Informationssysteme und Datenanalyse | SoSe 25

Ersttermin-Klausur vom 1. August 2025

Bearbeitungszeit: 120 Minuten

Zu erreichende Punktzahl: 116 Punkte

Einsicht war ungefähr eine halbe Stunde nach Ende der Klausur online möglich. Für Anmerkungen bezüglich der Bewertung oder Fragen gab es im ISIS-Kurs ein Formula, welches man ausfüllen musste. Diese wurden von der Modulleitung gesichtet und ggf. wurde die Bewertung angepasst.

Bei weiteren Fragen gab es eine Präsenzeinsicht ungefähr eine Woche später.

Punkteschema

Frage	Punkte
Modellierung	14,75
Relationaler Entwurf	10
Normalisierung	9
DDL	13
Relationale Algebra	12
DQL	19
Data Warehousing	8
Data Streams	12
Data Science (Fallstudie)	8
Data Science (Allgemein)	10,25
	116

Notenschlüssel:

Note	ab ... Punkten
1.0	100
1.3	85
1.7	75
2.0	65
2.3	60
2.7	57
3.0	54
3.3	51
3.7	48
4.0	45
5.0	0

1. Modellierung

Frage m-mod
Teilweise richtig
Erreichte Punkte 9,50 von 11,00
Frage markieren

Wissenschaftsdatenbank (Modellierung)

Christine und Erhard sind Vorständinnen für die Gesellschaft für Informatik e.V.. Sie organisieren regelmäßig Konferenzen und Workshops, um den Austausch zwischen Wissenschaftler:innen zu fördern. Dabei werden viele wissenschaftliche Arbeiten veröffentlicht, die in einer Datenbank erfasst werden sollen. Bei der letzten Dungeons & Dragons-Runde skizzierten sie dafür ein erweitertes Entity-Relationship-Diagramm, welches im epischen Boss-Kampf leider zerstört wurde.

Helfen Sie Christine und Erhard, das erweiterte ER-Diagramm wiederherzustellen!

Hinweise

- Sie können hierzu die auf den Notizzetteln ausgedruckten Tabellen nutzen. Diese finden Sie als Kopie auch weiter unten in dieser Aufgabe.
- Kardinalitäten von Relationshiptypen müssen eigentlich mit unterschiedlichen Buchstaben angegeben werden. Dies ist in dieser Aufgabe aus technischen Gründen nicht möglich. Deshalb werden z.B. die Kardinalitäten eines many to many Relationshiptyps ausnahmsweise mit zwei n angegeben.

Spezifikation 1: *Repräsentativer* Auszug der Tabellen

Institution

Name	Anschrift	Direktor:in
TU Berlin	Straße des 17. Juni 135, Berlin	Dr. Eva Schulz
LMU München	Geschwister-Scholl-Platz 1, München	Prof. Markus Weber
RWTH Aachen	Templergraben 55, Aachen	Dr. Lena Hoffmann
Uni Heidelberg	Grabengasse 1, Heidelberg	Prof. Dr. Paul König
TU Berlin	El Gouna Campus, Hurghada, Egypt	Dr. Eva Schulz

Autor:in

Name	ORCID	Anschrift
Amina Hassan	0000-0006-1122-3344	Hurghada, Egypt
Jonas Schmidt	0000-0002-3456-7890	München
Hiroshi Tanaka	0000-0007-2233-4455	Tokyo, Japan
Clara Becker	0000-0003-4567-8901	Heidelberg
Samuel Okoye	0000-0008-3344-5566	Lagos, Nigeria

Dissertation

Name → Autor:in	Titel	Verteidigung
Amina Hassan	Solarenergie im urbanen Raum	2021-12-10
Jonas Schmidt	KI in der medizinischen Diagnostik	2021-06-15
Hiroshi Tanaka	Multi-Agenten-Systeme in Smart Cities	2023-02-14
Clara Becker	Nachhaltige IT-Infrastrukturen	2023-03-22

Paper

DOI	Seiten	ist_online	Typ	Turnus	Impact-Factor	Publisher	Name	PKonferenz → Konferenz	WKonferenz → Konferenz
10.1000/solar2021	12	TRUE	Journal	jährlich	2.6	*NULL*	*NULL*	*NULL*	*NULL*
10.1000/medai2021	9	TRUE	Journal	halbjährlich	3.2	*NULL*	*NULL*	*NULL*	*NULL*
10.1000/sustech2023	15	FALSE	Proceedings	*NULL*	*NULL*	IEEE	*NULL*	Sustainable IT Forum	*NULL*
10.1000/agents2023	11	TRUE	Workshop	*NULL*	*NULL*	*NULL*	Smart Multi-Agents	*NULL*	*NULL*
10.1000/dbsys2024	8	TRUE	Workshop	*NULL*	*NULL*	*NULL*	DB Systems Today	*NULL*	Quantum Days
10.1000/edtechai2022	14	FALSE	Journal	monatlich	2.9	*NULL*	*NULL*	*NULL*	*NULL*
10.1000/aiethics2023	11	TRUE	Proceedings	*NULL*	*NULL*	IEEE	*NULL*	AI Summit Europe	*NULL*
10.1000/greengrid2024	13	FALSE	Proceedings	*NULL*	*NULL*	ACM	*NULL*	Sustainable IT Forum	*NULL*

Konferenz

Name	Impact-Factor	Ort
AI Summit Europe	2.7	Berlin
Sustainable IT Forum	1.8	München
Quantum Days	3.1	Aachen
EduTech 2024	2.4	Hamburg

besteht_aus

DOI → Paper	Name → Dissertation
10.1000/solar2021	Clara Becker
10.1000/solar2021	Hiroshi Tanaka
10.1000/medai2021	Jonas Schmidt
10.1000/agents2023	Clara Becker
10.1000/edtechai2022	Hiroshi Tanaka

schreibt

Name → Autor:in	DOI → Paper	Position
Amina Hassan	10.1000/solar2021	1
Clara Becker	10.1000/solar2021	2
Jonas Schmidt	10.1000/medai2021	1
Hiroshi Tanaka	10.1000/agents2023	1
Hiroshi Tanaka	10.1000/edtechai2022	2
Samuel Okoye	10.1000/edtechai2022	1

arbeitet_in

Name → Autor:in	Name → Institution	Anschrift → Institution	Datum
Amina Hassan	TU Berlin	Straße des 17. Juni 135, Berlin	2020-01-12
Jonas Schmidt	LMU München	Geschwister-Scholl-Platz 1, München	2019-11-15
Clara Becker	Uni Heidelberg	Grabengasse 1, Heidelberg	2021-05-20
Hiroshi Tanaka	TU Berlin	El Gouna Campus, Hurghada, Egypt	2022-07-01
Samuel Okoye	RWTH Aachen	Templergraben 55, Aachen	2023-04-10
Amina Hassan	TU Berlin	El Gouna Campus, Hurghada, Egypt	2022-08-01

zitiert

Selbst → Paper	Ziel → Paper
10.1000/edtechai2022	10.1000/solar2021
10.1000/greengrid2024	10.1000/sustech2023
10.1000/aiethics2023	10.1000/medai2021
10.1000/dbsys2024	10.1000/agents2023
10.1000/dbsys2024	10.1000/edtechai2022

veranstaltet

Name → Institution	Anschrift → Institution	Name → Autor:in	Name → Konferenz	Jahr
TU Berlin	El Gouna Campus, Hurghada, Egypt	Amina Hassan	AI Summit Europe	2023
LMU München	Geschwister-Scholl-Platz 1, München	Clara Becker	Sustainable IT Forum	2022
RWTH Aachen	Templergraben 55, Aachen	Samuel Okoye	Quantum Days	2023
TU Berlin	Straße des 17. Juni 135, Berlin	Hiroshi Tanaka	EduTech 2024	2024

editiert

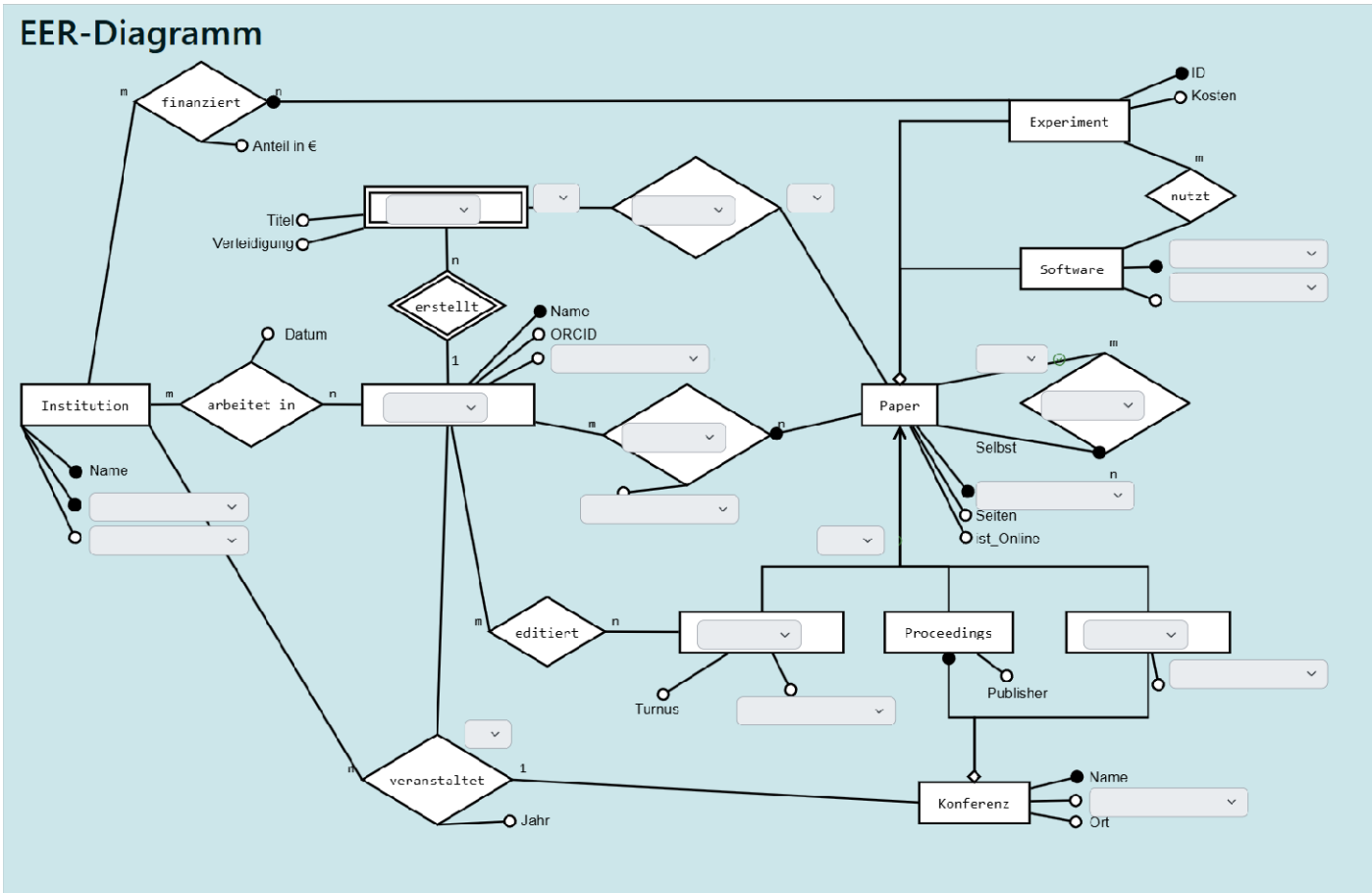
Name → Autor:in	DOI → Paper
Samuel Okoye	10.1000/aiethics2023
Hiroshi Tanaka	10.1000/dbsys2024
Clara Becker	10.1000/greengrid2024
Hiroshi Tanaka	10.1000/medai2021

