Informationssysteme und Datenanalyse

Test (International Version)

22.07.2017

Dies ist der Test der Lehrveranstaltung *Informationssysteme und Datenanalyse*. Bitte füllen Sie die Tabelle auf diesem Deckblatt aus und unterschreiben Sie den untenstehenden Hinweis.

Hinweise:

- Die Bearbeitungszeit für diesen Test beträgt 60 Minuten plus 10 Minuten Einlesezeit. Es können in 7 Fragen insgesamt 50 Punkte erreicht werden.
- Wenn Sie mehr als den zur Bearbeitung einer Aufgabe vorgesehenen Platz benötigen, können Sie ihre Antwort auf einer der freien Seiten fortsetzen. Machen Sie eine Weiterführung ihrer Antwort eindeutig kenntlich.
- Dieser Test besteht aus **16** Seiten. Bitte zählen Sie die Vollständigkeit der Seiten direkt nach Beginn der Einlesezeit.
- Bitte schreiben Sie außerdem direkt nach Beginn der Schreibzeit ihren Namen und ihre Matrikelnummer auf jede Seite.
- Die Verwendung von eigenem Papier ist **nicht** erlaubt. Zusätzliche leere Blätter werden auf Nachfrage ausgeteilt.
- Auf Ihrem Platz dürfen sich lediglich mehrere *dokumentenechte* Stifte sowie ihr Personal- und Studierendenausweis befinden. Einträge mit roten oder grünen Stiften sowie Füller und/oder Bleistift werden nicht gewertet. Weitere Hilfsmittel sind nicht zugelassen. Sämtliche elektronischen Geräte müssen sich ausgeschaltet in Ihrer Tasche befinden. Diese müssen Sie in der Reihe vor Ihnen oder anderweitig entfernt von Ihrem Platz abstellen.
- Klingelnde elektronische Geräte (Smartphones, Smartwatches o.Ä.) gelten als Täuschungsversuch.

Matrikelnummer	
Nachname(n)	
Vorname(n)	
Studiengang	

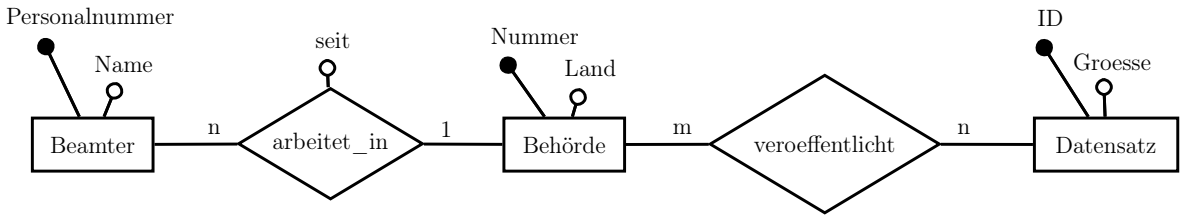
Hiermit bestätige ich, dass ich die oben genannte Hinweise verstanden haben und mich in der Lage fühle, diesen Test durchzuführen.

Unterschrift:

Aufgabe	Punkte	Erreicht	Korrektor
Datenbankentwurf	6		
Relationaler Entwurf	6		
Anfragesprachen	12		
Data Streams Management	5		
Data Warehousing	4		
Data Analysis	7		
Multiple Choice	10		
Summe	50		

Aufgabe 1: Datenbankentwurf (6 Punkte)

Gegeben Sei das folgende Entity-Relationship-Diagramm für eine *Open Government Data*-Implementierung.



