Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

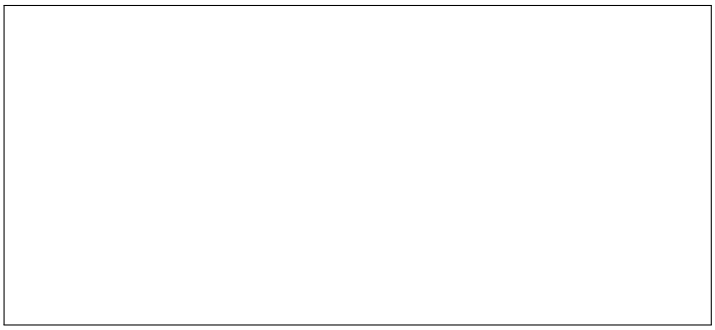
Universität Hamburg
Fachbereich Informatik
Datenbanken und Informationssysteme
Prof. Dr. Norbert Ritter, Dr. Fabian Panse

Module

Databases and Information Systems

Summer Term 2019

Mock Exam
11.07.2019, 8:00 to 10:00, ESA B



Part	1	2	3	4	5	6	7	8	Total	Grade
Points	18	18	12	12	15	18	17	10	120	
Result										

Rechtsmittelbelehrung (Instructions on Right of Appeal):

Gegen die Bewertung dieser Prüfungsleistung kann innerhalb eines Monats nach ihrer Bekanntgabe Widerspruch erhoben werden. In diesem Zeitraum kann die Bewertung der Klausur eingesehen werden. Der Widerspruch ist schriftlich oder zur Niederschrift beim Vorsitzenden des B.Sc./M.Sc.-Prüfungsausschusses einzulegen. Es wird darauf hingewiesen, dass ein erfolgloses Widerspruchsverfahren kostenpflichtig ist.

For your convenience, we provide an **English translation** below.
Please note that the English version is not legally binding:

You may appeal the examination result within one month after its publication. During this time period, the evaluated examination will be made accessible to you. The appeal must be made in writing or declared for recording at the B.Sc./M.Sc. examinations board ("B.Sc./M.Sc.-Prüfungsausschuss").
Please note that an unsuccessful appeal may be subject to a charge.

Bearbeitungshinweise:

- **Sie können die Aufgaben wahlweise in deutsch oder Englisch bearbeiten.**
- Die Bearbeitungsdauer der Klausur beträgt insgesamt 120 Minuten.
- Es werden maximal 120 Punkte vergeben. Zum Bestehen der Klausur sind 50% der erreichbaren Punkte hinreichend.
- Überprüfen Sie sofort nach Erhalt der Klausur die Vollständigkeit der Unterlagen (27 Seiten).
- Bitte lassen Sie die Klausur zusammengeheftet. Notieren Sie alle Lösungen direkt auf den Aufgabenblättern.
- Verwenden Sie kein eigenes Papier, sondern nur die zur Verfügung gestellten Blätter.
- Bitte schreiben Sie mit blauem oder schwarzem Schreiber – Bleistift und Rotstift sind nicht erlaubt.
- Schreiben Sie auf jedes Blatt Ihren Namen und Ihre Matrikelnummer.
- Es sind keine Hilfsmittel zugelassen.
- Bitte legen Sie Ihren Studentenausweis und einen Lichtbildausweis auf den Tisch.
- Schalten Sie Ihr Mobiltelefon aus!

Working Instructions:

- **You may use German or English according to your preference.**
- You will be given 120 minutes to complete the exam.
- You can achieve at most 120 points. 50% of the achievable points will suffice to pass the exam.
- Immediately upon receipt, please check whether your exam is complete (27 pages).
- Please do not remove the staples from the exam. Only use the provided document to write down solutions.
- Please do not use your own paper, but the provided paper instead.
- Please write with a pen or biro using blue or black ink – lead pencils or pens with red ink are not allowed.
- Please write your name and student number ("Matrikelnummer") on each and every sheet of paper.
- No aids are allowed.
- Please put your student ID card ("Studentenausweis") and a photo ID card ("Lichtbildausweis") on the table.
- Please turn off your mobile phone!