Spezifikation 2: Gesprächsnotizen

Aus der Video-Aufzeichnung des D&D-Spiels konnte eine KI außerdem folgende Notizen extrahieren:

- Teil eines Papers sind ein oder mehrere Software und / oder Experimente.
- Software wird anhand einer URL indentifiziert und besitzt zusätzlich eine DOI.
- Experimente besitzen eine eindeutige ID sowie einen Eintrag für entstandene Kosten.
- Beliebig viele Experimente können beliebig viel Software nutzen.

a)



b)

Sie sollen die Generalisierung/Spezialisierung des Entity-Typs `Paper` in der Datenbank umsetzen. Geben Sie für jeden der Stile an, ob mit ihm "Proceedings" `editiert` werden können oder nicht.

kann editiert werden	kann *nicht* editiert werden	
○	○	Objektorientierter Stil
○	○	ER-Stil
○	○	Null-Stil

c)

Frage **m-ins**
Richtig
Erreichte Punkte 1.50 von 1.50
Frage markieren

Im Kontext der Datenbankerstellung haben Christine und Erhard außerdem bereits Daten für die Tabelle "finanziert" und "Experiment" eingegeben. Vervollständigen Sie die noch offenen Lücken so, dass sämtliche aus dem EER-Diagramm abgeleiteten Integritätsbedingungen erfüllt sind.

Experiment

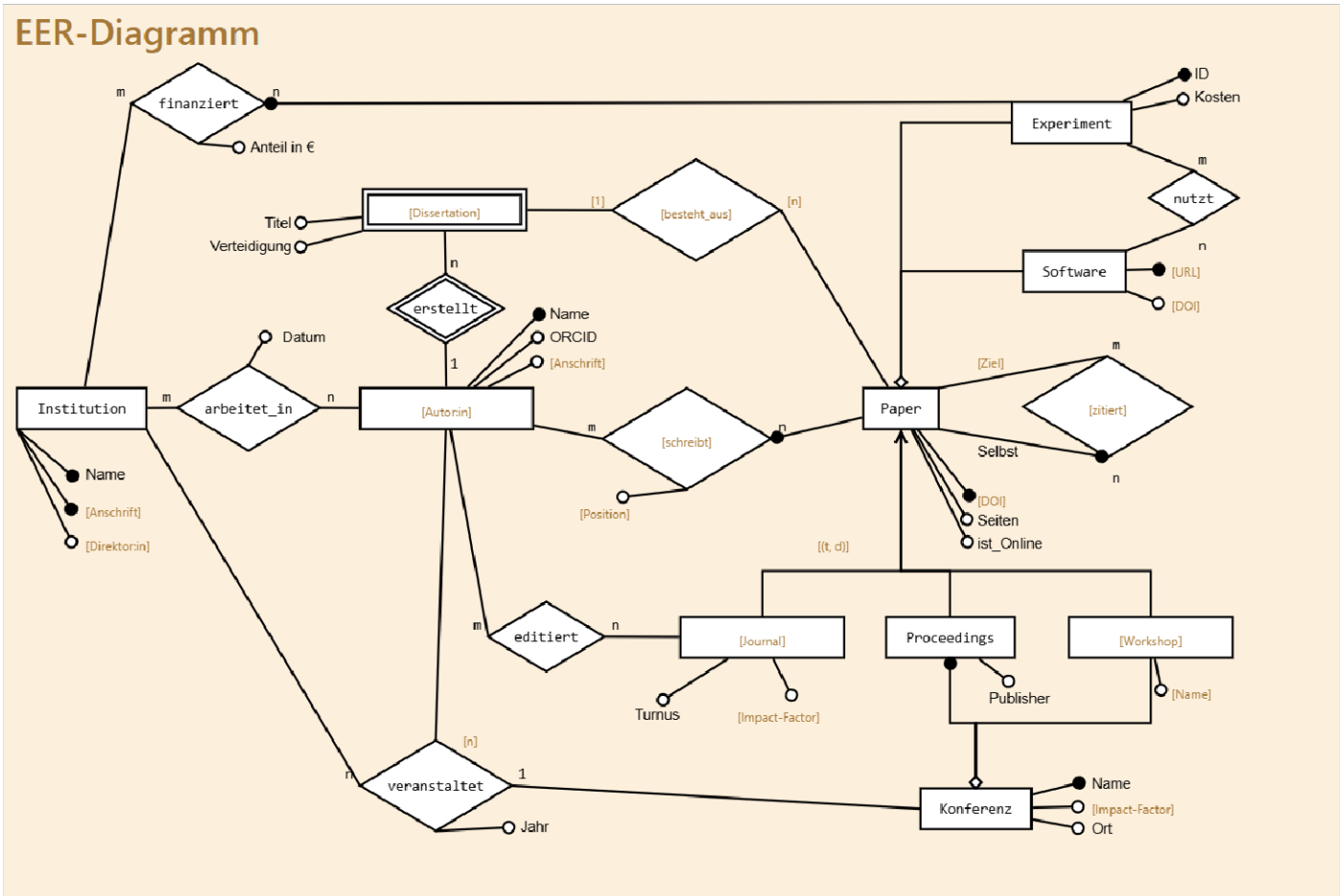
ID	Kosten
1	1500.00
2	2500.00
5	3000.00
6	2000.00

finanziert

Name → Institution	Anschrift → Institution	ID → Experiment	Anteil in €
TU Berlin	Straße des 17. Juni 135, Berlin	1	*NULL*
			NULL
			NULL
			NULL
			NULL

Lösung:

a)



b)

Objektorientierter Stil: kann *nicht* editiert werden
ER-Stil: kann *nicht* editiert werden
Null-Stil: kann editiert werden

c)

.

Alle Einträge sind unterschiedlich. Die Totalität zu "Experiment" ist erfüllt, da alle Experimente in der Tabelle "finanziert" vorkommen.

.

Eine richtige Antwort ist: `"TU Berlin"`

Eine richtige Antwort ist: `"El Gouna Campus, Hurghada, Egypt"`

Eine richtige Antwort ist **1**. Sie kann so eingegeben werden: `1`

Eine richtige Antwort ist: `"LMU München"`

Eine richtige Antwort ist: `"Geschwister-Scholl-Platz 1, München", "El Gouna Campus, Hurghada, Egypt"`

Eine richtige Antwort ist **2**. Sie kann so eingegeben werden: `2`

Eine richtige Antwort ist: `"RWTH Aachen"`

Eine richtige Antwort ist: `"Templergraben 55, Aachen", "El Gouna Campus, Hurghada, Egypt"`

Eine richtige Antwort ist **5**. Sie kann so eingegeben werden: `5`

Eine richtige Antwort ist: `"Uni Heidelberg"`

Eine richtige Antwort ist: `"Grabengasse 1, Heidelberg"`

Eine richtige Antwort ist **6**. Sie kann so eingegeben werden: `6`

2. Relationaler Entwurf

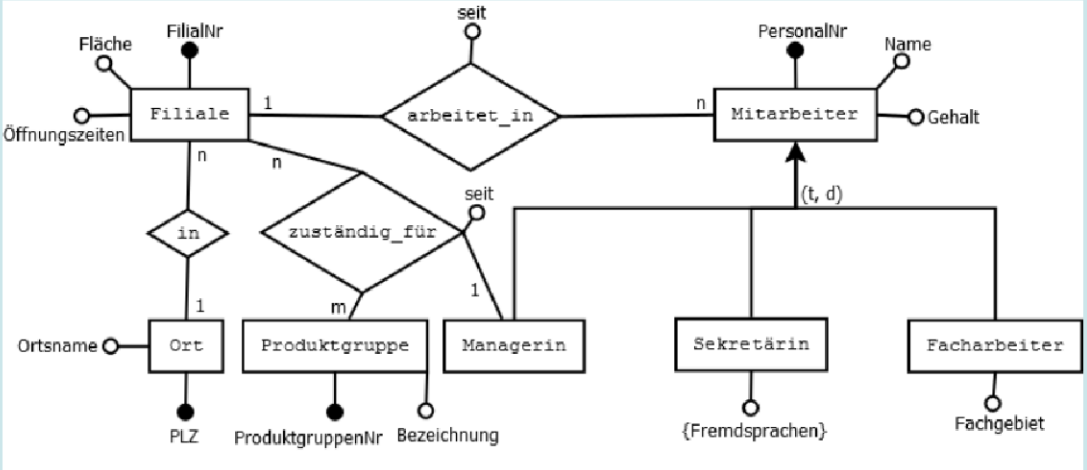
Information

Frage markieren

Personaldatenbank (Relationaler Entwurf)

Gegeben ist das folgende Entity-Relationship-Diagramm (siehe Abbildung).

Hinweis: Sie finden das untenstehende Diagramm auch auf dem ausgedruckten Notizzettel.



Frage **re-num**

Teilweise richtig

Erreichte Punkte von 10,00

Frage markieren

Das obige Schema soll in das relationale Modell überführt werden. Die Umsetzung erfolgt im Entity/Relationship-Stil (ER-Stil) mit Zusammenlegung. Es wird nicht normalisiert.

Alle folgenden Fragen beziehen sich jeweils auf das finale relationale Schema (ohne Normalisierung), das aus dem für die Frage relevanten Teil des ER-Diagramms resultiert.

Relation *zuständig_für* (3 Punkte)	**Antwort**
Wie viele Attribute hat die Relation *zuständig_für*?	
Wie viele Attribute bilden den Primärschlüssel der Relation *zuständig_für*?	
Wie viele Attribute referenzieren in *zuständig_für* als Fremdschlüssel auf andere Relationen?	

Fragen zu weiteren Relationen und ganzen Entwurf (7 Punkte)	**Antwort**
Wie viele Attribute hat die Relation *Facharbeiter*?	
Wie viele Relationen entstehen aus den Relationship-Typen des Schemas insgesamt?	
Wie viele Relationen entstehen aus den Entity-Typen des Schemas insgesamt?	
Ist die Transformation dieses Schemas in das relationale Modell im ER-Stil ohne Normalisierung verlustfrei?	
Führt die Umsetzung einer totalen Generalisierung/Spezialisierung im ER-Stil zu einem höheren Speicherbedarf im relationalen Modell als bei partieller Generalisierung/Spezialisierung?	

Lösung:

Frage **re-num**

Teilweise richtig

Erreichte Punkte von 10,00

Frage markieren

Das obige Schema soll in das relationale Modell überführt werden. Die Umsetzung erfolgt im Entity/Relationship-Stil (ER-Stil) mit Zusammenlegung. Es wird nicht normalisiert.