1.1. Ergänzen Sie das obenstehende Entity-Relationship-Diagramm um die folgenden Angaben. Achten Sie dabei auch auf mögliche Integritätsbedingungen .

a) Ein Beamter kann einen Beamten als Vorgesetzten haben. Ein Beamter kann der Vorgesetzte für beliebig viele Beamte sein. (1)

b) Jeder Beamte arbeitet in einer Behörde. (0,5)

1.2. Gegeben seien außerdem die folgenden Relationen. Erweitern Sie das ER-Diagramm aus Aufgabe 1 durch Verwendung eines Abstraktionskonzeptes zu einem Erweiterten ER-Diagramm (EER-Diagramm), indem Sie die Informationen aus den gegebenen Relationen verwerten. Weitere Datentupel als die angegebenen existieren nicht. Achten Sie dabei darauf, dass Ihre Modellierung nicht kapazitätserhöhend oder kapazitätsvermindernd ist. (3)

Datensatz

ID	Groesse
1A5	1875294
3D7	45782
14G8	27364
7E10	152
18H34	64821
1F78	51724

Text

ID → Datensatz	Herkunft	Format
3D7	NASA	XML
14G8	ESA	ODF
7E10	DeStatis	PDF

Tabelle

ID → Datensatz	AnzSpalten
1A5	7
18H34	4

Bild

ID → Datensatz	{keywords}
1A5	{Earth, BlueDot}
1F78	{Apollo}

1.3. Sind die folgenden Integritätsbedingungen im ER-Entwurf abgebildet ?

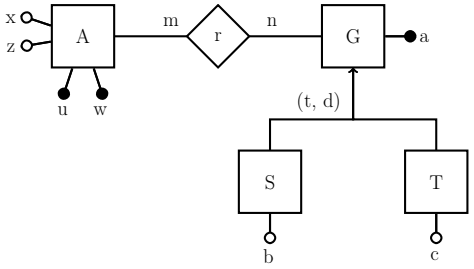
a) Eine Behörde muss Datensätze veröffentlichen. ◯ Ja ◯ Nein (0,5)

b) Ein Beamter kann in maximal einer Behörde arbeiten. ◯ Ja ◯ Nein (0,5)

c) Ein Datensatz darf eine maximale Größe von 25MB haben. ◯ Ja ◯ Nein (0,5)

Aufgabe 2: Relationaler Entwurf (6 Punkte)

Gegeben sei das folgende erweiterte Entity-Relationship-Modell.



2.1. Erstellen Sie zum oben gegebenen EER-Diagramm den Relationalen Entwurf , nicht jedoch für die Relation $A(\underline{u, w}, x, z)$. Benutzen Sie dabei den objektorientierten Stil zur Abbildung der Generalisierung/Spezialisierung . (4)

2.2. Weiterhin ist die Relation A mit $A(\underline{u, w}, x, z)$ sowie der funktionalen Abhängigkeiten $(u, w) \to x$, $(u, w) \to z$ sowie $z \to x$ gegeben. Weitere funktionalen Abhängigkeiten existieren nicht.

