+1/1/60+

Informationssysteme und Datenanalyse

Schriftlicher Test 2, 18.07.2020

Woman Muster, 123456, Raum CH I, Platz 1

Hinweise:

- Die Bearbeitungszeit für diesen Test beträgt 60 Minuten. Darin enthalten sind 10 Minuten, um Ihre Antworten auf den separaten Antwortbogen zu übertragen.
- Es werden ausschließlich Antworten auf dem **separaten Antwortbogen** gewertet. Zudem werden ausschließlich **vollständig ausgefüllte** Boxen gewertet. Kreuze oder sonstige Markierungen werden nicht gewertet.
- Dieser Test besteht aus insgesamt **13** Seiten: 12 Seiten Aufgabenstellung und eine separate Seite Antwortbogen. Bitte überprüfen Sie die Vollständigkeit der Seiten direkt nach Beginn der Bearbeitungszeit.
- Es können in 5 Themenbereichen insgesamt 40 Punkte erreicht werden.
- Dieser Test beinhaltet **zwei Fragetypen**. Bei Fragen von Typ 1 ist genau **eine Antwortmöglichkeit** korrekt. Bei Fragen von Typ 2 sind entweder **eine oder mehrere Antwortmöglichkeiten** korrekt. Fragen von Typ 2 sind mit dem Symbol ♣ markiert.
- Bei Fragen von Typ 2 vergeben wir Teilpunkte, wenn Sie einen Teil der richtigen Antwortmöglichkeiten ankreuzen. Wenn Sie eine oder mehrere falsche Antwortmöglichkeit ankreuzen, erhalten Sie 0 Punkte für die Frage.
- Die Verwendung von eigenem Papier ist **nicht** erlaubt. Sie finden leere Blätter auf Ihrem Platz.
- 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 entfernt von Ihrem Platz abstellen.
- Klingelnde elektronische Geräte (Smartphones, Smartwatches o.Ä.) gelten als Täuschungsversuch.

Themenbereich	**Punkte**
Transaktionen	8
Data Warehousing	8
Explorative Datenanalyse	9
Prädiktive Datenanalyse	7
Datenstrommanagement	8
Insgesamt	**40**

Transaktionen

Frage 1 ♣ (1 Punkt) Welche der folgenden Aussagen treffen auf ACID-Transaktionen zu?

- [A] Atomarität bedeutet, dass alle Änderungen einer abgebrochenen Transaktion immer rückgängig gemacht werden.
- [B] Konsistenz bedeutet, dass die Datenbank gespeicherte Daten nie willkürlich löscht.
- [C] Wenn genau eine Nutzerin Transaktionen ausführt, benötigen Transaktionen keine Isolation.
- [D] Eine Transaktion überführt die Datenbank immer von einem konsistenen Zustand in einen anderen konsistenten Zustand.
- [E] Alle SQL-Isolations-Level garantieren die vollständigen ACID-Eigenschaften.
- [F] Atomarität und Isolation implizieren direkt die Konsistenz einer Transaktion.
- [G] *Keine dieser Antworten ist korrekt.*

Frage 2 ♣ (3 Punkte) Gegeben ist eine Datenbank mit Studierendendaten:

Studierende	Name	Note
	Ines Becker	1.3
	Marcel Wechsler	4.0
	Kathrin Ackermann	3.7

Auf dieser Datenbank werden die folgenden Transaktionen zeitgleich ausgeführt.

```
START TRANSACTION ISOLATION LEVEL REPEATABLE READ;
INSERT INTO Studierende VALUES ('Michael Pfeifer', 1.0);
COMMIT;

START TRANSACTION ISOLATION LEVEL REPEATABLE READ;
SELECT SUM(Note) / COUNT(Note) AS Durchschnitt FROM Studierende;
COMMIT;
```

Wählen Sie alle Ergebnisse, welche die Datenbank als `Durchschnitt` ausgeben kann.

- [A] 2.0
- [B] 2.25
- [C] 2.5
- [D] 3.0
- [E] 3.33
- [F] *Keine dieser Antworten ist korrekt.*

Frage 3 ♣ (1 Punkt) Wählen Sie alle korrekten Aussagen über das angegebene Schedule:

T_1	T_2
$sl_1(A)$	
	$sl_2(B)$
$read_1(A)$	
	$write_2(B)$
	$sl_2(A)$
	$read_2(A)$
$ul_1(A)$	
	$ul_2(B)$

A Das Schedule ist sperr-konsistent.
B Das Schedule ist legal.
C Das Schedule ist lese-konsistent.
D *Keine dieser Antworten ist korrekt.*