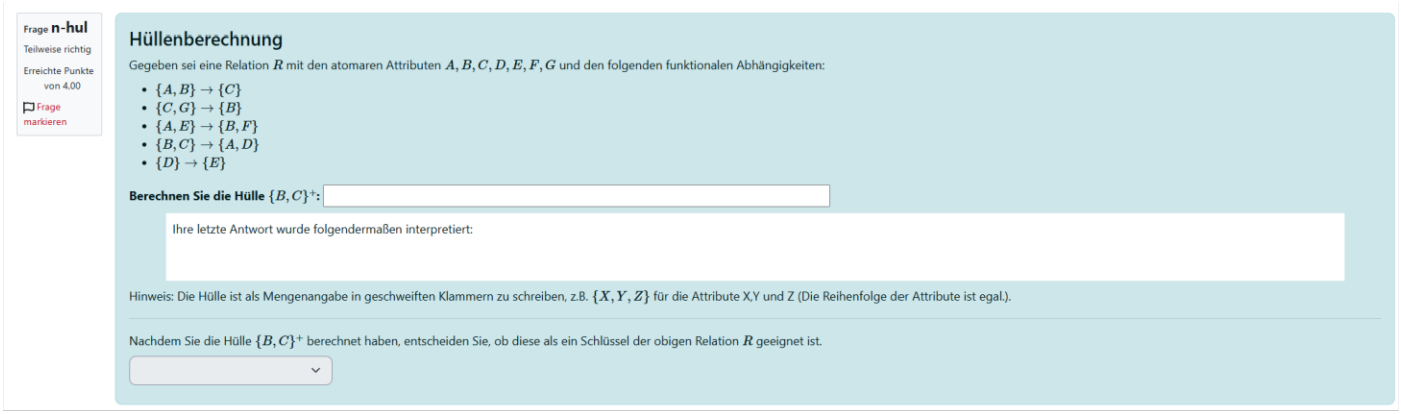
Alle folgenden Fragen beziehen sich jeweils auf das finale relationale Schema (ohne Normalisierung), das aus dem für die Frage relevanten Teil des ER-Diagramms resultiert.

Relation *zuständig_für* (3 Punkte)	**Antwort**
Wie viele Attribute hat die Relation *zuständig_für*?	4
Wie viele Attribute bilden den Primärschlüssel der Relation *zuständig_für*?	2
Wie viele Attribute referenzieren in *zuständig_für* als Fremdschlüssel auf andere Relationen?	3

Fragen zu weiteren Relationen und ganzen Entwurf (7 Punkte)	**Antwort**
Wie viele Attribute hat die Relation *Facharbeiter*?	2
Wie viele Relationen entstehen aus den Relationship-Typen des Schemas insgesamt?	1
Wie viele Relationen entstehen aus den Entity-Typen des Schemas insgesamt?	7
Ist die Transformation dieses Schemas in das relationale Modell im ER-Stil ohne Normalisierung verlustfrei?	Ja
Führt die Umsetzung einer totalen Generalisierung/Spezialisierung im ER-Stil zu einem höheren Speicherbedarf im relationalen Modell als bei partieller Generalisierung/Spezialisierung?	Ja

3. Normalisierung

a)



Frage **n-hul**
Teilweise richtig
Erreichte Punkte von 4.00
Frage markieren

Hüllenberechnung

Gegeben sei eine Relation R mit den atomaren Attributen A, B, C, D, E, F, G und den folgenden funktionalen Abhängigkeiten:

- $\{A, B\} \rightarrow \{C\}$
- $\{C, G\} \rightarrow \{B\}$
- $\{A, E\} \rightarrow \{B, F\}$
- $\{B, C\} \rightarrow \{A, D\}$
- $\{D\} \rightarrow \{E\}$

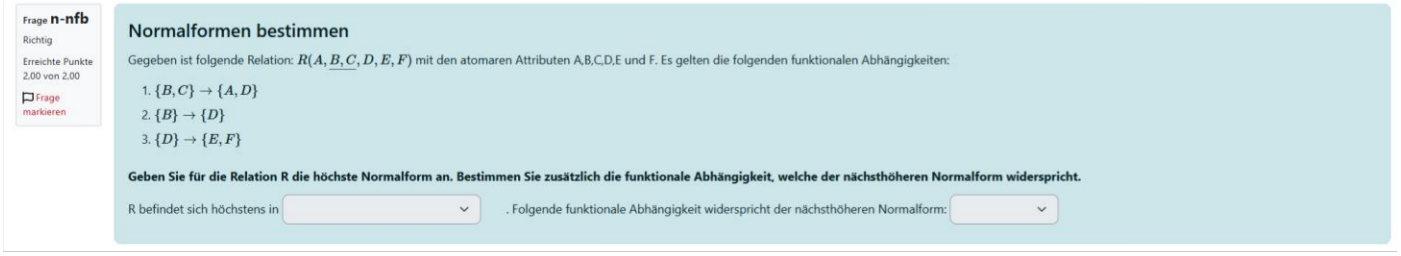
Berechnen Sie die Hülle $\{B, C\}^+$:

Ihre letzte Antwort wurde folgendermaßen interpretiert:

Hinweis: Die Hülle ist als Mengenangabe in geschweiften Klammern zu schreiben, z.B. $\{X, Y, Z\}$ für die Attribute X,Y und Z (Die Reihenfolge der Attribute ist egal.).

Nachdem Sie die Hülle $\{B, C\}^+$ berechnet haben, entscheiden Sie, ob diese als ein Schlüssel der obigen Relation R geeignet ist.

b)



Frage **n-nfb**
Richtig
Erreichte Punkte 2.00 von 2.00
Frage markieren

Normalformen bestimmen

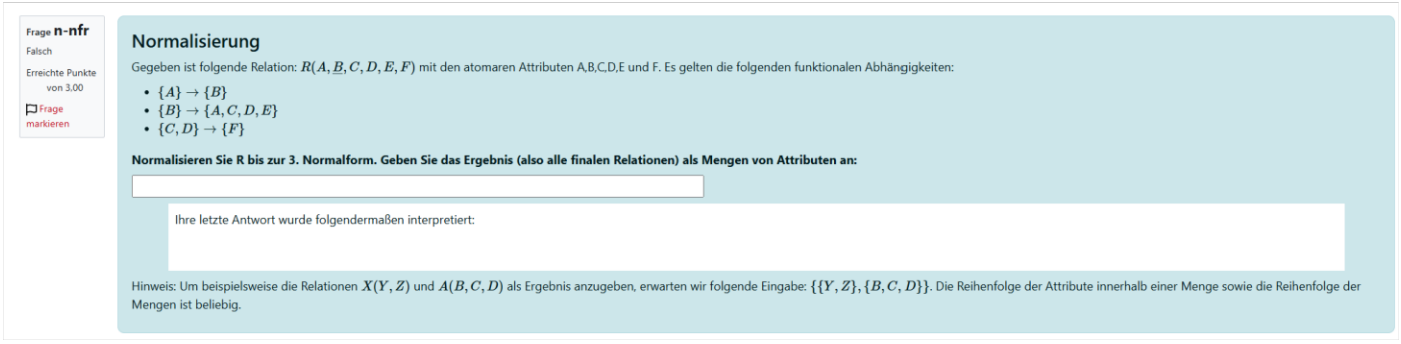
Gegeben ist folgende Relation: $R(A, \underline{B}, \underline{C}, D, E, F)$ mit den atomaren Attributen A,B,C,D,E und F. Es gelten die folgenden funktionalen Abhängigkeiten:

1. $\{B, C\} \rightarrow \{A, D\}$
2. $\{B\} \rightarrow \{D\}$
3. $\{D\} \rightarrow \{E, F\}$

Geben Sie für die Relation R die höchste Normalform an. Bestimmen Sie zusätzlich die funktionale Abhängigkeit, welche der nächsthöheren Normalform widerspricht.

R befindet sich höchstens in . Folgende funktionale Abhängigkeit widerspricht der nächsthöheren Normalform:

c)



Frage **n-nfr**
Falsch
Erreichte Punkte von 3.00
Frage markieren

Normalisierung

Gegeben ist folgende Relation: $R(A, \underline{B}, C, D, E, F)$ mit den atomaren Attributen A,B,C,D,E und F. Es gelten die folgenden funktionalen Abhängigkeiten:

- $\{A\} \rightarrow \{B\}$
- $\{B\} \rightarrow \{A, C, D, E\}$
- $\{C, D\} \rightarrow \{F\}$

Normalisieren Sie R bis zur 3. Normalform. Geben Sie das Ergebnis (also alle finalen Relationen) als Mengen von Attributen an:

Ihre letzte Antwort wurde folgendermaßen interpretiert:

Hinweis: Um beispielsweise die Relationen $X(Y, Z)$ und $A(B, C, D)$ als Ergebnis anzugeben, erwarten wir folgende Eingabe: $\{\{Y, Z\}, \{B, C, D\}\}$. Die Reihenfolge der Attribute innerhalb einer Menge sowie die Reihenfolge der Mengen ist beliebig.

Lösung:

a)

Die Attributmenge {B,C} ist kein Schlüssel von R, da nicht alle Attribute von R in der Hülle enthalten sind. Die Attributmenge {B,C} ist kein Schlüssel von R, da nicht alle Attribute von R in der Hülle enthalten sind.
Alle Attribute außer G können mit $\{B, C\}$ erreicht werden.

Eine richtige Antwort ist $\{A, B, C, D, E, F\}$. Sie kann so eingegeben werden: `{A,B,C,D,E,F}`
Eine richtige Antwort ist: `"{B, C} ist kein Schlüssel von R"`

b)

R befindet sich höchstens in der 1. Normalform, da D von einem Teilschlüssel abhängig ist $(2) : \{B\} \rightarrow \{D\}$ und somit nicht die 2. Normalform erfüllt.

Die richtige Antwort lautet:

Normalformen bestimmen

Gegeben ist folgende Relation: $R(A, B, C, D, E, F)$ mit den atomaren Attributen A,B,C,D,E und F. Es gelten die folgenden funktionalen Abhängigkeiten:

1. $\{B, C\} \rightarrow \{A, D\}$
2. $\{B\} \rightarrow \{D\}$
3. $\{D\} \rightarrow \{E, F\}$

Geben Sie für die Relation R die höchste Normalform an. Bestimmen Sie zusätzlich die funktionale Abhängigkeit, welche der nächsthöheren Normalform widerspricht.

R befindet sich höchstens in [der 1. Normalform]. Folgende funktionale Abhängigkeit widerspricht der nächsthöheren Normalform: [{B} → {D}]

c)

Die Relation ist in der 2. NF, die FD {\{C,D\} \rightarrow \{F\} } verhindert die 3. NF und muss daher extrahiert werden.

