A, return count) with *stream window aggregation*. Compute the output stream with schema (A, count, T_c), where T_c is the creation time (first output at full window size). (**4 points**)

- Input Stream:
  (X,0.5s), (Y,1.1s), (X,2.1s), (Y,2.9s), (X,4.1s), (X,4.4s), (X,4.5s), (X,5.2s), (X,5.9s), (Y,7.1s), (Y,8.8s), (X,8.9s), (X,10.1s), (X,10.7s), (Y,11.8s), (Z,11.9s).

- Tumbling Window (size 3s):

(b) Given the input stream S and continuous query below, compute the latency of individual tuples (in milliseconds), and maximum tuple throughput (in tuples/second). (**4 points**)



- Tuple Latency [ms]:

- Tuple Throughput [tuples/s]:

Task 10 Machine Learning Systems (6 points)

Describe the basic overall system architecture of a *data-parallel parameter server*, explain its components and interaction among these components.