Frage 4 ♣ (2 Punkte) Wählen Sie alle korrekten Aussagen über das folgende Schedule:

$$R_1(A), R_1(B), W_2(A), R_1(A), R_2(B)$$

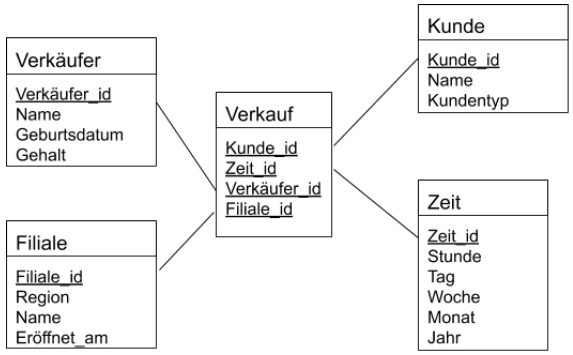
A Das Schedule ist weder seriell noch konfliktserialisierbar.
B Das Schedule ist konfliktäquivalent zu: $R_1(A), W_2(A), R_2(B), R_1(B), R_1(A)$
C Das Schedule ist seriell.
D Das Schedule ist konfliktserialisierbar.
E *Keine dieser Antworten ist korrekt.*

Frage 5 ♣ (1 Punkt) Wählen Sie alle Anomalien, die für das SQL Isolations-Level `REPEATABLE READ` möglich sind.

A Phantom Read
B Non-repeatable Read
C Dirty Read
D *Keine dieser Antworten ist korrekt.*

Data Warehousing

Frage 6 (1 Punkt) Ein Verkauf wird als Fakt mit den Dimensionen Zeit, Kunde, Verkäufer, und Filiale betrachtet. Die Dimension Zeit ist hierarchisch organisiert: Stunde, Tag, Woche, Monat, Jahr. Das Schema wurde wiefolgt entworfen:



Es handelt sich um ein:

A Schneeflockenschema B Sternschema C Fullfactschema

Frage 7 ♣ (2 Punkte) Gegeben ist eine Faktenrelation `Bestellung(Bestellung_Id, Zeit_Id, Preis)`, eine Dimensionsrelation `Zeit(Zeit_Id, Tag, Monat, Jahr)` (die hierarchisch nach Tag, Monat und Jahr organisiert ist), sowie folgende Anfrage:

```
SELECT Jahr, Monat, Tag, SUM(Preis) AS Teilsumme
  FROM Bestellung, Zeit
  WHERE Bestellung.Zeit_Id = Zeit.Zeit_Id
  GROUP BY Jahr, Monat, Tag
UNION ALL
SELECT Jahr, Monat, "ALL", SUM(Preis) AS Teilsumme
  FROM Bestellung, Zeit
  WHERE Bestellung.Zeit_Id = Zeit.Zeit_Id
  GROUP BY Jahr, Monat
UNION ALL
SELECT Jahr, "ALL", "ALL", SUM(Preis) AS Teilsumme
  FROM Bestellung, Zeit
  WHERE Bestellung.Zeit_Id = Zeit.Zeit_Id
  GROUP BY Jahr
```

Hinweis: `Bestellung(Zeit_Id)` ist ein Fremdschlüssel für `Zeit(Zeit_Id)`.

Welche der folgenden Aussagen über diese Anfrage sind korrekt?

A Die Anfrage führt eine Dice-Operation aus.

B Die Gesamtsumme aller Teilsummen des Ergebnisses entspricht dem Ergebnis folgender Anfrage: `SELECT SUM(Preis) FROM Bestellung`

C Als Ergebnis liefert die Anfrage die Summe der Preise aller Bestellungen für jede Kombination der hierarchischen Ebenen der Dimension Zeit.

D Die Anfrage führt eine Roll-Up-Operation aus.

E *Keine dieser Antworten ist korrekt.*

Frage 8 ♣ (2 Punkte) Welche der folgenden Aussagen zu mehrdimensionaler Modellierung und OLAP sind korrekt?

A Dimensionen beschreiben messbare Daten.

B Fakten sind hierarchisch organisiert.

C Data Scientists können OLAP-Systeme nutzen, um Informationen aus Datensätzen zu extrahieren.

D Fokus der mehrdimensionalen Modellierung ist es, schnelle Aggregationen zu unterstützen.

E Ein System für Bankautomaten benutzt typischerweise ein OLAP-System, um Zahlungsdaten zu verwalten.

F *Keine dieser Antworten ist korrekt.*

Frage 9 ♣ (2 Punkte) Welche der folgenden Aussagen zu einem Data Warehouse sind korrekt?

A Ein Data Warehouse wird üblicherweise periodisch mit Daten aus mehreren Datenquellen gefüllt.