Eine richtige Antwort ist $\{\{A, B, C, D, E\}, \{C, D, F\}\}$. Sie kann so eingegeben werden: `{{A,B,C,D,E},{C,D,F}}`

4. DDL

Information
Frage markieren

Bibliotheksdatenbank (SQL DDL/DML)

Ausgangssituation

In einer Bibliotheksdatenbank gibt es bereits folgende zwei Tabellen mit beispielhaften Inhalt:

Buch

BuchID::INT	Title::VARCHAR(255)	Erscheinungsjahr::INT
1	1984	1949
2	Harry Potter und der Stein der Weisen	1997

Autor

AutorID::INT	Name::VARCHAR(100)
1	Orwell
2	Joane

In den folgenden vier Aufgaben sollen Sie die Datenbank erweitern und verändern. **Erstellen Sie dafür die notwendigen SQL-Ausdrücke, um die Datenbank so zu modifizieren, dass sie die Vorgaben erfüllt.**

Hinweise

- Sie finden die Tabellen auch ausgedruckt auf den Notizzetteln.
- Sie können davon ausgehen, dass Sie keine Tabellen löschen müssen.
- Trennen Sie einzelne SQL-Ausdrücke in Ihrer Lösung durch ein Semikolon.
- In den Tests zur Korrektheit der Informationen in einer Tabelle prüfen wir, ob die Tabelle den Modellierungsvorgaben entspricht und ob die aus der Vorlage und Aufgabenstellung gegebenen Daten korrekt enthalten sind. Bei den Tests auf Integritätsbedingungen fügen wir darüber hinaus Ihnen unbekannte Daten ein oder löschen existierende Daten und prüfen das Verhalten Ihrer Datenbank.

a)

Frage **ddl-c**
Teilweise richtig
Erreichte Punkte 4.00 von 6.00
Frage markieren

Erweitern Sie die bestehende Datenbankstruktur wie folgt:

- Erstellen Sie eine neue Tabelle `Schreibt`, die eine N:M-Beziehung zwischen Autor und Buch darstellt. (Ein Buch kann also von mehreren Autoren geschrieben werden, und ein Autor kann mehrere Bücher geschrieben haben.)
- Beachten Sie die Regeln für Primär- und Fremdschlüssel sowie deren Referenzen auf die bestehenden Tabellen.
- Verwenden Sie für die referenzierenden Spalten **die gleichen Namen wie die Primärschlüssel** der referenzierten Tabellen (z.B. AutorID).
- Zusätzlich soll die Tabelle eine Spalte `Beitragsanteil` enthalten, der den prozentualen Anteil bestimmt, mit dem ein Autor zum jeweiligen Buch beigetragen hat. Der Beitragsanteil soll daher als Ganzzahlwert im Bereich `1 ≤ Beitragsanteil ≤ 100` liegen. Dieser Wertebereich soll durch eine geeignete CHECK-Bedingung sichergestellt werden.
- Alle Attribute der Tabelle dürfen nicht NULL sein.

b)

Fügen Sie der Tabelle `Buch` eine neue Spalte `Bewertung` hinzu, in der eine numerische Bewertung mit insgesamt bis zu 5 Stellen und bis zu zwei Nachkommastellen gespeichert werden sollen. Verwenden Sie dafür einen geeigneten Datentyp, welcher Werte wie 999.99 oder 3.00 erlaubt.

c)

Tragen Sie in die Tabelle `Autor` einen Datensatz ein: `Name = 'Martin'`, mit der `AutorID = 3`.

d)

Ändern Sie in die Tabelle `Autor` bei dem Datensatz mit der `AutorID = 2` und dem `Namen = 'Joane'` den Namen der Autorin zu `Rowling`.

Lösung:

a)

```sql
CREATE TABLE Schreibt (
    AutorID INT NOT NULL,
    BuchID INT NOT NULL,
    Beitragsanteil INT NOT NULL CHECK (Beitragsanteil BETWEEN 1 AND 100),
    PRIMARY KEY (AutorID, BuchID),
    FOREIGN KEY (AutorID) REFERENCES Autor(AutorID),
    FOREIGN KEY (BuchID) REFERENCES Buch(BuchID)
);
```

b)

```sql
ALTER TABLE Buch
ADD COLUMN Bewertung DECIMAL(5,2);
```

c)

```sql
INSERT INTO Autor (AutorID, Name) VALUES (3, 'Martin');
```

d)

```sql
UPDATE Autor SET Name = 'Rowling' WHERE AutorID = 2;
```

5. Relationale Algebra

a)

Frage **ra-tf1**

Richtig

Erreichte Punkte 1,00 von 1,00

Frage markieren

Bewerten Sie die folgenden Aussagen.

True	False	
○	○	Die Selektion reduziert die Anzahl der Tupel, nicht die Anzahl der Attribute.
○	○	Alle Operationen der relationalen Algebra sind auch in SQL direkt verfügbar.
○	○	Eine Projektion ist notwendig, um Relationen zu verknüpfen.
○	○	Die relationale Algebra arbeitet mengenbasiert.

b)

Frage **ra-tf2**

Richtig

Erreichte Punkte 2,00 von 2,00

Frage markieren

Weisen Sie den folgenden relationalen Operatoren die passenden Namen zu. Nutzen Sie die Drag & Drop-Funktion, um die Begriffe in die Lücken einzufügen.

Die Operation [ ] filtert Tupel aus einer Relation anhand einer Bedingung.

Mit der Operation [ ] werden nur bestimmte Attribute einer Relation ausgewählt.

Die Operation [ ] kombiniert alle Tupel zweier Relationen ohne Duplikate.

Bei der Operation [ ] erhält man nur die Tupel, die in der ersten Relation, aber nicht in der zweiten vorkommen.

Die Operation [ ] erzeugt alle möglichen Kombinationen von Tupeln aus zwei Relationen.

Die Operation [ ] dient dazu, Relationen oder Attribute umzubenennen.

Die Operation [ ] verbindet Relationen automatisch über gemeinsame Attribute.

Die Operation [ ] gibt alle Tupel zurück, für die es zu jedem Tupel einer zweiten Relation eine passende Kombination gibt.

Information

Frage markieren

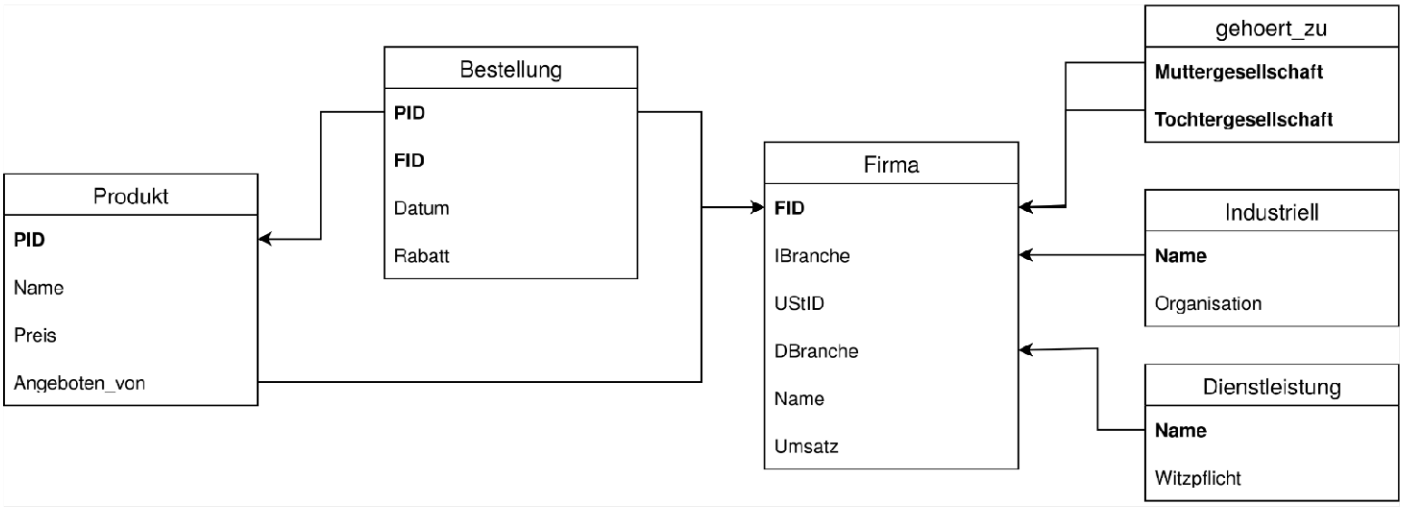
B2B-Onlineshop (Relationale Algebra)

Gegeben sei das folgende Datenbankschema eines B2B-Onlineshops inklusive der entstandenen Relationen und beispielhafter Tupel.

Hinweise

- Umsatzzahlen sind in Tausend angegeben.
- Das Zeichen , steht nicht für die Dezimaltrennung, sondern signalisiert die Tausenderabstände.
- In manchen Aufgaben wird die Relationale Algebra ausnahmsweise auf Multimengen und *nicht* auf Mengen definiert.

Sie finden das untenstehende Diagramm sowie die Relationen auch auf den ausgedruckten Notizzetteln.



Bestellung

PID → Produkt	FID → Firma	Datum	Rabatt
1	1	2023-06-01 10:00:00	0.1
2	2	2023-06-02 14:30:00	0.0
3	3	2023 06 03 09:15:00	0.05
1	4	2023-06-04 16:45:00	0.0
6	3	2023-06-05 12:30:00	0.2
2	4	2023-06-06 11:00:00	0.0
5	2	2023 06 07 15:45:00	0.1
3	4	2023-06-08 13:15:00	0.0
4	1	2023-06-09 17:30:00	0.15
5	3	2023-06-10 10:30:00	0.0

Dienstleistung

Name	Witzpflicht
Versicherung	0
Tattoo	1
Haare	1

Firma

FID	IBranche	UStID	DBranche	Name	Umsatz
1		FI123456789	Versicherung	Hannoversche	245,390
2	Wein	FI987654321		Pahlgruber & Söhne	4,976,200
3	Haushaltsstahlwaren	FI567891234		Saugblaser Heinzelmann	119,970
4		FI246813579	Tattoo	VerkHAIRt	89,120
5		FI374287151	Tattoo	Stechpunkt	337,620
6	Haushaltsstahlwaren	FI842571678		Saugblaser Heinzelmann	3,245,390

gehoert_zu

Muttergesellschaft → gehoert zu	Tochtergesellschaft → gehoert zu
1	2
1	3
2	3

Industriell

Name → Firma	Organisation
Haushaltsstahlwaren	Matrix
Kupfer	Linie
Wein	Stab
Elektronik	Matrix

Produkt

PID	Name	Preis	Angeboten_von
1	DataShield	249.99	6
2	EcoPower Solaris	399.99	4
3	DataShield	19.99	1
4	RoboAssist	149.99	5
5	DataShield	79.99	4
6	Saugblaser Heinzelmann	199.95	3
7	Burgunder	12.23	2
8	Bordeaux	42.87	2

c)

Frage **ra-b2**
Falsch
Erreichte Punkte 0,00 von 2,00
Frage markieren

Wie viele Tupel liefert das Ergebnis der folgenden Anfrage? Sollte die Anfrage fehlerhaft sein, geben Sie im Antwortfeld -1 an.

Annahme: Operation auf **Mengen**

$\sigma_{Umsatz<500000}(Firma \times \pi_{Wirtschaftlichkeit}(Dienstleistung))$

Anzahl Tupel

d)

Frage **ra-b1**
Falsch
Erreichte Punkte 0,00 von 2,00
Frage markieren

Wie viele Attribute liefert das Ergebnis der folgenden Anfrage? Sollte die Anfrage fehlerhaft sein, geben Sie im Antwortfeld -1 an.

$\gamma_{PBranche,SUM(Umsatz)}(\delta((Produkt \bowtie_{PID=FID} Firma)))$

Anzahl Attribute

e)

Frage **ra-b3**
Falsch
Erreichte Punkte 0,00 von 2,00
Frage markieren

Wie viele Tupel liefert das Ergebnis der folgenden Anfrage? Sollte die Anfrage fehlerhaft sein, geben Sie im Antwortfeld -1 an.

