Raum	
Platz	

Informationssysteme und Datenanalyse

Schriftlicher Test (International Version)

21.07.2018

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:

- You can answer any question in English or German.
- Die Bearbeitungszeit für diesen Test beträgt 60 Minuten plus 10 Minuten Einlesezeit. Es können in 7 Fragen insgesamt 50 Punkte erreicht werden. Während der Einlesezeit darf **nicht** geschrieben 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	

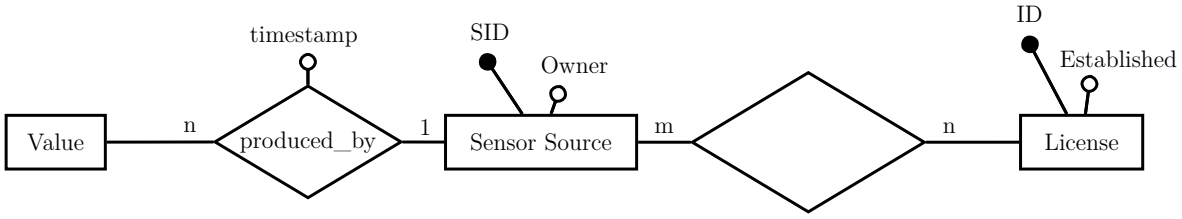
Hiermit bestätige ich, dass ich die oben genannten Hinweise verstanden haben und mich in der Lage fühle, diesen Test durchzuführen.

Unterschrift:

Aufgabe	Punkte	Erreicht	Korrektor
Datenbankentwurf	7		
Relationaler Entwurf	8		
Anfragesprachen	9		
XML	5,5		
Data Streams Management & DWH	5		
Data Analysis	5,5		
Multiple Choice	10		
Summe	50		

Aufgabe 1: Datenbankentwurf (7 Punkte)

Gegeben sei das folgende Entity-Relationship-Diagramm für das *opensense.network*, einer Sammlung von frei verfügbaren Sensordaten.



1.1. Vervollständigen Sie das gegebene ER-Diagramm zu einem syntaktisch korrekten (*syntactically correct*) ER-Diagramm. (1)

1.2. Ergänzen Sie das gegebene ER-Diagramm um die folgenden Angaben. Achten Sie dabei auch auf mögliche Integritätsbedingungen (*integrity constraints*).

a) Eine Sensordatenquelle besteht aus Teperatur- und Feinstaubsensoren. Feinstaub- und Temperatursensoren verfügen jeweils über eindeutige Gerätekennungen (TID, FID). (1,5)
(*A sensor source consists of temperature sensors and fine dust sensors. Temperature sensors and fine dust sensors are uniquely identified by seperate device IDs (TID, FID).*)

b) *Sämtliche* Werte müssen von einem Sensor produziert worden sein. (0,5)
(*All values have to be produced by a sensor.*)

1.3. Gegeben seien außerdem die folgenden Relationen. Erweitern Sie das ER-Diagramm aus Aufgabe 1 durch Verwendung eines Abstraktionskonzeptes zu einem erweiterten (*Extended*) ER-Diagramm (EER-Diagramm), das dem gegebenen relationalen Modell entspricht. Nutzen Sie die Informationen aus den gegebenen Relationen. Weitere Tupel als die angegebenen existieren nicht. Achten Sie dabei darauf, dass Ihre Modellierung nicht kapazitätserhöhend (*capacity-increasing*) oder kapazitätsvermindernd (*capacity-decreasing*) ist. (3)

License

ID	Established
CC	2001
MIT	1988
BSD	1999
IBM	1971
Terzio	2001
TUBS	2011
TLDR	1990

Open

ID → License	Version	Comm_use
CC	4.0	nein
MIT	1.0	ja
BSD	Free	ja

Closed

ID → License	Evilness
IBM	235
TUBS	9001

Own

ID → License	Meta(Charset, Length)
Terzio	(UTF-16, 654327)
TUBS	(Win-1251, 25403)

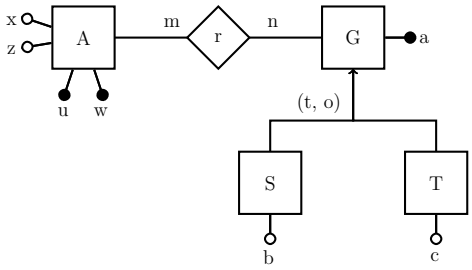
1.4. Sind die folgenden Integritätsbedingungen (*integrity constraints*) im ER-Entwurf abgebildet (*represented*)?