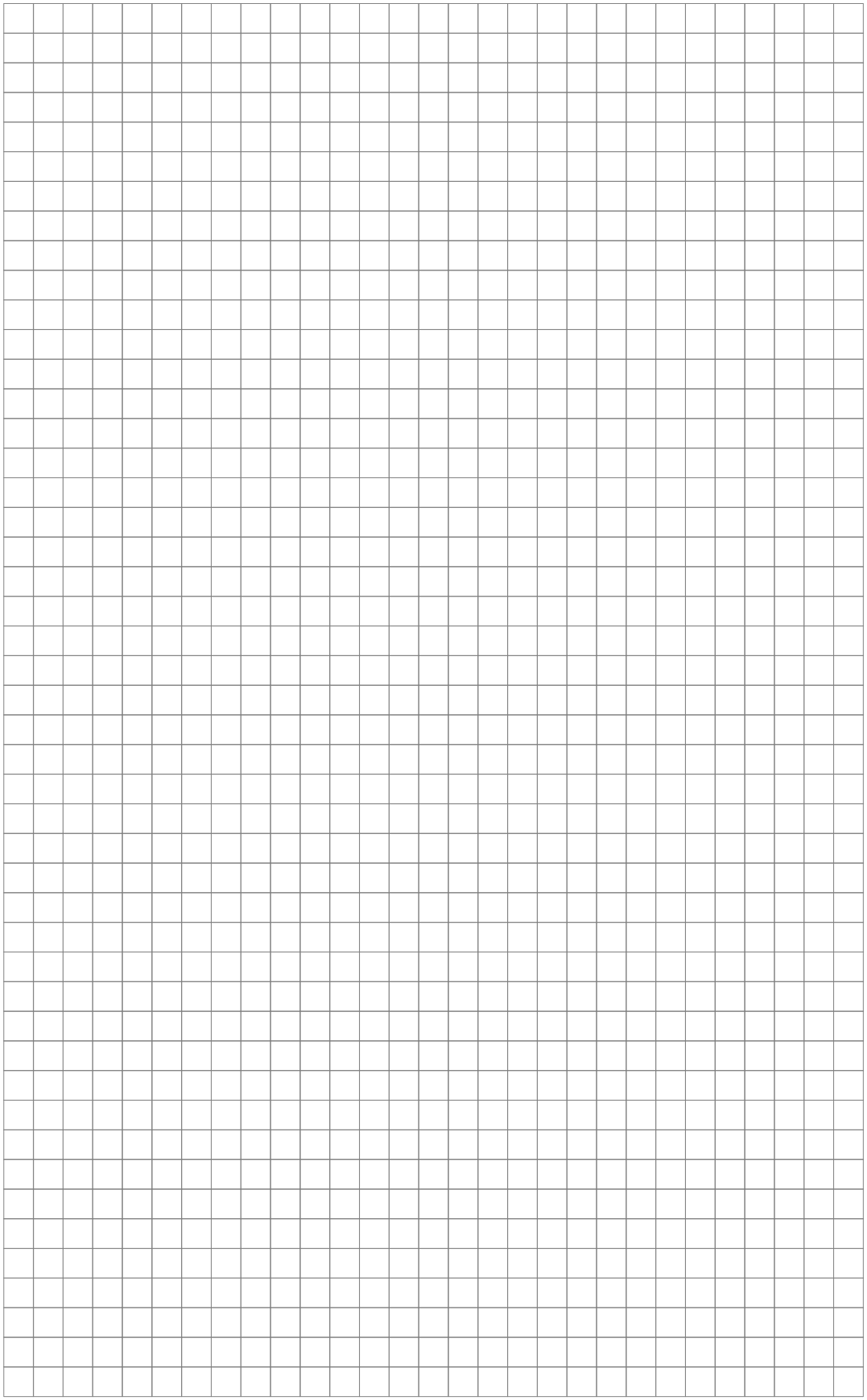
a) Begründen Sie, warum sich A in der ersten Normalform befindet. (0,5)

. . . . . .

. . . . . .

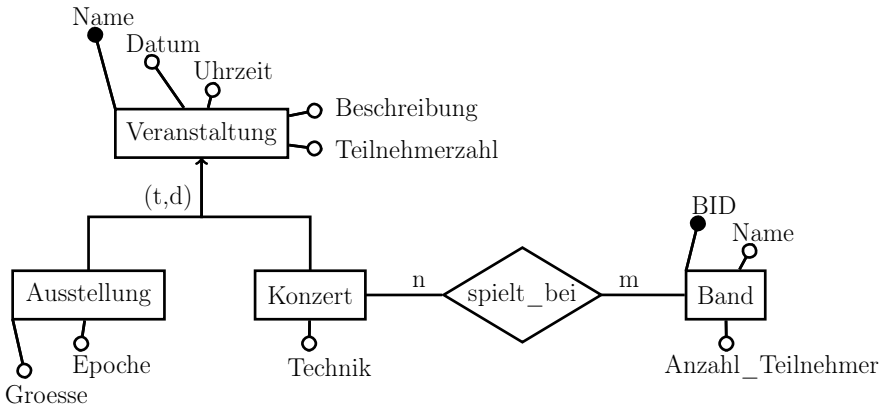
. . . . . .

b) Normalisieren Sie die Relation A bis zur dritten Normalform. Geben Sie für jeden Normalisierungsschritt die Zerlegung der Relation(en) sowie die möglicherweise Normalformverletzende funktionalen Abhängigkeit an. Unterstreichen Sie Schlüsselattribute . Sie können davon ausgehen, dass sich A bereits in der ersten Normalform befindet. (1,5)



Aufgabe 3: Anfragesprachen (12 Punkte)

Gegeben sei folgendes Datenbankschema eines Kulturzentrums, das bereits beispielhafte Tupel enthält.