B Ein Data Warehouse bearbeitet hauptsächlich kurze, direkt aufeinanderfolgende Transaktionen.

C Ein Data Warehouse hilft einem Unternehmen Datenanalysen effizient durchzuführen.

D In einem Data Warehouse werden historische Daten mit frischen Daten ersetzt.

E *Keine dieser Antworten ist korrekt.*

Frage 10 ♣ (1 Punkt) Welche der folgenden Aussagen zu ETL im Kontext eines Data Warehouses sind korrekt?

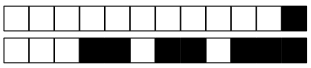
A Der Transformationsprozess definiert unter anderem, wie die Quelldaten gereinigt werden sollen, bevor diese in das Data Warehouse geladen werden.

B Das Schema der Quelldatenbanken (von OLTP-Systemen) und des Data Warehouses unterscheiden sich nicht.

C ETL steht für Extract, Transform, Load.

D Der Prozess *Extract* wandelt die Daten vom Schema der Quelldatenbanken in das Schema des Data Warehouses um.

E *Keine dieser Antworten ist korrekt.*



Explorative Datenanalyse

Frage 11 ♣ (3 Punkte) Gegeben ist ein zweidimensionales Datenset D:

$$D = \begin{matrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{matrix} \begin{bmatrix} 1 & 0 \\ 3 & 2 \\ 0 & 6 \\ 1 & 8 \\ 2 & 9 \end{bmatrix}$$

Führen Sie hierarchisches Clustering mit euklidischer Distanz und single linkage durch, sodass Sie 3 Cluster erhalten. Welche der folgenden Cluster sind Teil des sich ergebenden Clusterings?

Hinweis: Wählen Sie eine Antwortmöglichkeit nur, wenn der Cluster der Antwortmöglichkeit exakt mit einem der Ergebniscluster übereinstimmt.

Hinweis: Diese Frage kann beantwortet werden, ohne die Distanzmatrix vollständig zu berechnen.

A $\{ d_1 \}$	E $\{ d_1, d_4, d_5 \}$	I $\{ d_2, d_4 \}$	M $\{ d_3, d_4, d_5 \}$	Q *Keine dieser Antworten ist korrekt.*
B $\{ d_1, d_2 \}$	F $\{ d_2 \}$	J $\{ d_3 \}$	N $\{ d_4 \}$	
C $\{ d_1, d_2, d_3 \}$	G $\{ d_2, d_3 \}$	K $\{ d_3, d_4 \}$	O $\{ d_4, d_5 \}$	
D $\{ d_1, d_4 \}$	H $\{ d_2, d_3, d_4 \}$	L $\{ d_3, d_5 \}$	P $\{ d_5 \}$	

Frage 12 (2 Punkte) Gegeben ist ein eindimensionales Datenset D und 2 initiale Centroids C_1 und C_2:

$$D = \begin{bmatrix} 1 \\ 11 \\ 5 \\ 13 \\ 15 \end{bmatrix} \qquad C_1 = \begin{bmatrix} 1 \end{bmatrix} \qquad C_2 = \begin{bmatrix} 2 \end{bmatrix}$$

Führen Sie zwei Iterationen von K-Means durch. Was ist die Distanz zwischen den zwei Centroids nach diesen Iterationen?

A 3	C 4	E 5	G 6	I 7	K 8	M 9	O 10
B 3.5	D 4.5	F 5.5	H 6.5	J 7.5	L 8.5	N 9.5	P 10.5

Frage 13 ♣ (2 Punkte) Welche der folgenden Aussagen über K-Means sind korrekt?

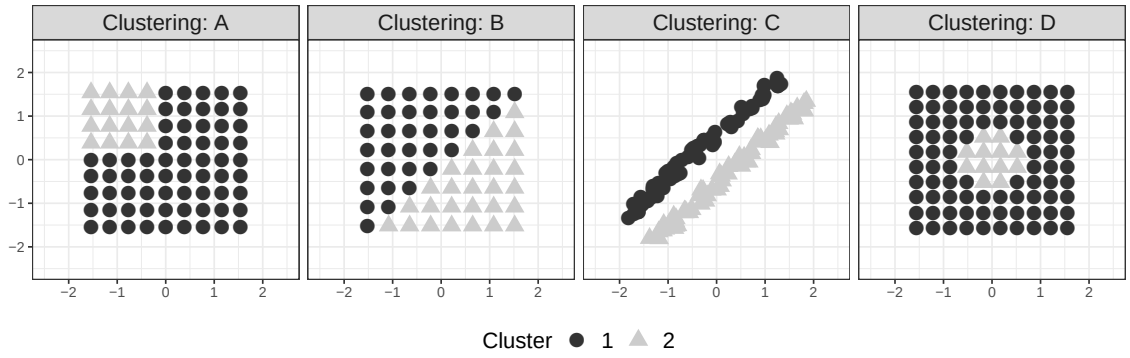
A Die Wahl der initialen Centroids kann Auswirkungen auf die Laufzeit haben.