a) Ein Wert kann von mehreren Sensorquellen erzeugt worden sein. (0,5)
(*A value can be prduced by multiple sensors.*) ◯ Ja ◯ Nein

b) Einer Sensorquelle müssen mindestens drei Werte zugeordnet sein. (0,5)
(*A sensor source has to have produced at least three values.*) ◯ Ja ◯ Nein

Aufgabe 2: Relationaler Entwurf (8 Punkte)

2.1. Gegeben sei das folgende erweiterte Entity-Relationship-Modell.



a) Erstellen Sie nun zum oben gegebenen EER-Diagramm den Relationalen Entwurf (*relational design*), nicht jedoch für die Relation $A(\underline{u, w}, x, z)$. Benutzen Sie zur Abbildung der Generalisierung/Spezialisierung den ER-Stil. (2)

b) Wie viele Relationen entstehen, wenn das oben gegebene EER-Diagramm im Null-Stil und im objektorienterten Stil abgebildet wird? (1)

Null-Stil		OO-Stil	

c) Die Abbildung der Generalisierung/Spezialisierung im gegebenen ER-Diagramm mit dem ER-Stil ist (0,5)

- ◯ kapazitätserhaltend. (*capacity-preserving*)
- ◯ kapazitätsvermindernd. (*capacity-decreasing*)
- ◯ kapazitätserhöhend. (*capacity-increasing*)

d) Weiterhin ist die Relation A mit $A(\underline{u, w}, x, z)$ sowie den funktionalen Abhängigkeiten (*functional dependencies*) (2,5)

$(u, w) \to x$, $(u, w) \to z$, $z \to u$, $u \to x$

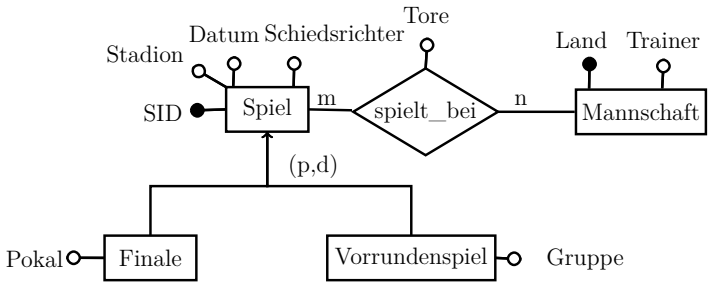
gegeben. Weitere funktionale Abhängigkeiten existieren nicht.

Normalisieren Sie die Relation A bis zur Boyce-Codd-Normalform. Geben Sie für jeden Normalisierungsschritt die Zerlegung (*decompostion*) der Relation(en) sowie die möglicherweise Normalform-verletzende (*violating*) funktionalen Abhängigkeit (*functional dependency*) an. Unterstreichen Sie Schlüsselattribute (*key attributes*). Sie können davon ausgehen, dass sich A bereits in der ersten Normalform befindet.

2.2. Ist es sinnvoll, funktionale Abhängigkeiten (*functional dependencies*) aus einer Instanz, also dem Zustand einer Relation, abzuleiten? Begründen Sie in höchstens fünf Sätzen. (2)

Aufgabe 3: Anfragesprachen (9 Punkte)

Gegeben sei folgendes Schema einer *Fußballdatenbank* mit Beispieltupeln.
(*Spiel - game, Mannschaft - team, Tor - Goal, Schiedrichter - referee, Vorrundenspiel - group match*)



spielt_bei

SID → Spiel	Land → Mannschaft	Tore
1	Schweden	0
1	England	2
2	Russland	3
2	Kroatien	4
3	Frankreich	4
3	Kroatien	2
4	Frankreich	1
4	Belgien	0
5	Kroatien	2
5	England	1
6	Schweden	1
6	Schweiz	0

Mannschaft

Land	Trainer
Russland	Tschertschessow
Kroatien	Dalić
Schweiz	Petković
Belgien	Martínez
Schweden	Andersson
Frankreich	Deschamps
England	Southgate

Spiel

SID	Stadion	Datum	Schiedsrichter
1	Samara	2018-07-07	Kuipers
2	Sotchi	2018-07-07	Ricci
3	Moskau	2018-07-15	Pitana
4	St. Petersburg	2018-07-10	Cunha
5	Moskau	2018-07-11	Çakır
6	St. Petersburg	2018-07-03	Skomina