Spielt_Bei

BID	VName
1	Open Flair Festival
2	Eurovision Songcontest
2	Open Flair Festival
3	Eurovision Songcontest
4	Musikantenstadl
4	Rammstein Live
5	Rammstein Live
2	Musikantenstadl
5	Open Flair Festival

Band

BID	Bandname	Anzahl_Musiker
1	Rise Against	4
2	Alligatoah	1
3	Von Wegen Lisbeth	5
4	Helene Fischer	3
5	Rammstein	6
6	Phil Collins	3

Ausstellung

VName	Epoche	Groesse
Sommerausstellung	Gegenwart	klein
Vernissage Berlin-Mitte	`NULL`	`NULL`

Konzert

VName	Technik
Open Flair Festival	LVX99 Bundle
Musikantenstadl	Soundmaster XL
Eurovision Songcontest	Stereoanlage ZZZ
Rammstein Live	Dosentelefon Nofeletnesod

Veranstaltung

VName	Datum	Uhrzeit	Teilnehmerzahl	Beschreibung
Open Flair Festival	2017-08-09	20:00:00	30000	Rockfestival
Musikantenstadl	2017-03-12	17:00:00	2500	BR-Abendprogramm
Eurovision Songcontest	2016-05-10	20:15:00	9999	Wettbewerb
Rammstein Live	2016-12-07	16:00:00	100	Tourneestart
Vernissage Berlin-Mitte	2017-07-20	08:00:00	42	Hipsterstuff
Sommerausstellung	2017-07-01	08:30:00	1337	Action Painting

3.1. Formulieren Sie sowohl eine gültige SQL-Query als auch eine äquivalente Anfrage in relationaler Algebra: *Eine duplikatfreie Liste der Bandnamen der Bands, die bei einem Konzert spielen und mehr als 3 Musiker haben.* (2)

SQL:

Relationale Algebra:

3.2. Formulieren Sie sowohl eine gültige SQL-Query als auch eine äquivalente Anfrage in relationaler Algebra: *Name der Veranstaltung(en) mit den meisten Teilnehmern (also der größten Teilnehmerzahl).* (3)

SQL:

Relationale Algebra:

3.3. Formulieren Sie sowohl eine gültige SQL-Query als auch eine äquivalente Anfrage in relationaler Algebra: *Eine Aufstellung aller Bandnamen und der Summe ihrer Konzertbesucher bei allen Konzerten.* (3)

SQL:

Relationale Algebra:

3.4. Geben Sie das Ergebnis (inklusive Spaltentitel) auf Basis der vorliegenden Daten zur folgenden Anfrage an. Formulieren Sie zusätzlich die Anfrage in natürlicher Sprache. (2)

SELECT v.Teilnehmerzahl
FROM AUSSTELLUNG a **NATURAL JOIN** Veranstaltung v
WHERE v.VName **LIKE** ´V%´;

Ergebnistabelle:

Natürlichsprachliche Formulierung:

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

3.5. Geben Sie das Ergebnis (inklusive Spaltentitel) auf Basis der vorliegenden Daten zur folgenden Anfrage an. Formulieren Sie zusätzlich die Anfrage in natürlicher Sprache. (2)

SELECT Bandname
FROM Band
WHERE BID **NOT IN** (**SELECT DISTINCT** BID **FROM** SPIELT_BEI);

Ergebnistabelle:

Natürlichsprachliche Formulierung:

.....

.....

.....

.....

.....

Aufgabe 4: Data Streams Management (5 Punkte)

4.1. Nennen Sie zwei typische Einsatzgebiete von Data Streams Management-Systemen und beschreiben Sie kurz in höchstens drei Sätzen, warum es in diesem Gebiet vorteilhaft sein kann, ein Data Streams Management-System zu verwenden. (2)

.....

.....

.....

.....

.....

.....

.....

.....