B Die Wahl der initialen Centroids kann Auswirkungen auf das finale Clustering haben.

C k-Means hat eine eindeutige Lösung.

D *Keine dieser Antworten ist korrekt.*

Frage 14 ♣ (2 Punkte) Welche der folgenden Clusterings sind als Ergebnis einer konvergierten k-Means-Clusteranalyse **möglich**?



A Clustering A
B Clustering B
C Clustering C
D Clustering D
E *Keine dieser Antworten ist korrekt.*

Prädiktive Datenanalyse

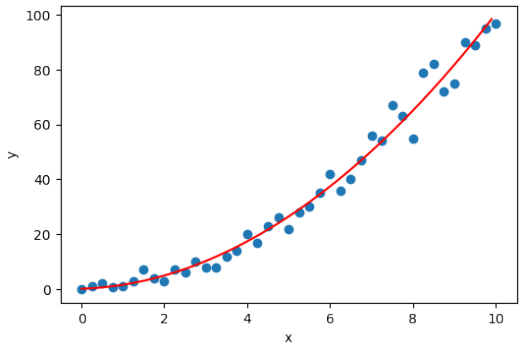
Frage 15 ♣ (2 Punkte) Welche der folgenden Techniken können verwendet werden, um Overfitting eines Modells zu **erkennen**?

- [A] Analysieren von Korrelationen
- [B] Kreuzvalidierung
- [C] Hinzufügen von polynomiellen Attributen
- [D] Zahl der Centroids erhöhen
- [E] Test-/Train-Datensplit
- [F] Lineare Regression
- [G] Polynomielle Regression
- [H] *Keine dieser Antworten ist korrekt.*

Frage 16 ♣ (1 Punkt) Welche der folgenden Verfahren werden zur Klassifikation eingesetzt?

- [A] DBSCAN
- [B] k-Means
- [C] Lineare Regression
- [D] k-Nearest-Neighbour
- [E] *Keine dieser Antworten ist korrekt.*

Frage 17 (2 Punkte) Welche der Gleichungen beschreibt das Modell, das in folgendem Plot durch eine Linie dargestellt wird?



- [A] $y = 0.12 + 0.5x + 0.95x^2$
- [B] $y = 0.05 + 9x + 0.1x^2$
- [C] $y = -0.05 + 9x$
- [D] $y = 0.34 + 4x$
- [E] $y = -4.53 + 0.2x + 0.85x^2$
- [F] $y = 12 - 4x$

Frage 18 (2 Punkte) Ein E-Mail-Anbieter nutzt ein Klassifikationsmodell, um Spam zu identifizieren. Unten sehen Sie für ein E-Mail-Datenset für jede E-Mail die Spam-Bewertung einer Nutzerin und des Spamfilters.

Betreff	Bewertung	
	Nutzerin	Spamfilter
Brandenburg Bilder	kein spam	kein spam
1M Dollar warten für Sie	spam	kein spam
2M Dollar warten für Sie	spam	kein spam
Ihr Fahrrad ist repariert	kein spam	kein spam
Happy27 voucher	kein spam	spam
jdjwrewiwerjlwjk	spam	spam
Unser Geschenk für Sie	spam	kein spam

Evaluieren Sie den Spamfilter anhand der Bewertungen der Nutzerin. Was ist der F1-Score dieses Spamfilters?

A 0.0	C 0.25	E 0.375	G 0.5	I 0.625	K 0.75	M 0.9
B 0.125	D 0.333	F 0.4	H 0.6	J 0.677	L 0.875	N 1.0

Datenstrommanagement

Frage 19 ♣ (1 Punkt) Was sind typische Merkmale eines Datenstrommanagementsystems?

- [A] SQL-Abfragen
- [B] Zeitbasierte Operationen
- [C] Graphverarbeitung
- [D] Kontinuierlich laufende Abfragen
- [E] *Keine dieser Antworten ist korrekt.*

Frage 20 ♣ (3 Punkte) Der Datenstrom (13, 14, 15) wird mit einem Bloomfilter mit 10 Bits und den Hashfunktionen $h_0(x)$ und $h_1(x)$ aufgezeichnet.

$$h_0(x) = ((x + 2) \bmod 15) \bmod 10$$
$$h_1(x) = ((2x) \bmod 12) \bmod 10$$

Welche der folgenden Aussagen über den Datenstrom sind anhand des Bloomfilters möglich?