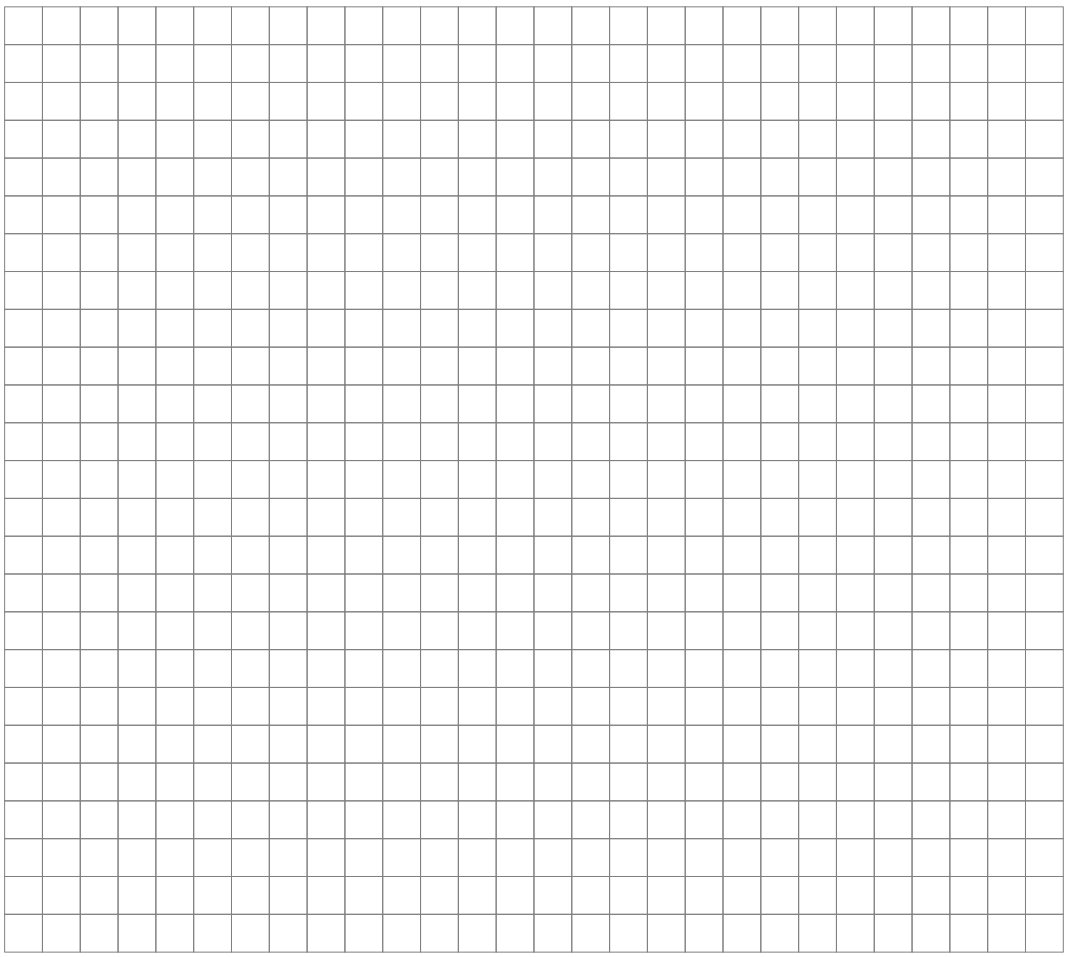
Finale

SID → Spiel	Pokal
3	FIFA-WM-Pokal

Vorrundenspiel

SID → Spiel	Gruppe
4	A
5	B

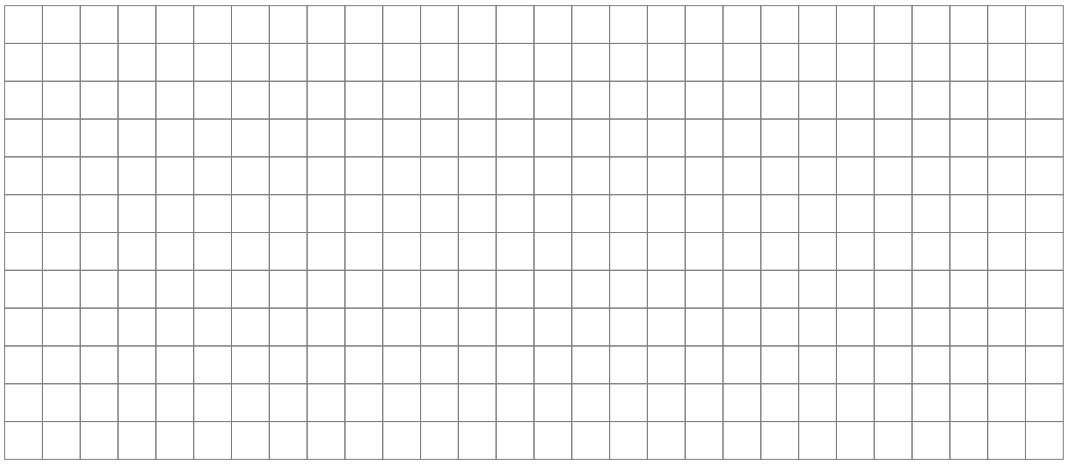
3.1. Formulieren Sie eine gültige SQL-Abfrage: *Eine Auflistung aller Länder und der Summe ihrer Tore über alle Spiele.* (2)