4.2. Gegeben sei die folgende Ergebnismatrix eines Count-Min-Sketch-Durchlaufs: (2)

$$\begin{array}{c} \\ h_0 \\ h_1 \\ h_2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 17 & 49 & 18 \\ 41 & 29 & 14 \\ 26 & 19 & 39 \end{bmatrix} \end{array}$$

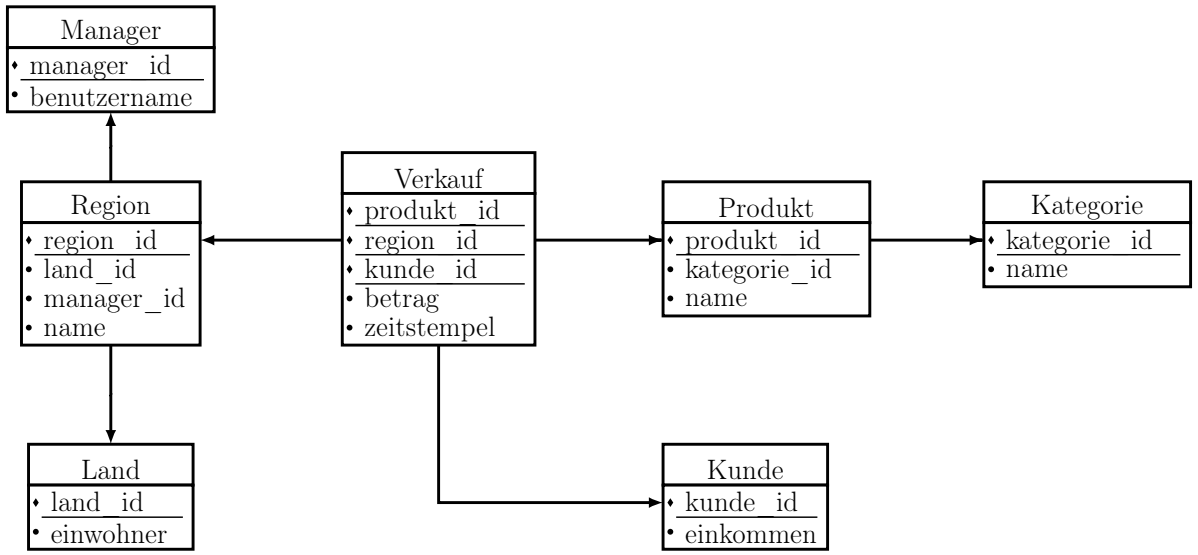
Geben Sie die Schätzung des Count-Min Sketches $\hat{f}(v)$ für die folgenden Werte v an:

v	$h_0(v)$	$h_1(v)$	$h_2(v)$	$\hat{f}(v)$
„Koala“	0	0	1	
„Colamat“	2	1	1	

4.3. Angenommen Sie verwenden einen Bloom-Filter in Ihrer Anwendung. Vorherige Berechnungen haben für eine Anzahl Bits $b = 512$ eine optimale Anzahl von Hash-Funktionen $h = 3$ ergeben. Messungen ergaben, dass Sie die Anzahl der auszuwertenden Hashfunktionen auf $h = 12$ erhöhen könnten, ohne die Anwendung zu beeinträchtigen. Geben Sie die optimale Anzahl Bits b für den Parameter an. (1)

Aufgabe 5: Data Warehousing (4 Punkte)

Gegeben sei das folgende relationale Diagramm eines OLAP-Würfels :



5.1. Markieren Sie im obenstehenden Diagramm die Fakten- sowie Dimensionstabellen mittels einer eindeutigen Beschriftung. (1)

5.2. In der Vorlesung wurden drei Darstellungen vorgestellt, um einen OLAP-Würfels auf ein relationales Schema abzubilden. Welcher der vorgestellten Darstellungen entspricht das oben gennante Schema? (0,5)

.....

5.3. Nennen Sie eine weitere Darstellung sowie die Anzahl der Relationen, die bei der Verwendung der von Ihnen gewählten Darstellung aus dem obigen relationalen Schema entstehen. (1)

.....