- [A] 15 kommt mindestens einmal vor.
- [B] 15 kommt möglicherweise vor.
- [C] 15 kommt nicht vor.
- [D] 27 kommt mindestens einmal vor.
- [E] 27 kommt möglicherweise vor.
- [F] 27 kommt nicht vor.
- [G] 40 kommt mindestens einmal vor.
- [H] 40 kommt möglicherweise vor.
- [I] 40 kommt nicht vor.
- [J] *Keine dieser Antworten ist korrekt.*

Frage 21 ♣ (2 Punkte) Eine Firma verkauft vier Produkte: T-Shirts, Jeans, Jacken und Schuhe. Die Firma benutzt zum Aufzeichnen ihrer Verkaufszahlen einen Count-Min Sketch mit den Hashfunktionen h_0 und h_1:

	h_0	h_1
T-Shirts	0	1
Jeans	1	0
Jacken	1	1
Schuhe	0	0

Um die Genauigkeit des Sketches zu überprüfen, misst eine Mitarbeiterin die genauen Verkaufszahlen und vergleicht sie mit den Werten des Count-Min Sketches. Am Ende des Arbeitstages hat die Mitarbeiterin folgende reale Verkaufszahlen gemessen:

	Anzahl
T-Shirts	17
Jeans	9
Jacken	0
Schuhe	12

Welche Obergrenzen für die Verkaufszahlen ergeben sich aus dem erzeugten Count-Min Sketch?

- [A] 12 T-Shirts
- [B] 17 T-Shirts
- [C] 29 T-Shirts
- [D] 9 Jeans
- [E] 12 Jeans
- [F] 21 Jeans
- [G] 0 Jacken
- [H] 9 Jacken
- [I] 17 Jacken
- [J] 17 Schuhe
- [K] 21 Schuhe
- [L] 29 Schuhe
- [M] *Keine dieser Antworten ist korrekt.*

Frage 22 ♣ (2 Punkte) Gegeben sei der folgende Reservoir-Sampling-Algorithmus für ein Reservoir der Größe t. Der Algorithmus führt für das k-te Element eines Datenstroms folgende Schritte aus:

1. Erzeuge eine Zufallszahl p zwischen 1 und k (inklusive 1 und k).
2. Falls $p \leq t$: Schreibe Element k an Stelle p in das Reservoir.

Einem Reservoir der Größe $t = 10$ werden mit diesem Algorithmus nacheinander 100 Elemente eines Datenstroms hinzugefügt. Welche der folgenden Aussagen treffen zu?

Hinweis: Alle Nummerierungen in dieser Aufgabe beginnen bei 1.

A Das 5. Element wird dem Reservoir hinzugefügt.

B Das 11. Element wird dem Reservoir nicht hinzugefügt.

C Das 15. Element wird dem Reservoir mit einer Wahrscheinlichkeit von 90% hinzugefügt.

D Das 20. Element wird dem Reservoir mit einer Wahrscheinlichkeit von 50% hinzugefügt.

E Das 40. Element wird dem Reservoir mit einer Wahrscheinlichkeit von 50% hinzugefügt.

F *Keine dieser Antworten ist korrekt.*

+1/13/48+

Antwortbogen

Woman Muster
123456, Raum CH I, Platz 1

Es werden ausschließlich Antworten auf diesem Blatt und ausschließlich deutlich ausgefüllte Boxen gewertet (Beispiel: ■). Kreuze oder sonstige Markierungen werden nicht gewertet.

Transaktionen

Frage 1 : A B C D E F G
Frage 2 : A B C D E F
Frage 3 : A B C D
Frage 4 : A B C D E
Frage 5 : A B C D

Data Warehousing

Frage 6 : A B C
Frage 7 : A B C D E
Frage 8 : A B C D E F
Frage 9 : A B C D E
Frage 10 : A B C D E

Explorative Datenanalyse

Frage 11 : A B C D E F G H I J K L M N O P Q
Frage 12 : A B C D E F G H I J K L M N O P
Frage 13 : A B C D
Frage 14 : A B C D E

Prädiktive Datenanalyse

Frage 15 : A B C D E F G H
Frage 16 : A B C D E
Frage 17 : A B C D E F
Frage 18 : A B C D E F G H I J K L M N

Datenstrommanagement

Frage 19 : A B C D E
Frage 20 : A B C D E F G H I J
Frage 21 : A B C D E F G H I J K L M
Frage 22 : A B C D E F