Hiermit versichere ich, dass ich die Klausur selbstständig und unter ausschließlicher Nutzung der erlaubten Hilfsmittel bearbeitet und die Rechtsmittelbelehrung zur Kenntnis genommen habe.

I hereby declare that I have written the exam by myself and under exclusive use of the permitted aids. I am aware of my right to appeal.

Unterschrift (Signature): _____

Blättern Sie erst weiter, wenn die Bearbeitungszeit offiziell begonnen hat.

Do not turn the page, before the examination has started.

Part 1: Concurrency Control: Correctness

[18 P.]

a) Consider the following transaction schedules. For each schedule, do the following:

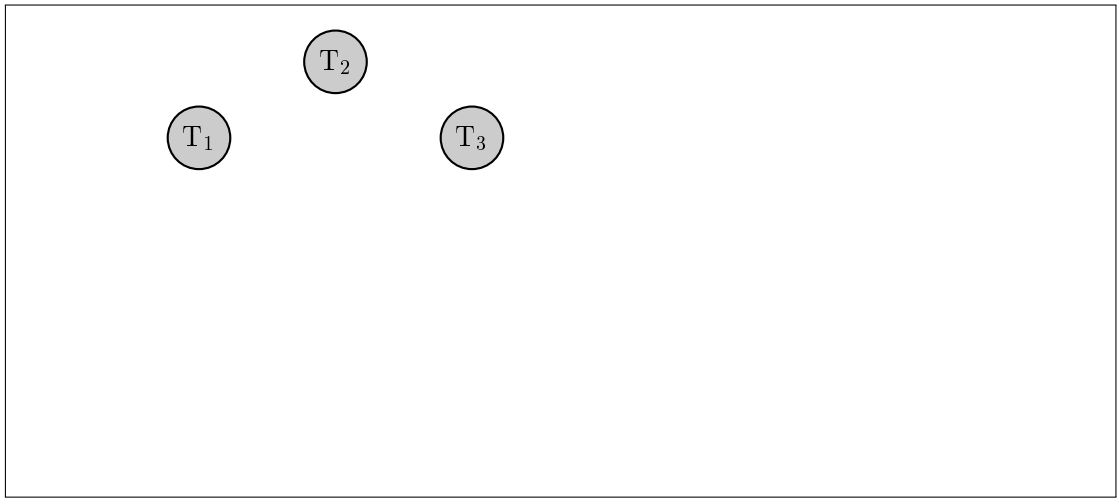
- draw the conflict (precedence) graph,
- indicate whether or not the schedule is **conflict-serializable (CSR)**, and if so give a conflict-equivalent serial schedule (you just need to list the order of transactions), and
- indicate whether or not the schedule is **view-serializable (VSR)**.

Briefly justify your answers.

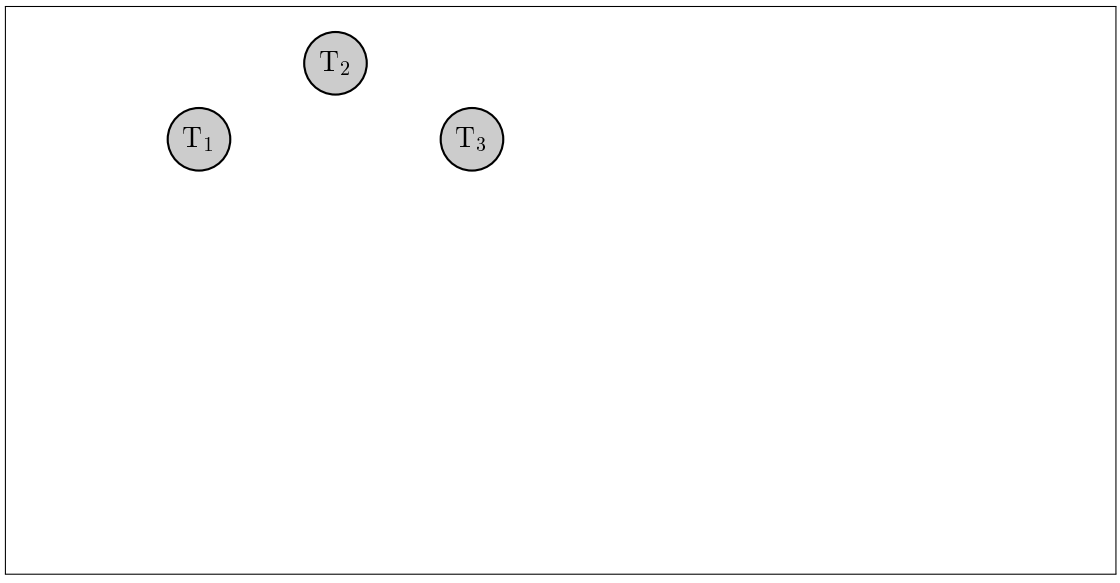
Consider the following notation for operations of transactions:

$w_i(x)$: transaction i writes object x
$r_i(x)$: transaction i reads object x
c_i : transaction i commits
a_i : transaction i aborts

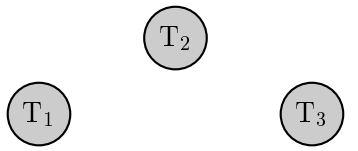
[4 P.] i) $\mathbf{S_1 = r_1(x)\ r_2(x)\ w_1(x)\ r_2(y)\ w_2(y)\ r_3(v)\ w_3(v)\ r_2(v)\ w_2(v)\ c_1\ c_2\ c_3}$



[4 P.] ii) $\mathbf{S_2 = r_1(x)\ w_1(x)\ r_2(x)\ c_2\ r_3(x)\ r_3(y)\ r_1(y)\ w_1(y)\ c_1\ c_3}$



iii) $\mathbf{S_3 = r_1(x)\ w_1(x)\ r_2(x)\ c_2\ r_3(y)\ w_3(y)\ r_1(y)\ w_1(y)\ c_3\ c_1}$ [4 P.]



b) Give short answers to the following questions.

i) Is schedule S_3 order-preserving conflict serializable (OCSR)? Briefly justify your answer. [2 P.]

ii) Is schedule S_1 commit order-preserving conflict serializable (COCSR)? Briefly justify your answer. [2 P.]

iii) Given the schedule $\mathbf{S_4 = r_1(x)\ w_1(x)\ r_2(x)\ w_2(y)\ a_1\ c_2}$, indicate whether or not the schedule can produce canonical synchronization problems (i. e. Dirty-Read, Lost-Update, Non-repeatable Read, Phantom-Problem). Which of them? [2 P.]

Part 2: Concurrency Control: Synchronization

[18 P.]

a) Consider a database with the following schema:

Book(ISBN, Title, Language, Pages)
Seller(SID, Name, Address)
Inventory(Book → Book.ISBN, Seller → Seller.SID, Price, Quantity)

Assume further that table Inventory contains the following data:

Inventory			
Book	Seller	Price	Quantity
53	2	9.33	1
87	1	45.35	1
48	4	45.99	1
42	4	200.00	2
12	9	39.05	1
92	9	28.22	4

Consider the transactions T_1 and T_2. T_1 always has the highest ANSI isolation level **Serializable (level 3)** (DB2's Repeatable Read).

[4 P.] i) Suppose T_2 uses **Read Committed (level 1)** (DB2's Cursor Stability) on the following SQL statements. What is the price of the book 42? Briefly explain whether or not a canonical synchronization problem (i. e. Dirty-Read, Lost-Update, Non-repeatable Read, Phantom-Problem) occurs and if so, name the problem and describe which ANSI isolation level must be set at least, to prevent it.