5.4. Wofür steht das Akronym *ETL* im Kontext von Data-Warehouses (3 Begriffe)? (0,5)

.....

5.5. Zur Analyse von Textdaten in relationalen Datenbanksystemen müssen diese zunächst in ein relationales Modell überführt werden. Ist dieser Schritt auch *zwingend* für die Analyse in MapReduce-Systemen nötig? Begründen Sie Ihre Antwort in höchstens drei Sätzen. (1)

.....

.....

.....

.....

.....

.....

.....

Aufgabe 6: Data Analysis (7 Punkte)

6.1. Sortieren Sie die Euklidische , Manhattan- und Maximumdistanz zwischen zwei beliebigen Punkten aufsteigend von der garantiert kürzesten zur garantiert längsten Distanz. (1)

_____ $\leq$ _____ $\leq$ _____

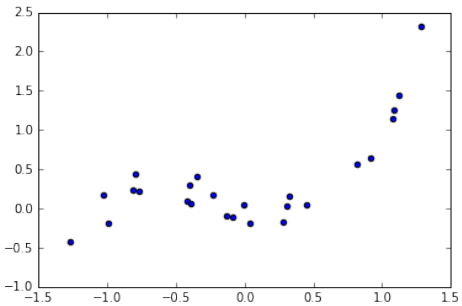
6.2. Gegeben seien die folgenden eindimensionalen Datenpunkte: $\{1, 1, 1, 4, 5, 6\}$ (3)

Führen Sie eine Iteration des k-Means-Algorithmus anhand dieses Beispiels durch und geben Sie die Clusterzentren an. Wählen Sie dazu die Punkte $c_1 = 1$ und $c_2 = 6$ als initiale Clusterzentren.

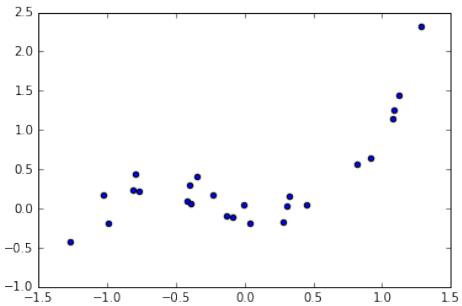
6.3. a) Auf welchen Datenklasse(n) kann man Regression anwenden? (0,5)

. . . . . .

b) Skizzieren Sie in Bild (a) eine Regressionsfunktion, die den Datensatz *underfitted* und in Bild (b) eine Regressionsfunktion, die den Datensatz *overfitted*. (1)



(a)



(b)

c) Nennen und beschreiben Sie eine Möglichkeit, wie man *Over- und Underfitting* beim Trainieren eines Modells feststellen kann. (1,5)

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Aufgabe 7: Multiple Choice (10 Punkte)

Diese Aufgabe umfasst 10 Multiple-Choice-Fragen. Sie bestehen aus jeweils drei Antwortmöglichkeiten, wobei mindestens eine Antwortmöglichkeit richtig und mindestens eine Antwortmöglichkeit falsch ist. Jede Frage, in der alle richtigen Antwortmöglichkeiten angekreuzt und alle falschen Antwortmöglichkeiten nicht angekreuzt sind, wird mit genau einem Punkt bewertet. Sollten nicht alle richtigen Antwortmöglichkeiten angekreuzt worden sein oder wurde mindestens eine falsche Antwortmöglichkeit angekreuzt, wird die Frage mit 0 Punkten bewertet.

7.1. Kreuzen Sie die richtigen Aussagen über Entity-Relationship-Diagramme an. (1)

- ☐ Zwei Entitytypen können nicht ohne einen Relationshiptypen miteinander verbunden werden.
- ☐ Ein Relationshiptyp ist immer mit genau zwei Entitytypen verbunden.
- ☐ Eine Entity eines schwachen Entitytyps kann nur in Verbindung mit einer Entity eines nicht-schwachen Entitytyps existieren.