Annahme: Operation auf **Mengen**

$\pi_{Name}(Produkt) \cap \pi_{Name}(Firma)$

Anzahl Tupel

f)

Frage **ra-aq**
Richtig
Erreichte Punkte 3,00 von 3,00
Frage markieren

Gegeben sei die folgende natürlichsprachliche Anfrage:

> Je Firma der durchschnittliche Preis von Bestellungen, die vor dem 8. Juni 2023 aufgegeben werden, abzüglich des Rabatts der Bestellungen.

Ist die folgende Anfrage in Relationaler Algebra zu diesem Ausdruck äquivalent?

$$\gamma_{FName,AVG(BPreis)\rightarrow Durchschnitt}(\pi_{FName,Preis*(1-Rabatt)\rightarrow BPreis}(\pi_{PID,Preis}(Produkt) \bowtie \sigma_{Datum<'2023-06-08'}(Bestellung) \bowtie \rho_{FID,IB,ID,DB,FName,U}(Firma)))$$

- Wahr
- Falsch

Lösung:

a)

Antworten wurden abgegeben.

Die Selektion reduziert die Anzahl der Tupel, nicht die Anzahl der Attribute.: True
Alle Operationen der relationalen Algebra sind auch in SQL direkt verfügbar.: False
Eine Projektion ist notwendig, um Relationen zu verknüpfen.: False
Die relationale Algebra arbeitet mengenbasiert.: True

b)

Die richtige Antwort lautet:
Weisen Sie den folgenden relationalen Operatoren die passenden Namen zu. Nutzen Sie die Drag & Drop-Funktion, um die Begriffe in die Lücken einzufügen.

Die Operation [Selektion] filtert Tupel aus einer Relation anhand einer Bedingung.

Mit der Operation [Projektion] werden nur bestimmte Attribute einer Relation ausgewählt.

Die Operation [Vereinigung] kombiniert alle Tupel zweier Relationen ohne Duplikate.

Bei der Operation [Differenz] erhält man nur die Tupel, die in der ersten Relation, aber nicht in der zweiten vorkommen.

Die Operation [Kreuzprodukt] erzeugt alle möglichen Kombinationen von Tupeln aus zwei Relationen.

Die Operation [Umbenennung] dient dazu, Relationen oder Attribute umzubenennen.

Die Operation [Natürlicher Join] verbindet Relationen automatisch über gemeinsame Attribute.

Die Operation [Division] gibt alle Tupel zurück, für die es zu jedem Tupel einer zweiten Relation eine passende Kombination gibt.

c)

Anzahl Attribute: -1
Es kann nur auf existierende Attribute gruppiert werden.

d)

Anzahl Tupel: 8

e)

Anzahl Tupel: 1

f)

.

Die richtige Antwort ist 'Wahr'.

6. DQL

Information

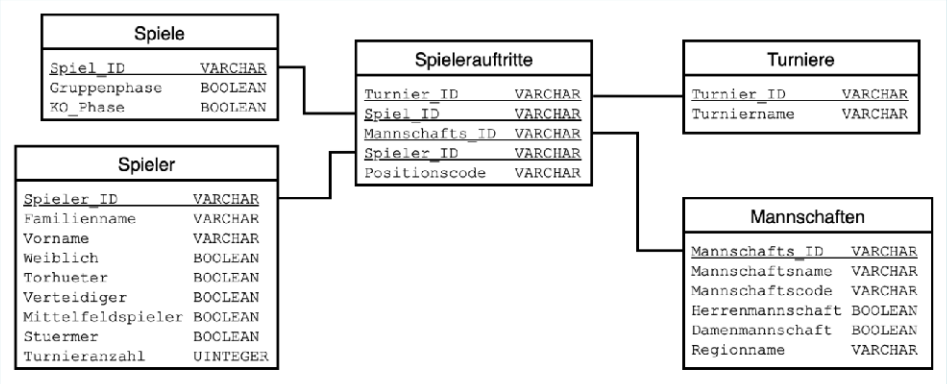
Frage markieren

In der folgenden Aufgabe werden Sie einen Datensatz behandeln, der Daten zu den bisherigen FIFA-Weltmeisterschaften enthält. Hierbei handelt es sich um eine leicht angepasste Version der Datenbank, die Sie bereits aus der Hausaufgabe kennen.

Alle Aufgaben sind mit dem SQL '92 Standard lösbar, aber Sie dürfen auch alle zusätzlichen Funktionen von DuckDB nutzen.

Hinweis: Sie finden das folgende Schema auch auf dem Notizzettel. **Beachten Sie, dass auf dem Notizzettel in „Spielerauftritte" fälschlicherweise keine Schlüsselattribute markiert wurden.**

Datenbank-Schema



a)

Frage **dql-tf1**

Richtig

Erreichte Punkte 1,00 von 1,00

Frage markieren

Beim vorliegenden Schema handelt es sich um...

Wählen Sie eine Antwort:

- ein Star Schema
- keine der vorliegenden Optionen
- ein Snowflake-Schema
- ein Fullfact Schema

b)

Frage **dql-tf2**

Richtig

Erreichte Punkte 1,00 von 1,00

Frage markieren

Welche der folgenden SQL-Anweisungen gibt alle Spalten von der Tabelle Turniere zurück?

True	False	
○	○	`SELECT * FROM TURNIERE;`
○	○	`SELECT * FROM Turniere;`
○	○	`SELECT ALL FROM Turniere;`
○	○	`SELECT Turniere FROM *;`

c)

Frage **dql-tf3**

Teilweise richtig

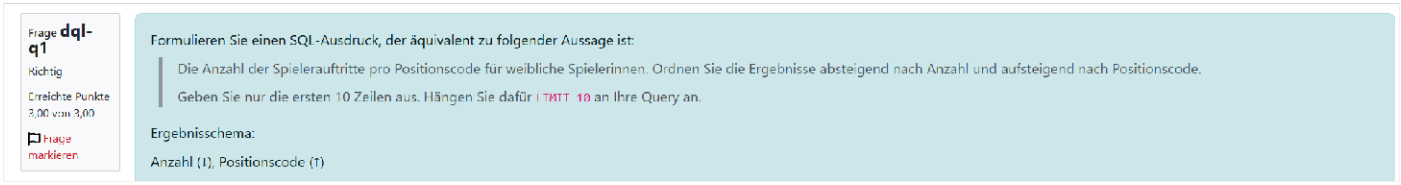
Erreichte Punkte 1,50 von 2,00

Frage markieren

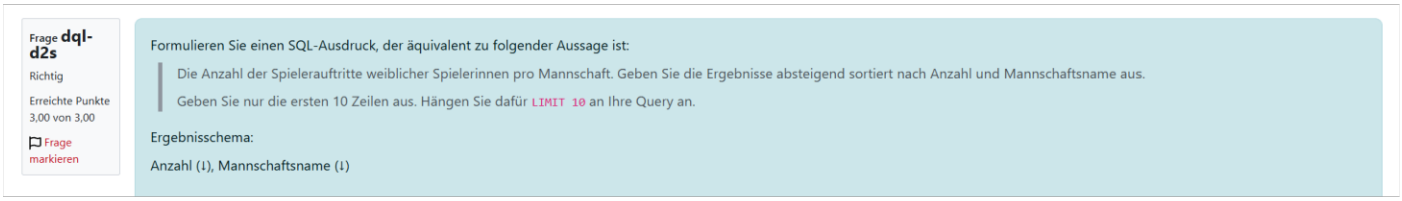
Beurteilen Sie, ob die folgenden Aussagen zutreffen.

True	False	
○	○	Es ist möglich, dass `SELECT count(*) FROM Spielerauftritte` mehr Zeilen ausgibt als `SELECT count(Positionscode) FROM Spielerauftritte`.
○	○	Die HAVING-Klausel kann in SQL auch ohne GROUP BY verwendet werden, um Bedingungen auf einzelne Zeilen anzuwenden.
○	○	Mit `SELECT count(*) FROM Turniere` wird die Anzahl aller Zeilen in der Turniere-Tabelle ausgegeben.
○	○	SQL ist eine deklarative Sprache.

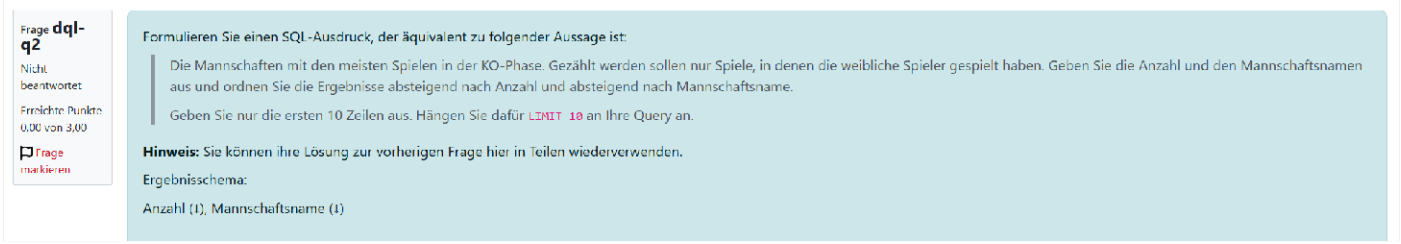
d)


Frage dql-q1
Richtig
Erreichte Punkte 3,00 von 3,00
Frage markieren
Formulieren Sie einen SQL-Ausdruck, der äquivalent zu folgender Aussage ist:
Die Anzahl der Spielerauftritte pro Positionscode für weibliche Spielerinnen. Ordnen Sie die Ergebnisse absteigend nach Anzahl und aufsteigend nach Positionscode.
Geben Sie nur die ersten 10 Zeilen aus. Hängen Sie dafür LIMIT 10 an Ihre Query an.
Ergebnisschema:
Anzahl (↓), Positionscode (↑)

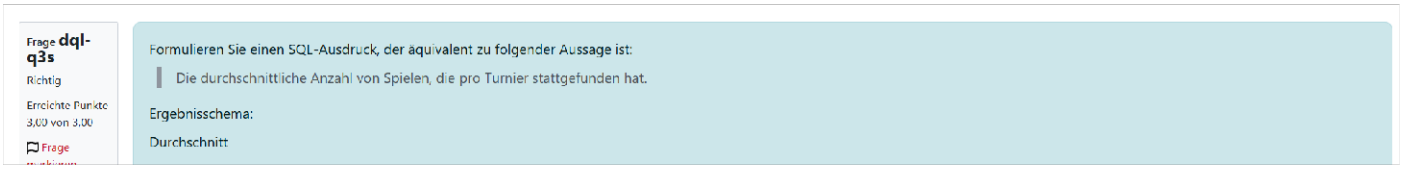
e)


Frage dql-d2s
Richtig
Erreichte Punkte 3,00 von 3,00
Frage markieren
Formulieren Sie einen SQL-Ausdruck, der äquivalent zu folgender Aussage ist:
Die Anzahl der Spielerauftritte weiblicher Spielerinnen pro Mannschaft. Geben Sie die Ergebnisse absteigend sortiert nach Anzahl und Mannschaftsname aus.
Geben Sie nur die ersten 10 Zeilen aus. Hängen Sie dafür LIMIT 10 an Ihre Query an.
Ergebnisschema:
Anzahl (↓), Mannschaftsname (↓)