3.2. Geben Sie das Ergebnis (inklusive Attributnamen) auf Basis der vorliegenden Daten zur folgenden Anfrage an. (1)

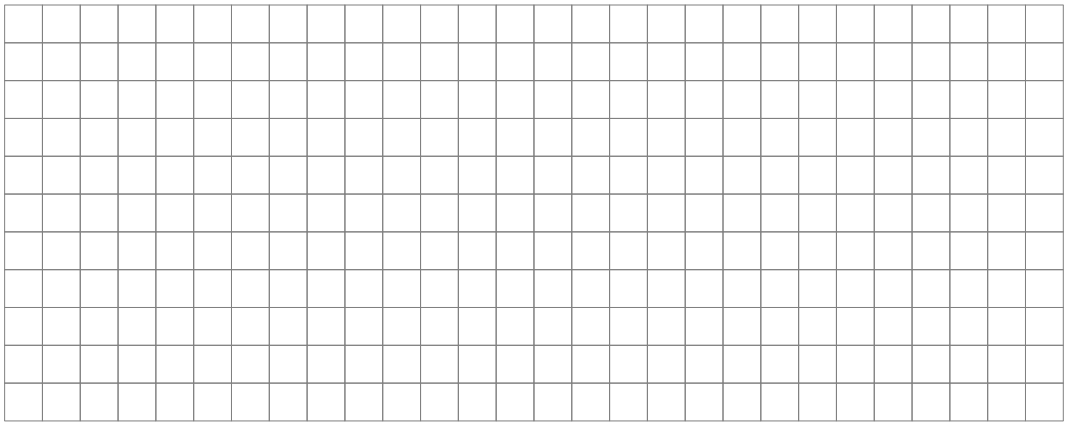
```
SELECT s.Datum AS D, s.Schiedsrichter AS S
FROM Spiel s
WHERE s.Schiedsrichter LIKE '%i%';
```

Ergebnistabelle:

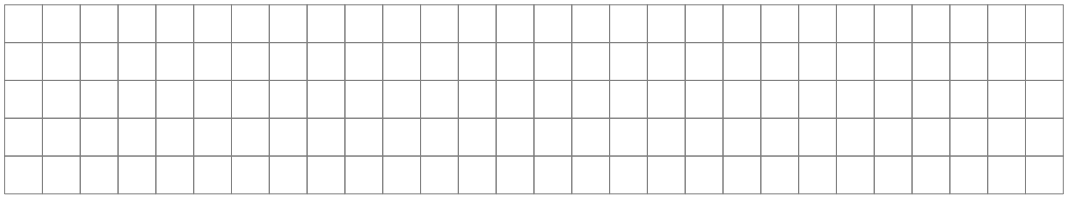


3.3. Formulieren Sie sowohl eine gültige SQL-Query als auch eine äquivalente Anfrage in relationaler Algebra: *Eine duplikatfreie (*duplicate-free*) Auflistung der SIDs aller Spiele, die weder ein Finale noch ein Vorrundenspiel sind.* (3)

SQL:

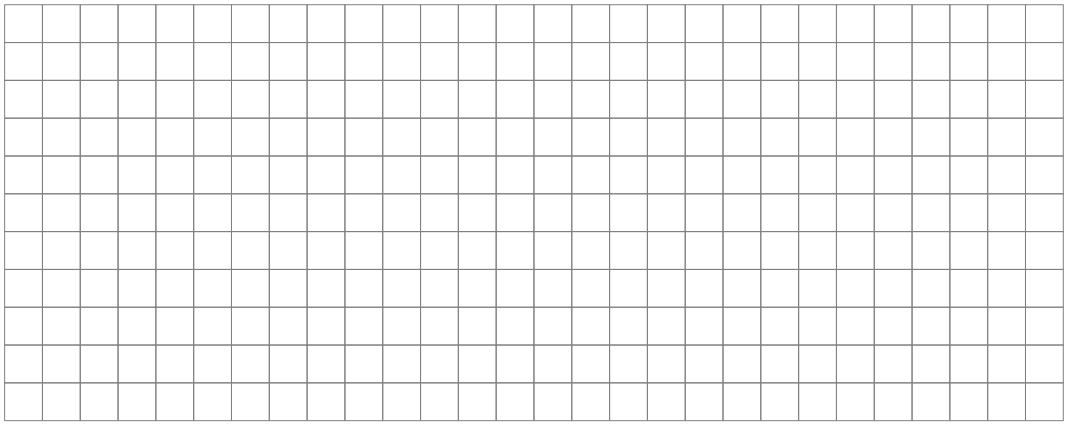


Relationale Algebra:

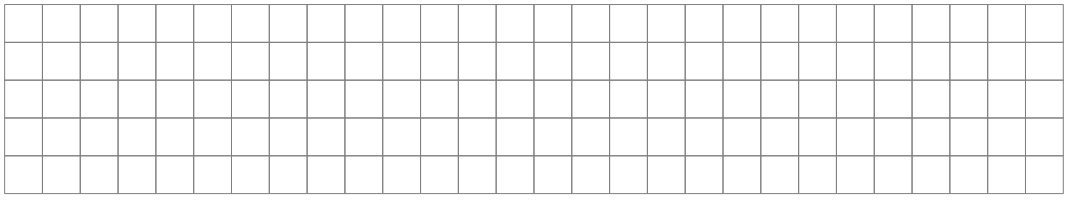


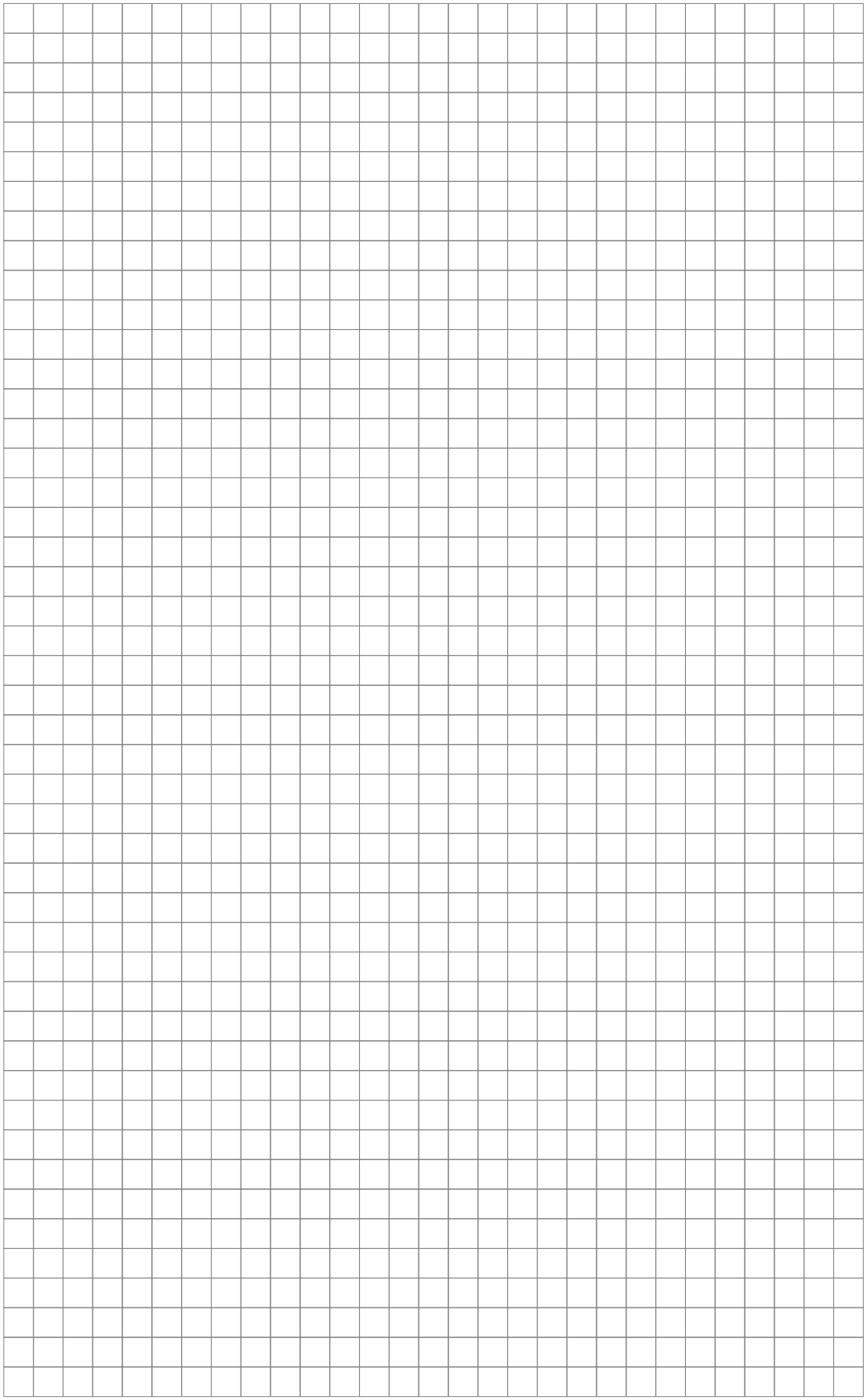
3.4. Formulieren Sie sowohl eine gültige SQL-Query als auch eine äquivalente Anfrage in relationaler Algebra: *Die Länder, die insgesamt die höchste Anzahl Tore geschossen haben.* (3)