7.2. Was bezeichnet man als Relation? (1)

- ☐ Eine geordnete Menge von Attributen und ungeordnete Menge von Tupeln.
- ☐ Eine geordnete Menge von Tupeln und ungeordnete Menge von Attributen.
- ☐ Eine geordnete Menge von Attributen und geordnete Menge von Tupeln.

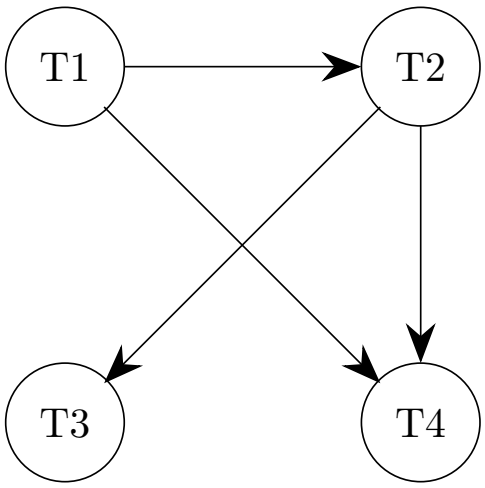
7.3. Welche/r Superschlüssel ist/sind für $R(A, B, C, D)$ nach folgenden funktionalen Abhängigkeiten möglich? $A \to C$, $C \to A$, $(A, B) \to D$, $C \to B$, $D \to B$ (1)

- ☐ C
- ☐ (A, D)
- ☐ B

7.4. Gegeben sei eine Relation R und der Ausdruck $(\sigma_c(R))$, wobei c ein beliebiges Prädikat für die Selektion ist. Geben Sie alle korrekten Aussagen über die mögliche Kardinalität des Ausdrucksergebnisses an. (1)

- ☐ Die Kardinalität ist höchstens $|R|$.
- ☐ Die Kardinalität ist immer $\frac{|R|}{2}$.
- ☐ Auch für nicht-leere R kann die Kardinalität 0 sein.

7.5. Gegeben sei der folgende Konfliktgraph für einen Schedule S. Kreuzen Sie die korrekten Aussagen an. (1)



- ☐ Der Graph enthält keinen Zyklus , ein äquivalenter serieller Schedule hat die Reihenfolge $T1 \to T2 \to T4 \to T3$.
- ☐ Der Graph enthält keinen Zyklus, ein äquivalenter serieller Schedule hat die Reihenfolge $T1 \to T2 \to T3 \to T4$.
- ☐ Der Graph enthält einen Zyklus und ist somit nicht konfliktserialisierbar .

7.6. Wofür steht das Akronym *ACID* im Kontext von Datenbanksystemen? (1)

- ☐ Atomicity, Consistency, Isolation, Durability
- ☐ Afri Cola Is Delicious
- ☐ Alter Consistent Irrational Databases

7.7. Was gilt im Allgemeinen für eine Hashfunktion $f : K \to S$? (1)

- ☐ $|K| \leq |S|$
- ☐ Jedes Element aus K wird auf ein Element aus S abgebildet.
- ☐ $|S|$ ist immer genau 42.

7.8. Welches Verhältnis gilt im Allgemeinen zwischen der Schätzung des Count-Min-Sketches $\hat{f}(v)$ und der tatsächlichen Häufigkeit $f(v)$? (1)

- ☐ $\hat{f}(v) \leq f(v)$
- ☐ $\hat{f}(v) = f(v)$
- ☐ $\hat{f}(v) \geq f(v)$

7.9. Kreuzen Sie die richtigen der folgenden Aussagen über XML und XPath an. (1)

- ☐ Jeder XML-Knoten (self) kann maximal einen *ancestor* und beliebig viele *descendants* haben.
- ☐ Ein XPath-Lokalisierungsschritt folgt der Syntax `node-test::axis[predicate 1]`.
- ☐ XML-Dokumente enthalten sowohl beschreibende Metadaten als auch Daten selbst.

7.10. Die Funktionalität welcher SQL-Operatoren kann die Reduce-Funktion eines MapReduce-Durchlaufs übernehmen? (1)

- ☐ WHERE
- ☐ GROUP BY
- ☐ ORDER BY