T_1	T_2
	SELECT Price **FROM** Inventory **WHERE** Book = 42;

UPDATE Inventory
  SET Price = Price * 0.50
  WHERE Book = 42;
 	
COMMIT	

	UPDATE Inventory
	  SET Price = Price * 0.80
	  WHERE Book = 42;
	
	COMMIT

ii) Given the following SQL statements of T_1 and T_2. T_2 returns 2 for the count of inventory entries and 7 for the sum of the quantity. Briefly explain whether a canonical synchronization problem occurred and if so, name the problem and say which is the highest possible ANSI isolation level used by T_2 and which ANSI isolation level have to be set at least to prevent it. [4 P.]

T_1	T_2

	`SELECT COUNT(*)`
	`FROM Inventory`
	`WHERE Seller = 9;`

`INSERT INTO Inventory` `VALUES(54, 9, 123.33, 2);`	
`COMMIT`	

	`SELECT SUM(Quantity)`
	`FROM Inventory`
	`WHERE Seller = 9;`

	`COMMIT`

[4 P.]

iii) Given the following SQL statements of T_1 and T_2. If T_2 encounters a Dirty Read problem, what will be the result set of the select statement. Briefly explain which ANSI isolation level has been set for T_2 and which one has to be set at least to prevent the Dirty Read.

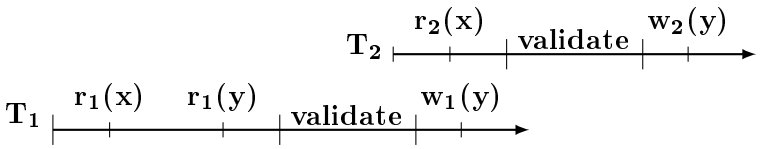
T_1	T_2
UPDATE Inventory **SET** Price = Price * 1.20;	
	SELECT Book, Seller **FROM** Inventory **WHERE** Price >= 50;
	COMMIT
ROLLBACK	

b) Give short answers to the following questions.

i) Which locks are compatible with the RIX lock? [1 P.]

ii) Which variant of 2PL prevents deadlocks altogether? Briefly explain why. [2 P.]

iii) Consider optimistic concurrency control as applied to the following transactions T_1 and T_2: [3 P.]



Explain the outcome of **BOCC** for T_1 and T_2.

Part 3: Logging and Recovery

[12 P.]

Consider a database using a no-force/steal/non-atomic recovery procedure with fuzzy checkpoints as described in the lecture. The table below shows the log on disk for a period of time during which 2 checkpoints were taken. The system crashes just after the last log record has been written.
At the time of Checkpoint 1, there were no running transactions and no dirty pages in the system. No pages have been written out to disk after Checkpoint 1.

LSN	Transaction	Type	Page	
– Checkpoint 1 –				
11	T1	BOT		
12	T1	UPDATE	P1	
13	T2	BOT		
14	T3	BOT		
15	T2	UPDATE	P5	
16	T2	COMMIT		
17	T3	UPDATE	P3	
– Checkpoint 2 –				
18	T3	UPDATE	P1	
19	T3	COMMIT		

[1 P.] a) What are the winner transactions?

[1 P.] b) What are the loser transactions?

[2 P.] c) At what LSN does the analysis phase begin? Briefly justify your answer.

[2 P.] d) At what LSN does the redo phase begin? Briefly justify your answer.

e) Please list all log records that are undone during undo phase in the order that the system undoes them. Briefly justify your answer. [2 P.]

f) What information does the system store for Checkpoint 2? [4 P.]

Part 4: Distributed Transactions & NoSQL

[12 P.]

a) In a distributed DB, there are three database servers A, B, C. The three servers co-operate in a distributed transaction using the **Two-Phase Commit (2PC)** protocol. A is the coordinator.

[2 P.] i) Suppose that A has sent all the prepare messages but has not yet received a prepared (or abort) message from server B. Would it be correct for A to commit the transaction at this point? Why or why not?

[2 P.] ii) Same situation, would it be correct for A to abort the transaction at this point, sending abort messages to the servers and an failed message to the client? Why or why not?