SQL:



Relationale Algebra:

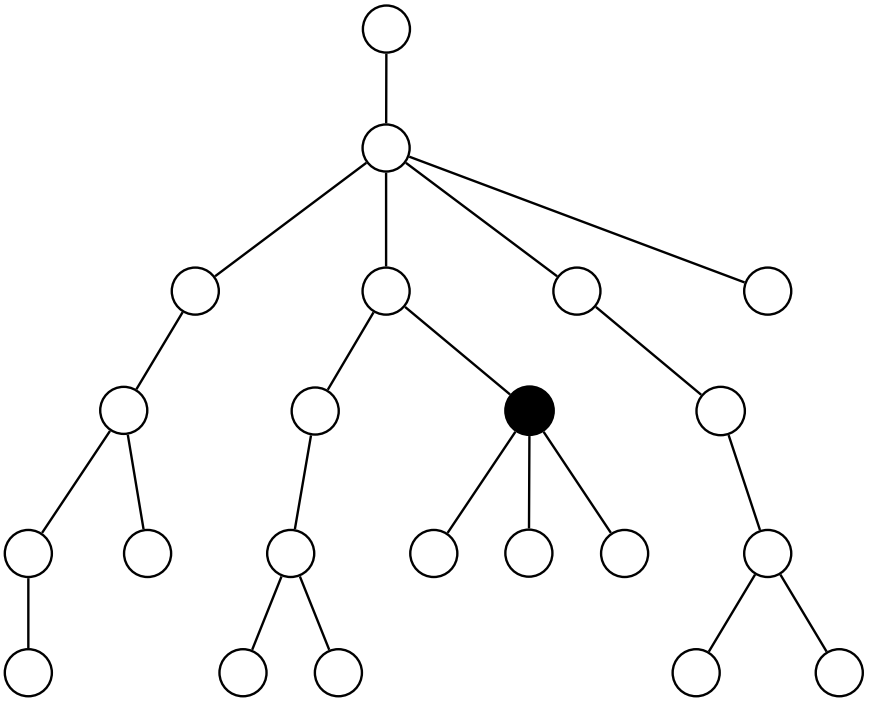




Aufgabe 4: XML (5,5 Punkte)

4.1. Gegeben ist die Baum-Repräsentation (*tree representation*) eines XML-Dokuments. Markieren Sie die angegebenen Achsen (*axes*) ausgehend vom schwarz markierten Knoten (*node*), indem Sie die enthaltenen Knoten umkreisen (*(to) circle*) und das Umkreiste mit dem Namen der Achse beschriften. Sollte eine Achse keine Knoten enthalten, brauchen Sie für diese keine Markierung vorzunehmen. (2,5)

`child`, `ancestor-or-self`, `following-sibling`, `preceiding`, `descendant`



4.2. XML-Dokumente können in bestimmte Kategorien klassifiziert werden.

a) Nennen Sie eine Eigenschaft, in der sich ein datenzentriertes (*data-centric*) von einem dokumentenzentrierten (*document-centric*) XML-Dokument unterscheidet. (0,5)

.....

b) Geben Sie kurz beispielhaft ein dokumentenzentriertes XML-Dokument an. (1)