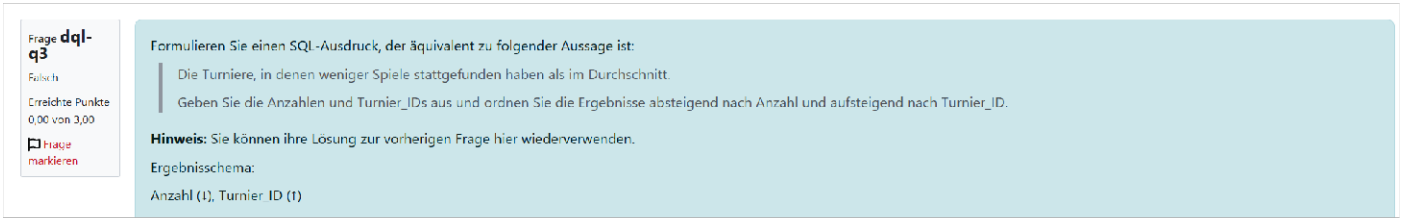
f)


Frage dql-q2
Nicht beantwortet
Erreichte Punkte 0,00 von 3,00
Frage markieren
Formulieren Sie einen SQL-Ausdruck, der äquivalent zu folgender Aussage ist:
Die Mannschaften mit den meisten Spielen in der KO-Phase. Gezählt werden sollen nur Spiele, in denen die weibliche Spieler gespielt haben. Geben Sie die Anzahl und den Mannschaftsnamen aus und ordnen Sie die Ergebnisse absteigend nach Anzahl und absteigend nach Mannschaftsname.
Geben Sie nur die ersten 10 Zeilen aus. Hängen Sie dafür LIMIT 10 an Ihre Query an.
Hinweis: Sie können ihre Lösung zur vorherigen Frage hier in Teilen wiederverwenden.
Ergebnisschema:
Anzahl (↓), Mannschaftsname (↓)

g)


Frage dql-q3s
Richtig
Erreichte Punkte 3,00 von 3,00
Frage
Formulieren Sie einen SQL-Ausdruck, der äquivalent zu folgender Aussage ist:
Die durchschnittliche Anzahl von Spielen, die pro Turnier stattgefunden hat.
Ergebnisschema:
Durchschnitt

h)


Frage dql-q3
Falsch
Erreichte Punkte 0,00 von 3,00
Frage markieren
Formulieren Sie einen SQL-Ausdruck, der äquivalent zu folgender Aussage ist:
Die Turniere, in denen weniger Spiele stattgefunden haben als im Durchschnitt.
Geben Sie die Anzahlen und Turnier_IDs aus und ordnen Sie die Ergebnisse absteigend nach Anzahl und aufsteigend nach Turnier_ID.
Hinweis: Sie können ihre Lösung zur vorherigen Frage hier wiederverwenden.
Ergebnisschema:
Anzahl (↓), Turnier_ID (↑)

Lösung:

a)

Your answer is correct.

Die richtige Antwort ist: ein Star Schema

b)

`SELECT * FROM TURNIERE;`: True
`SELECT * FROM Turniere;`: True
`SELECT ALL FROM Turniere;`: False
`SELECT Turniere FROM *;`: False

c)

Es ist möglich, dass `SELECT count(*) FROM Spieleraufritte` mehr Zeilen ausgibt als `SELECT count(Positionscode) FROM Spieleraufritte`.: True
Die HAVING-Klausel kann in SQL auch ohne GROUP BY verwendet werden, um Bedingungen auf einzelne Zeilen anzuwenden.: False
Mit `SELECT count(*) FROM Turniere` wird die Anzahl aller Zeilen in der Turniere Tabelle ausgegeben.: True
SQL ist eine deklarative Sprache.: True

d)

```sql
SELECT count(*) AS Anzahl, Positionscode
FROM Spieleraufritte
NATURAL JOIN Spieler
WHERE Weiblich = TRUE
GROUP BY Positionscode
ORDER BY Anzahl DESC, Positionscode ASC
LIMIT 10;
```

e)

```sql
SELECT count(*) as Anzahl, Mannschaftsname
FROM Spieleraufritte
NATURAL JOIN Mannschaften
NATURAL JOIN Spieler
WHERE Weiblich = TRUE
GROUP BY Mannschafts_ID, Mannschaftsname
ORDER BY Anzahl DESC, Mannschaftsname DESC
LIMIT 10;
```

f)

```sql
SELECT count(distinct Spiel_ID) as Anzahl, Mannschaftsname
FROM Spielerauftritte
NATURAL JOIN Mannschaften
NATURAL JOIN Spiele
NATURAL JOIN Spieler
WHERE KO_Phase = TRUE AND Weiblich = TRUE
GROUP BY Mannschafts_ID, Mannschaftsname
ORDER BY Anzahl DESC, Mannschaftsname DESC
LIMIT 10;
```

g)

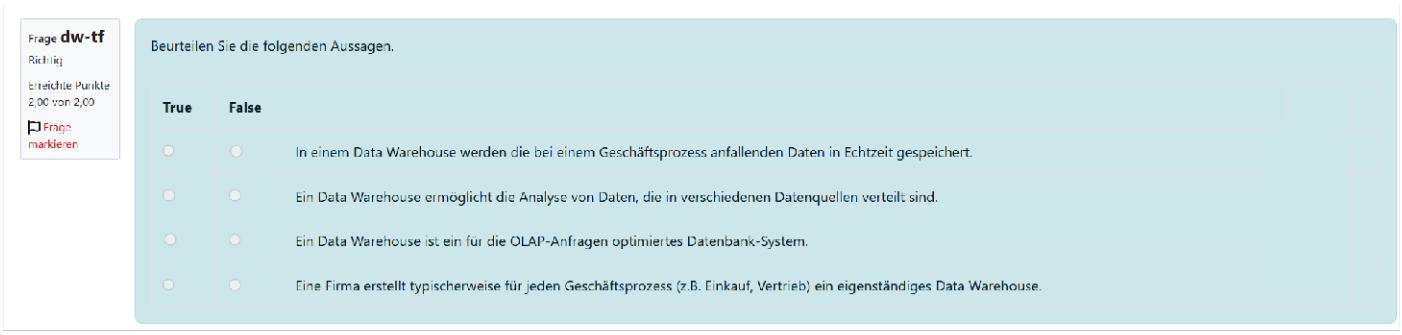
```sql
SELECT COUNT(DISTINCT Spiel_ID) / COUNT(DISTINCT Turniere.Turnier_ID) AS Durchschnitt
FROM Spielerauftritte JOIN Turniere ON Spielerauftritte.Turnier_ID = Turniere.Turnier_ID;
```

h)

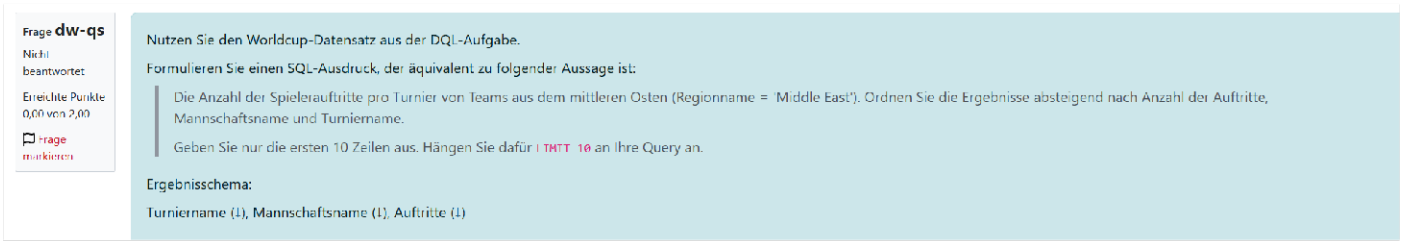
```sql
WITH games_per_tournament AS (SELECT Turnier_ID, count(DISTINCT Spiel_ID) AS Anzahl FROM Spielerauftritte
GROUP BY Turnier_ID),
avg_games AS (SELECT avg(Anzahl) As durchschnitt FROM games_per_tournament)
SELECT Anzahl, Turnier_ID
FROM games_per_tournament, avg_games
WHERE Anzahl < durchschnitt
ORDER BY Anzahl DESC, Turnier_ID ASC;
```

7. Data Warehousing

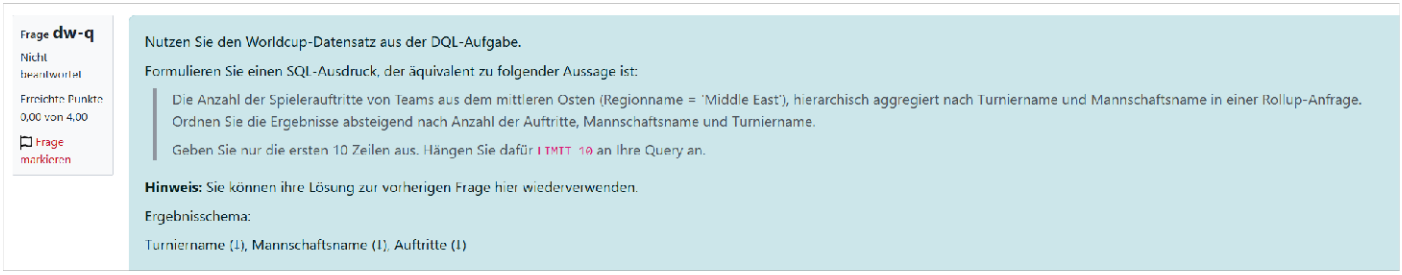
a)



b)



c)



Lösung:

a)

In einem Data Warehouse werden die bei einem Geschäftsprozess anfallenden Daten in Echtzeit gespeichert.: False
Ein Data Warehouse ermöglicht die Analyse von Daten, die in verschiedenen Datenquellen verteilt sind.: True
Ein Data Warehouse ist ein für die OLAP-Anfragen optimiertes Datenbank-System.: True
Eine Firma erstellt typischerweise für jeden Geschäftsprozess (z.B. Einkauf, Vertrieb) ein eigenständiges Data Warehouse.: False

b)

```sql
SELECT Turniername, Mannschaftsname, COUNT(Spielerauftritte) AS Auftritte
FROM Spielerauftritte
NATURAL JOIN Mannschaften
NATURAL JOIN Turniere
WHERE Regionname = 'Middle East'
GROUP BY Turniername, Mannschaftsname
ORDER BY Auftritte DESC, Mannschaftsname DESC, Turniername DESC
LIMIT 10;
```

c)

```sql
(
SELECT Turniername, Mannschaftsname, COUNT(Spielerauftritte) AS Auftritte
FROM Spielerauftritte
NATURAL JOIN Mannschaften
NATURAL JOIN Turniere
WHERE Regionname = 'Middle East'
GROUP BY Turniername, Mannschaftsname
)
UNION ALL
(
SELECT Turniername, NULL AS Mannschaftsname, COUNT(Spielerauftritte) AS Auftritte
FROM Spielerauftritte
NATURAL JOIN Mannschaften
NATURAL JOIN Turniere
WHERE Regionname = 'Middle East'
GROUP BY Turniername
)
UNION ALL
(
SELECT NULL AS Turniername, NULL AS Mannschaftsname, COUNT(Spielerauftritte) AS Auftritte
FROM Spielerauftritte
NATURAL JOIN Mannschaften
NATURAL JOIN Turniere
WHERE Regionname = 'Middle East'
)
ORDER BY Auftritte DESC, Mannschaftsname DESC, Turniername DESC
LIMIT 10;
```

8. Data Streams

a)

Frage **str-win**

Richtig

Erreichte Punkte 2,00 von 2,00

Frage markieren

In einer Anfrage werden zeit-basierte Sliding Windows mit einer Fensterlänge von 5 Zeiteinheiten und einer Verschiebung (Slide) aller 2 Zeiteinheiten berechnet.