[2 P.] iii) Suppose that server B has received the prepare, sent the ready message, but has not yet received a commit or abort message. Server B contacts server C and discovers that server C has received a commit message. Would it be correct for server B to commit its part of the transaction at this point? Why or why not?

iv) Suppose again that server B has received the prepare, contacts server C and discovers that server C has received the prepare, sent the ready message, but has not yet received a commit or abort, too. Would it be correct for Server B to commit its part of the transaction at this point? Why or why not? [2 P.]

b) In NoSQL systems, the BASE paradigm is usually used instead of the ACID paradigm.

i) The E in BASE stands for eventually consistent. What does that mean? [2 P.]

ii) Describe an example application scenario that produces inconsistent behavior due to the lack of causal consistency. [2 P.]

Part 5: Data Warehouse

[15 P.]

a) Consider the following fact table:

Sales

product	region	customer	year	#sales	profit
TV	North	Doe	2000	1	500
Radio	South	Smith	2000	2	100
Radio	East	Bush	1999	1	50
TV	East	Bush	2000	2	1000
Radio	South	Doe	1999	3	150
TV	North	Smith	1999	1	500

[1 P.] i) What is the primary key of this table?

[1 P.] ii) Is this fact table cumulative or is it a snapshot table?

[6 P.] iii) Consider the following SQL query:

```
SELECT    product, region, SUM(#sales) as totalSales
FROM      Sales
GROUP BY  GROUPING SETS ((product),(region),());
```

Fill in the result of this query in the empty table below.

product	region	totalSales

b) Name three of the four strategies that can be used to extract data from a source into a Data Warehouse. Briefly describe each of these three strategies with a short sentence. [3 P.]

c) While an operational database is called application-oriented, a Data Warehouse is considered to be subject-oriented. Briefly describe the difference between these two properties. [2 P.]

d) Determine which properties MOLAP and ROLAP do have by marking the corresponding boxes with a cross (note, each property belongs to exactly one of both concepts). [2 P.]

	MOLAP	ROLAP
fast read access	☐	☐
low storage requirements	☐	☐
many joins	☐	☐
supports galaxy schema	☐	☐

Part 6: Data Mining

[18 P.]

[6 P.] a) Consider the following eight one-dimensional data objects:

$$9, 12, 18, 21, 28, 31, 35, 38$$

Cluster these objects by using the canopy clustering algorithm where

- the two thresholds are set to $T_1 = 8$ and $T_2 = 4$, and
- in each iteration step the smallest object in the candidate list is selected as new centroid.

Describe the clustering algorithm by naming the new centroid, the computed canopy as well as the objects which are removed from the candidate list for each iteration step (see layout shown below).

1. Iteration Step:

- Centroid:
- Canopy: { }
- Objects removed from the candidate list: { }

b) Consider the following five market basket transactions each of which contains some of the six items $\{A, B, C, D, E, F\}$: [12 P.]

k	T_k
001	$\{A, B, D\}$
002	$\{B, C, E\}$
003	$\{B, E\}$
004	$\{A, B, C, D, E\}$
005	$\{A, B, D, F\}$

i) Apply the Apriori algorithm from the lecture to compute all frequent itemsets while using the minimal support threshold $s_{cut} = 0.5$. [8 P.]

ii) Name two association rules that can be derived from the itemset $\{A, C, E\}$. [2 P.]

iii) Compute support and confidence of the association rule $r\colon \{A, B\} \rightarrow \{D\}$. [2 P.]

Part 7: Uncertain Databases

[17 P.]

a) Consider the following possible worlds representation of a relational database table called *Person*:

W_1, Pr(W_1)=0.6

	WK	name
t_1	p1	Alice
t_3	p2	Bob
t_4	p3	Charly

W_2, Pr(W_2)=0.3

	WK	name
t_2	p1	Ally
t_3	p2	Bob

W_3, Pr(W_3)=0.1

	WK	name
t_1	p1	Alice
t_5	p3	Charles

[2 P.] i) In which way do the two tuples t_3 and t_4 depend on each other? Briefly justify your answer.

[2 P.] ii) What is the marginal probability of tuple t_1? Briefly justify your answer.