4.3. Zur Durchführung einer Wahlanalyse sollen die Datensätze aus den verschiedenen Datenbanken in ein zentrales Data Warehouse geladen werden. Dazu wird als Austauschformat XML verwendet. Im Folgenden sehen Sie einen Auszug aus einer dieser XML-Dateien.

```
<?xml version="1.0" encoding="UTF-8"?>
<wahl name="bundestag" jahr="2017">
        <partei name="RotGelbWeiss" id="p1">
                <mitglied alter="25">Meier</mitglied>
                <mitglied alter="64">Soenne</mitglied>
        </partei>
        <partei id="p2" name="OrangeGruen">
                <mitglied alter="74">Hoenne</mitglied>
                <mitglied alter="21">Aref</mitglied>
                <mitglied alter="34">Kondler</mitglied>
        </partei>
        <partei id="p3" name="BlauSchwarzLila">
                <mitglied alter="34">Pahler</mitglied>
                <mitglied alter="52">Lehner</mitglied>
        </partei>
        <wahlbezirk id="wb1-1" stadt="Berlin">
                <kandidat name="Meier" alter="25"/>
                <kandidat name="Schneider" alter="54" hatSitz="Ja"/>
                <partei id="p1">34,2</partei>
                <partei id="p2">48,5</partei>
                <partei id="p3">17,3</partei>
        </wahlbezirk>
        <wahlbezirk id="wb1-2" stadt="Berlin">
                <kandidat name="Aref" alter="21" hatSitz="Ja"/>
                <kandidat name="Pahler" alter="34"/>
                <partei id="p2">75,5</partei>
                <partei id="p3">24,5</partei>
        </wahlbezirk>
        <wahlbezirk id="wb2-1" stadt="Hamm">
                <kandidat name="Kondler" alter="34" hatSitz="Ja"/>
                <kandidat name="Soenne" alter="64"/>
                <partei id="p1">34,2</partei>
                <partei id="p2">65,8</partei>
        </wahlbezirk>
</wahl>
```

Sie sollen nun zur manuellen Validierung der Daten einzelne Informationen aus der oben stehenden exportierten XML-Datei extrahieren. Entwickeln Sie zu diesem Zweck zu den folgenden Aufgabenstellungen XPath-Anfragen:

a) Alle Kandidaten. (0,5)

.....

b) Die Namen aller Kandidaten, die im XML-Dokument vor dem Kandidaten mit Name ”Schneider” stehen (1)

.....

Aufgabe 5: Data Streams Management & DWH (5 Punkte)

5.1. Gegeben sei die folgende Ergebnismatrix eines Count-Min Sketch-Durchlaufs:

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

a) Wie viele Werte wurden insgesamt in den Count-Min Sketch eingetragen? (0,5)

. . . . . .

b) Aktualisieren Sie den Count-Min Sketch aus der vorherigen Aufgabe für jeweils eine Observation der folgenden Werte: (2)

v	$h_0(v)$	$h_1(v)$	$h_2(v)$
„Hello“	2	0	1
„Trello“	2	0	0

$$\begin{array}{c} \\ h_0 \\ h_1 \\ h_2 \end{array}\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} \end{array}$$

5.2. Welche Art von Fenster (*window*) über einen Datenstrom mit KFZ-Kennzeichen (*license plate*) wird hier beschrieben: Das Fenster enthält alle KFZ-Kennzeichen der letzten 5 Minuten. (0,5)

. . . . . .

5.3. Das Star-Schema verletzt eine der Ihnen bekannten Normalformen. Um welche Normalform handelt es sich? Erklären Sie in höchsten drei Sätzen, warum diese Verletzung akzeptiert wird. (1,5)

5.4. Nennen sie einen Nachteil des Fullfact-Schemas gegenüber dem Snowflake-Schema. (0,5)

.....

Aufgabe 6: Data Analysis (5,5 Punkte)

6.1. Gegeben seien die Vektoren $\vec{v}_1 = (3, 2, 6, 5, 10)$ und $\vec{v}_2 = (3, 2, 1, 10, 10)$. Geben Sie die Maximum-, Euklidische, Manhattan- und Hammingdistanz der beiden Vektoren an. (2)

d_{max}		d_{eukl}		d_{manh}		d_{ham}	

6.2. Mit welcher Aggregatfunktion werden im K-Means Algorithmus neue Clusterzentren aus einer bestehenden Zuordnung der Punkte zu einem Cluster berechnet? Geben Sie den Namen der Funktion an. (0,5)

.....

6.3. Was ist der Unterschied zwischen Supervised Learning und Unsupervised Learning? Erklären Sie in höchstens vier Sätzen. (2)

.....
.....
.....
.....
.....
.....
.....
.....
.....

6.4. Vervollständigen Sie die Beschreibung für das Optimierungsproblem (*optimization problem*) für lineare Regression nach der Methode der kleinsten Quadrate (*least squares*). Gegeben sind Werte $x_1, \cdots, x_n \in \mathbb{R}$ sowie deren Label $y_1, \cdots, y_n \in \mathbb{R}$. (1)

$$\underset{a,b}{argmin}$$

Aufgabe 7: Multiple Choice (10 Punkte)

Diese Aufgabe umfasst 10 Multiple-Choice-Fragen. Sie bestehen aus jeweils mehreren 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. Welche der folgenden Konzepte in ER Modellen sind Beschränkungen (Constraints)? (1)

- ☐ Totalität (*totality*)
- ☐ Kardinalitäten (*cardinalities*)
- ☐ Attribute
- ☐ Markierung von Attributen als Schlüsselattribut (*key attributes*)

7.2. Welche Aussagen gelten für skalare Unterabfragen (*scalar subqueries*) in SQL? (1)

- ☐ Skalare Unterabfragen sind immer korreliert (*correlated*).
- ☐ Skalare Unterabfragen können korreliert (*correlated*) sein.
- ☐ Skalare Unterabfragen können mehre Tupel zurückgeben.
- ☐ Skalare Unterabfragen geben ein Tupel mit einem Attribut zurück.

7.3. Wie viele Relationen werden benötigt, um eine totale und überlappende (*overlapping*) Generalisierung/Spezialisierungsbeziehung mit n Spezialisierungen im objektorientierten Stil abzubilden? (1)

- ☐ $2^n - 1$
- ☐ 2^n
- ☐ n
- ☐ 1

7.4. Gegeben seien die Relationen R und S sowie der Ausdruck $\sigma_c(R \bowtie S)$. c ist ein beliebiges Selektionsprädikat. Geben Sie alle korrekten Aussagen über die mögliche Kardinalität des Ausdrucksergebnisses (Anzahl der Tupel) an. (1)

- ☐ Die Kardinalität ist nie größer als $|R| \cdot |S|$.
- ☐ Das Ergebnis kann leer sein.
- ☐ Die Kardinalität ist nie größer als $|R| + |S|$.
- ☐ Das Ergebnis kann nicht leer sein.

7.5. Wählen Sie allem zum Ausdruck $\sigma_c(R \times S)$ äquivalenten Ausdrücke. (1)

- ☐ $\sigma_c(R \bowtie_c S)$
- ☐ $\sigma_c(R) \bowtie_c S$
- ☐ $\sigma_c(\sigma_c(R \times S))$
- ☐ $R \bowtie_c S$

7.6. Welches der ACID-Kriterien beschreibt die folgende Aussage: „Eine Transaktion wird vollständig oder gar nicht ausgeführt. (*A transaction is either executed entirely or not executed at all.*)“ (1)

- ☐ Atomicity
- ☐ Consistency
- ☐ Isolation
- ☐ Durability

7.7. Welche Vorteile bietet die Verbindung von MapReduce und HDFS? (1)

- ☐ Skalierbarkeit (*scalability*)
- ☐ Ausfallsicherheit (*failure safety*)
- ☐ Schnelle Transaktionsverarbeitung (*fast transaction processing*)
- ☐ Deklarative Formulierung von Analyseaufgaben

7.8. Die Wahrscheinlichkeit im Reservoir Sampling dafür, dass das aktuelle Element in das Sample aufgenommen wird, ... (1)

- ☐ steigt monton (*increases montonically*) mit der Anzahl der gesehenen Elemente.
- ☐ sinkt monoton (*decreases montonically*) mit der Anzahl der gesehenen Elemente.
- ☐ ist unabhängig (*independent*) von der Anzahl der gesehenen Elemente.

7.9. Welche der folgenden Aussagen über Distanzfunktionen sind korrekt? (1)

- ☐ Die Euklidische Distanz kann zwischen zwei Vektoren mit jeweils unterschiedlicher Länge berechnet werden.
- ☐ Die Manhattan-Distanz kann für zwei Vektoren mit beliebiger aber gleicher Länge berechnet werden.
- ☐ Die Definition der Maximum-Distanz lautet: $D_{max}(x, y) = \max_i(|x_i - y_i|)$
- ☐ Die Hamming-Distanz kann auch auf nicht-numerische Daten angewendet werden.

7.10. Welche der folgenden Aussagen über den k-Means-Algorithmus sind korrekt? (1)

- ☐ Clusterzentren werden als Median ihrer zugehörigen Punkte berechnet.
- ☐ Der Wert k wird vom Algorithmus automatisch berechnet.
- ☐ Der K-Means Algorithmus verwendet die Euklidische Distanz als Distanzfunktion.
- ☐ Als Eingabe benötigt der Algorithmus eine Menge von Punkten sowie dazugehörige Label.