Berechnen Sie die Summe des 3. Fensters über den gegebenen Datenstrom:

Zeit:	**0**	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**16**	**18**
Tupelwert:	6	4	3	9	8	7	4	5	3	8

Hinweis: Das Tupel mit dem Wert 6 wurde zum Zeitpunkt 0 erstellt.

Antwort:

b)

Frage **str-cm**

Teilweise richtig

Erreichte Punkte von 4,00

Frage markieren

Um die Sicherheit des TU-Netzes zu erhöhen setzt die tubIT neuerdings Count-Min Sketche zur verbesserten Spam-Erkennung für E-Mails ein.

Hierfür nutzt der Count-Min Sketch zwei Hash Funktionen

$$h1(n)$$

und

$$h2(n)$$

$$h1(n) = (n+1) \mod 3$$

$$h2(n) = (n-1) \mod 3$$

Über einen Tag werden folgende dreistellige Martikelnummern aufgezeichnet und in den Count-Min-Sketch eingefügt:

1. 054
2. 065
3. 076
4. 011
5. 002
6. 035

Max. Punkte	**Frage**	**Antwort**
(0.5)	Anhand der gegebenen Information: Wie viele Zeilen sollte der Count-Min Sketch haben?	
(0.5)	Anhand der gegebenen Information: Wie viele Spalten sollte der Count-Min Sketch haben?	
(1.0)	Wie hoch ist die Summe aller Zellen einer Zeile des Count-Min Sketches, nachdem alle angegebenen Elemente eingefügt wurden?	
(2.0)	Was ist die geschätzte Häufigkeit für das Vorkommen der Matrikelnummer 057?	

Information

Frage markieren

Verkehrssteuerung auf Autobahnen (Data Streams)

Informationen zum Eingabestrom

In den folgenden Aufgaben verwenden wir einen Datenstrom zur intelligenten Verkehrssteuerung auf Autobahnen. Dieser Datenstrom ist ein TimedStream, bei dem jedes Element ein vom System erkanntes Fahrzeug repräsentiert. Der Timestamp der Elemente ist als Unixzeit in Sekunden codiert. Die Stream-Elemente sind als Python Tuple mit folgender Struktur implementiert:

`(lane, velocity, type, brand)`

- **lane**: Autobahnspur, auf der das Fahrzeug fährt. Diese sind nummeriert von 1 bis 3 (z.B. 1 für rechte Fahrspur).
- **velocity**: Gemessene Geschwindigkeit des Fahrzeugs in km/h.
- **type**: Art des Fahrzeugs (lkw oder pkw).
- **brand**: Marke des Fahrzeugs (z.B. BMW, Audi, ...).

Ihre Aufgabe ist es, Dataflow-Pipelines zu implementieren, die auf diesem Datenstrom basieren und verschiedene Anfragen beantworten. Zur Lösung stehen Ihnen alle nötige Funktionen erster Ordnung und die folgende Liste der Funktionen der ISDA Streaming API zur Verfügung:

DataStream

- map
- flat_map
- filter
- reduce
- key_by
- landmark_window
- tumbling_window
- sliding_window

KeyedStream

- map
- flat_map
- filter
- reduce

TimedStream

- map
- flat_map
- filter
- reduce
- key_by
- landmark_tuple_window
- tumbling_tuple_window
- sliding_tuple_window
- landmark_time_window
- tumbling_time_window
- sliding_time_window

WindowedStream

- aggregate
- reduce
- apply

Hinweise: Sie finden die gesamte Aufgabenbeschreibung auch ausgedruckt auf den Notizzetteln.

c)

Frage **str-q1**

Falsch

Erreichte Punkte 0.00 von 3.00

Frage markieren

Implementieren Sie die Funktion `pkw_max_velocity_per_lane`, die eine Dataflow-Pipeline mit folgendem Ergebnis erstellt:

Die laufende maximale Geschwindigkeit eines PKWs pro Autobahnspur.

Erwartete Elementstruktur pro Autobahnspur (`lane`) im Ausgabestrom: `rolling_pkw_max_velocity`

Code-Vorgabe

Falls Sie von Beginn an anfangen möchten, können Sie die folgende Code-Vorgabe verwenden:

▶ Code-Vorgabe

Antwort: (Abzugssystem: 0 %)

Antwort zurücksetzen

```
######## Code-Vorgabe ########
def is_pkw(stream_element: tuple) -> bool:
    if stream_element[2] == "pkw":
        return True
    return False

def get_lane(stream_element: tuple) -> str:
    return stream_element[0]

def gat_velocity(stream_element: tuple) -> float:
    return stream_element[1]

def max_velocity(velocity_1: float, velocity_2: float) -> float:
    if velocity_1 > velocity_2:
        return velocity_1
    return velocity_2

######## Ende der Code-Vorgabe ########

def pkw_max_velocity_per_lane(input_stream: TimedStream):
    return (input_stream
            # HIER! Dataflow-Pipeline ergänzen
        )

```

d)

Frage **str-q2**

Falsch

Erreichte Punkte 0.00 von 3.00

Frage markieren

Implementieren Sie die Funktion `lane_2_mean_velocity`, die eine Dataflow-Pipeline mit folgendem Ergebnis erstellt:

Die durchschnittlichen Geschwindigkeiten (`mean_velocity`) aus den letzten 10 Autos auf der mittleren Fahrspur. Diese Statistik soll alle 5 Autos aktualisiert werden.

Erwartete Elementstruktur im Ausgabestrom: `(mean_velocity, window_start, window_end)`

Code-Vorgabe

Falls Sie von Beginn an anfangen möchten, können Sie die folgende Code-Vorgabe verwenden:

▶ Code-Vorgabe

Antwort: (Abzugssystem: 0 %)

Antwort zurücksetzen

```
######## Code-Vorgabe ########
def is_lane_2(stream_element: tuple) -> bool:
    if stream_element[0] == 2:
        return True
    return False

def get_velocity(stream_element: tuple)  > float:
    return stream_element[1]

######## Ende der Code-Vorgabe ########

def lane_2_mean_velocity(input_stream: TimedStream) -> DataStream:
    return (input_stream
        # Hier Dataflow-Pipeline ergänzen
    )

```

Lösung:

a)

Das dritte Fenster: 3+9+8 = 20
Die richtige Antwort ist: 20

b)

Der Count Min Sketch hat folgende Werte:

	0	1	2
h1	4	1	1
h2	1	4	1

c)

Keine Lösung gegeben

d)

```python
# Filter
def is_lane_2(stream_element: tuple) -> bool:
    if stream_element[0] == 2:
        return True
    return False

# Map
def get_velocity(stream_element: tuple) -> float:
    return stream_element[1]

def lane_2_mean_velocity(input_stream: TimedStream) -> DataStream:
    return (
        input_stream
        .filter(is_lane_2)
        .map(get_velocity)
        .sliding_tuple_window(10, 5)
        .aggregate("mean")
    )
```

9. Data Science (Fallstudie)

Information

Frage markieren

Das fiktive Unternehmen DocuGenAI plant den Aufbau eines Retrieval-Augmented Generation (RAG) Systems, bei dem Dokumente anhand ihrer Inhalte gesucht und in ein Sprachmodell eingespeist werden, um bessere Antworten für Nutzerfragen zu generieren. Im Rahmen dieser Aufgabe sollen Sie das Team bei verschiedenen Fragen zur Modellwahl, Trainingsstrategie und Evaluation unterstützen.

a)

Frage **rag-01**

Teilweise richtig

Erreichte Punkte 0,50 von 1,00

Frage markieren

Das Team trainiert ein Modell zur Vorhersage der Relevanz von Dokumenten auf der Grundlage historischer Benutzerinteraktionen. Nach dem Einsatz stellt das Team fest, dass das Modell bei den Trainingsdaten gut, bei neuen Dokumenten jedoch schlecht abschneidet.

Beurteilen Sie den Wahrheitsgehalt der folgenden Aussagen über das trainierte Modell.

True	False	
○	○	Das Modell hat einen hohen Bias.
○	○	Das Modell ist overfitted, da es auf Trainingsdaten gut, auf Testdaten aber schlecht performt.
○	○	Ein kleiner Trainingsdatensatz schließt Overfitting aus.
○	○	Die Testdaten müssen falsch gelabelt sein.

b)

Frage **rag-02**

Teilweise richtig

Erreichte Punkte 0,50 von 1,00

Frage markieren

Das Team fragt sich, welche Strategien helfen können, Overfitting zu reduzieren. Beantworten Sie die folgenden Fragen.

True	False	
○	○	Komplexere Modelle mit mehr Parametern sind besser gegen Overfitting geschützt.
○	○	Man sollte Kreuzvalidierung vermeiden, um Rechenzeit zu sparen.
○	○	Mehr Trainings-Epochen führen immer automatisch zu besserer Generalisierung.
○	○	Mehr hochwertige und ausgewogene Trainingsdaten zu sammeln ist eine gute Methode gegen Overfitting.

c)

Frage **rag-03**

Teilweise richtig

Erreichte Punkte 0,50 von 1,00

Frage markieren

Welche der folgenden Ansätze eignen sich zur realistischen Bewertung der Generalisierungsfähigkeit eines Modells?

True	False	
○	○	Um die Leistung des Modells zu verbessern, sollten die Testdaten mit in das Training einbezogen werden.
○	○	Eine Konfusionsmatrix auf den Trainingsdaten zeigt immer hinreichende Informationen über die Modellqualität.
○	○	K-fache Kreuzvalidierung auf den Trainingsdaten ist eine gute Möglichkeit.
○	○	Der Fehler auf den Trainingsdaten reicht für gewöhnlich aus, um die Modellqualität zu bewerten.

Information

Frage markieren

Um relevante Dokumente für eine gegebene Anfrage zu identifizieren, möchte das Team einen k-Nearest-Neighbor-Ansatz auf der Grundlage von Vektor-Embeddings mit der euklidischen Distanz verwenden.

Gegeben sind die folgenden vereinfachten Dokument-Embeddings und eine Anfrage:

Dokument	Vektor-Embedding	Kategorie
D1	(1, 2, 1)	Recht
D2	(3, 1, 2)	Medizin
D3	(4, 4, 1)	Technik
D4	(2, 3, 2)	Medizin
D5	(0, 1, 0)	Recht
Anfrage Q	(2, 2, 1)	?

d)

Frage **rag-05**

Richtig

Erreichte Punkte 2,00 von 2,00

Frage markieren

Berechnen Sie die euklidische Distanz der Dokumente zur Anfrage. Welche drei Dokumente sind der Query Q am nächsten?

Schreiben Sie die ID der Dokumente Komma-separiert und in geschweiften Klammern in das Antwortfeld (z.B. `{D1, D2, D3}`, angezeigt als $\{D_1, D_2, D_3\}$).

Ihre letzte Antwort wurde folgendermaßen interpretiert:

e)

Frage **rag-06**

Falsch

Erreichte Punkte 0,00 von 1,00

Frage markieren

Jedes Dokument hat eine zugehörige Kategorie. Welches Label würde der kNN-Algorithmus Q für k = 3 zuweisen?

Wählen Sie eine Antwort:

- Es kann keine Klassifikation durchgeführt werden, da k kleiner als die Anzahl der Dokumente ist.
- Technik
- Medizin
- Q würde kein Label erhalten, da es keine Mehrheit gibt.
- Recht

f)

Frage **rag-07**

Falsch

Erreichte Punkte 0,00 von 1,00

Frage markieren

Wie würde sich die Klassifikation bei k = 5 ändern?

Wählen Sie eine Antwort:

- Die Anfrage schlägt fehlt, da k gleich der Anzahl der Dokumente ist.
- Zwei Kategorien haben gleich viele Nachbarn und das Team muss entschieden, wie ein Unentschieden behandelt wird.
- Q würde das Label "Recht" erhalten.
- Q würde das Label "Medizin" erhalten.
- Q würde das Label "Technik" erhalten.

g)

Frage **rag-08**

Falsch

Erreichte Punkte 0,00 von 1,00

Frage markieren

Welche Aussage beschreibt den Einfluss der Wahl von k auf Bias und Varianz am besten?

Wählen Sie eine Antwort:

- Die Wahl von k hat keinerlei Einfluss auf Bias und Varianz des Modells.
- Ein großer Wert für k reduziert sowohl Bias als auch Varianz.
- Ein kleiner Wert für k führt zu niedrigerem Bias, aber höherer Varianz.
- Ein kleiner Wert für k führt zu hohem Bias und niedriger Varianz.

Lösung:

a)

Das Modell hat einen hohen Bias.: False
Das Modell ist overfitted, da es auf Trainingsdaten gut, auf Testdaten aber schlecht performt.: True
Ein kleiner Trainingsdatensatz schließt Overfitting aus.: False
Die Testdaten müssen falsch gelabelt sein.: False

b)

Komplexere Modelle mit mehr Parametern sind besser gegen Overfitting geschützt.: False
Man sollte Kreuzvalidierung vermeiden, um Rechenzeit zu sparen.: False
Mehr Trainings-Epochen führen immer automatisch zu besserer Generalisierung.: False
Mehr hochwertige und ausgewogene Trainingsdaten zu sammeln ist eine gute Methode gegen Overfitting.: True

c)

Um die Leistung des Modells zu verbessern, sollten die Testdaten mit in das Training einbezogen werden.: False
Eine Konfusionsmatrix auf den Trainingsdaten zeigt immer hinreichende Informationen über die Modellqualität.: False
K-fache Kreuzvalidierung auf den Trainingsdaten ist eine gute Möglichkeit.: True
Der Fehler auf den Trainingsdaten reicht für gewöhnlich aus, um die Modellqualität zu bewerten.: False

d)

Eine richtige Antwort ist $\{D_1, D_2, D_4\}$. Sie kann so eingegeben werden: `{D1,D2,D4}`

e)

Your answer is incorrect.

Die richtige Antwort ist: Medizin

f)

Your answer is incorrect.

Die richtige Antwort ist: Zwei Kategorien haben gleich viele Nachbarn und das learn muss entschieden, wie ein Unentschieden behandelt wird.

g)

Your answer is incorrect.

Die richtige Antwort ist: Ein kleiner Wert für k führt zu niedrigerem Bias, aber höherer Varianz.

10. Data Science (Allgemein)

a)

Frage **ds-dat**
Richtig
Erreichte Punkte 0,50 von 0,50
Frage markieren

Welche der folgenden Aussagen sind wahr?

True	False	
○	○	Diskrete Daten wie zum Beispiel das Alter in Jahren sind quantitative Daten.
○	○	Bei einer Text-Datei handelt es sich um semi-strukturierte Daten.

b)

Frage **ds-dst**
Teilweise richtig
Erreichte Punkte 0,50 von 1,00
Frage markieren

Weisen Sie den folgenden Anwendungsfällen eine passende Distanzfunktion zu. Nutzen Sie die Drag & Drop Funktion, um die Begriffe in die Lücken einzufügen.

Anwendungsfall	Distanzfunktion
Erkennnung von Übertragungsfehlern in Bit Streams	
Erkennung von Betrugsversuchen in Programmieraufgaben	
Ähnlichkeit von Vektor-Embeddings	
Ähnlichkeit von Mengen	

c)

Frage **ds-dl**
Richtig
Erreichte Punkte 1,00 von 1,00
Frage markieren

Berechnen Sie die Levenshtein Distanz zwischen den Wörtern "Lernen" und "Lehrer".

Antwort:

d)

Frage **ds-km**
Falsch
Erreichte Punkte 0,00 von 0,75
Frage markieren

Welche der folgenden Aussagen über k-Means sind wahr?

True	False	
○	○	Um den besten Wert für den Paramter k zu bestimmen, muss man k-Means lediglich ein Mal pro möglichem Wert für k ausführen.
○	○	Die kumulative Distanz der Punkte zum nächsten Centroid hängt von dem anfangs festgelegten Parameter k ab.
○	○	Die kumulative Distanz der Punkte zum nächsten Centroid kann sich im Laufe des Optimierungsprozesses verschlechtern.

e)

Frage **ds-cent**
Nicht bewertet
Erreichbare Punkte: 2,00
Frage markieren

Gegeben ist ein eindimensionales Datenset D und zwei initiale Centroids C_1 und C_2:

$$D = \begin{bmatrix} 3 \\ 0 \\ 9 \\ 2 \\ 12 \\ 13 \\ 0 \end{bmatrix} \quad C_1 = |11| \quad C_2 = |14|$$

Führen Sie **drei** Iterationen von K-Means durch. Welchen Wert (auf **zwei** Dezimalstellen gerundet) haben die zwei Centroids nach diesen Iterationen?

$C_1 =$

C_2

f)

Frage **ds-spam**

Nicht beantwortet

Erreichbare Punkte: 3,00

Frage markieren

Ein E-Mail-Anbieter nutzt ein Klassifikationsmodell, um Spam zu identifizieren. In der folgenden Tabelle sehen Sie die Spam-Bewertung einer Nutzerin und die des Spamfilters für jede E-Mail eines Beispiel Datensatzes.

Betreff	Bewertung Nutzerin	Bewertung Spamfilter
Happy27 voucher	kein Spam	Spam
Klausurlö sungen online	kein Spam	Spam
1M Dollar warten für Sie	Spam	kein Spam
Komm in meine WhatsApp-Gruppe	Spam	kein Spam
Döner auf dem Campus	kein Spam	Spam
Ihre Erbschaft wartet auf Sie	Spam	Spam
Kursanmeldung DBPRA	kein Spam	kein Spam
Kritische Sicherheitslpcke auf Ihrem PC	Spam	Spam

Evaluieren Sie den Spamfilter anhand der Bewertungen der Nutzerin auf **zwei** Nachkommastellen genau. Alternativ dürfen Sie Ihre Antwort auch als Bruch angeben, z.B. `1/3` statt `0,33`.

Hinweis: "Spam" ist die positive, "kein Spam" die negative Klassifizierung.

Präzision

Was ist die Präzision des Spamfilters?

Recall

Was ist der Recall (die Sensitivität) des Spamfilters?

F1-Score

Was ist der F1-Score des Spamfilters?

g)

Frage **ds-regpy**

Teilweise richtig

Erreichte Punkte 1,00 von 2,00

Frage markieren

Ein Kollege von Ihnen hat folgende Pipeline für das Training und die Evaluation eines Regressionsmodells erstellt:

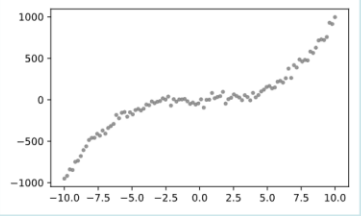
```
X, y = load_data()

low_degree = 0
mid_degree = 2
high_degree = 10

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25)
scores = []
for degree in [low_degree, mid_degree, high_degree]:
    model = create_regression_model(polynomial_feature_degree=degree)  # Modell mit polynomiellen Features erstellen
    model.fit(X_train, y_test)                                         # Modell trainieren
    y_pred = model.predict(X_test)                                     # Vorhersagen berechnen
    scores.append(mean_squared_error(y_test, y_pred))                  # Modellqualität berechnen

best_score = max(scores)
```

Darüber hinaus hat er die von `load_data()` geladenen Daten für eine explorative Datenanalyse visualisiert:



Welche der folgenden Aussagen sind wahr?

True	False	
○	○	Train- und Test-Split werden im For-Loop korrekt verwendet.
○	○	`best_score` enthält den Mean Squared Error des besten Modells.
○	○	Die Größe des Test-Splits ist angemessen gewählt.
○	○	Der Wert von `mid_degree` ist so hoch wie nötig und so niedrig wie möglich gewählt, um Under- und Overfitting zu vermeiden.

Lösung:

a)

Diskrete Daten wie zum Beispiel das Alter in Jahren sind quantitative Daten.: True
Bei einer Text-Datei handelt es sich um semi-strukturierte Daten.: False

b)

Sie haben 2 richtig ausgewählt.
Die richtige Antwort lautet:
Weisen Sie den folgenden Anwendungsfällen eine passende Distanzfunktion zu. Nutzen Sie die Drag & Drop-Funktion, um die Begriffe in die Lücken einzufügen.

Anwendungsfall	Distanzfunktion
Erkennung von Übertragungsfehlern in Bit-Streams	[Hamming Distanz]
Erkennung von Rechtschreibversuchen in Programmieraufgaben	[Levenshtein Distanz]
Ähnlichkeit von Vektor Embeddings	[Euklidische Distanz]
Ähnlichkeit von Mengen	[Jaccard Distanz]

c)

Die richtige Antwort ist: 3

d)

Um den besten Wert für den Parameter k zu bestimmen, muss man k-Means lediglich ein Mal pro möglichem Wert für k ausführen.: False
Die kumulative Distanz der Punkte zum nächsten Centroid hängt von dem anfangs festgelegten Parameter k ab.: True
Die kumulative Distanz der Punkte zum nächsten Centroid kann sich im Laufe des Optimierungsprozesses verschlechtern.: False

e)

Keine Lösung gegeben

f)

Eine richtige Antwort ist 0,4. Sie kann so eingegeben werden: `0,4`
Eine richtige Antwort ist 0,5. Sie kann so eingegeben werden: `0,5`
Eine richtige Antwort ist $\frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}}$. Sie kann so eingegeben werden: `2/(1/received_recall+1/received_precision)`

g)

Train- und Test-Split werden im For-Loop korrekt verwendet.: False
`best_score` enthält den Mean Squared Error des besten Modells.: False
Die Größe des Test-Splits ist angemessen gewählt.: True
Der Wert von `mid degree` ist so hoch wie nötig und so niedrig wie möglich gewählt, um Under- und Overfitting zu vermeiden.: False