[2 P.] iii) Can this possible worlds representation be modeled with a BID-database? Briefly justify your answer.

iv) Model this possible worlds representation by using a pc-database. Tip: Limit the world-table to one variable. [5 P.]

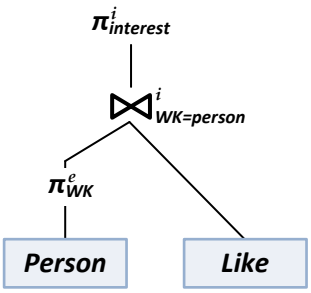
b) Consider the following BID-database and extensional query plan:

Person

	RK	WK	name	p
t_1	1	p1	Trump	0.8
t_2	2	p1	The Donald	0.2
t_3	3	p2	Merkel	1.0
t_4	4	p3	Schulz	0.4
t_5	5	p3	Scholz	0.4

Like

	RK	person	interest	p
t_6	1	p1	golf	1.0
t_7	2	p2	science	0.7
t_8	3	p3	birthdays	0.8
t_9	4	p2	rhombi	1.0
t_{10}	5	p1	science	0.2



Evaluate this query plan operator by operator and fill in the probabilities of the resultant tuples in the tables given below.

[1.5 P.] i) Subquery $Q_1 = \pi^e_{WK}(Person)$

	WK	p
t_{11}	p1	
t_{12}	p2	
t_{13}	p3	

[2.5 P.] ii) Subquery $Q_2 = Q_1 \bowtie^i_{WK=Person} Like$

	WK	RK	person	interest	p
t_{14}	p1	1	p1	golf	
t_{15}	p1	2	p1	science	
t_{16}	p2	3	p2	rhombi	
t_{17}	p2	4	p2	science	
t_{18}	p3	5	p3	birthdays	

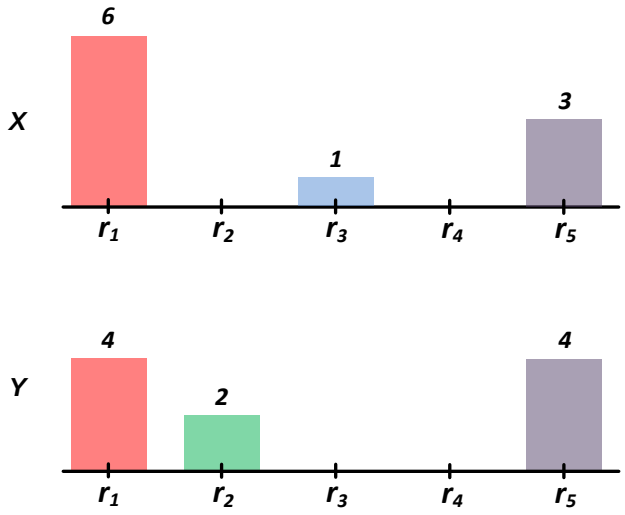
[2 P.] iii) Subquery $Q_3 = \pi^i_{interest}(Q_2)$

	interest	p
t_{19}	golf	
t_{20}	science	
t_{21}	rhombi	
t_{22}	birthdays	

Part 8: Similarity Search of Multimedia Objects

[10 P.]

a) Consider the following two feature signatures X and Y:



[3 P.] i) Compute the Earth Mover's Distance between X and Y by using the ground distance function $\delta(r_i, r_j) = |i - j|$.

$EMD_\delta(X, Y) =$

[3 P.] ii) Compute the Independent Minimization Lower Bound from X to Y by using the ground distance function $\delta(r_i, r_j) = |i - j|$.

$LB_{IM}(X, Y) =$

[2 P.] b) Why does storing feature histograms require less memory than storing feature signatures?

c) When does a distance function $\delta\colon \mathbb{X} \times \mathbb{X} \to \mathbb{R}^{\geq 0}$ satisfy the triangle inequality? [2 P.]
Complete the formal definition below.

Distance function δ satisfies the triangle inequality, iff

$\forall x, y, z \in \mathbb{X}$: