

Familienname, Vorname (bitte durch eine Leerspalte trennen)

[illegible]

Bereich		Berufsnummer				IHK-Nummer			Prüflingsnummer			
7	3	1	2	0	3							
Sp. 1-2		Sp. 3-6				Sp. 7-9			Sp. 10-14			

**Termin: Mittwoch, 27. November 2024**



# IHK

Abschlussprüfung Winter 2024/25  
1203

## 2 Sicherstellen der Datenqualität

Fachinformatiker  
Fachinformatikerin  
Daten- und Prozessanalyse

## Teil 2 der Abschlussprüfung

## 4 Aufgaben

90 Minuten Prüfungszeit

100 Punkte

## Bearbeitungshinweise

1. Bevor Sie mit der Bearbeitung der Aufgaben beginnen, überprüfen Sie bitte die **Vollständigkeit** dieses Aufgabensatzes. Die Anzahl der zu bearbeitenden Aufgaben ist auf dem Deckblatt links angegeben. Wenden Sie sich bei Unstimmigkeiten sofort an die Aufsicht, weil Reklamationen am Ende der Prüfung nicht anerkannt werden können.
2. Füllen Sie zuerst die **Kopfzeile** aus. Tragen Sie Ihren Familiennamen, Ihren Vornamen und Ihre Prüflings-Nr. in die oben stehenden Felder ein.
3. Lesen Sie bitte den **Text** der Aufgaben ganz durch, bevor Sie mit der Bearbeitung beginnen.
4. Halten Sie sich bei der Bearbeitung der Aufgaben genau an die **Vorgaben der Aufgabenstellung** zum Umfang der Lösung. Wenn z. B. vier Angaben gefordert werden und Sie sechs Angaben anführen, werden nur die ersten vier Angaben bewertet.
5. Tragen Sie die frei zu formulierenden **Antworten dieser offenen Aufgaben** in die dafür lt. Aufgabenstellung vorgesehenen Bereiche (Lösungszeilen, Formulare, Tabellen u. a.) des Arbeitsbogens ein.
6. Sofern nicht ausdrücklich ein Brief oder eine Formulierung in ganzen Sätzen gefordert werden, ist eine **stichwortartige Beantwortung** zulässig.
7. Schreiben Sie deutlich und gut lesbar. Ein nicht eindeutig zuzuordnendes oder **unleserliches Ergebnis** wird als **falsch** gewertet.
8. Zur Lösung der Rechenaufgaben darf ein nicht programmierter, netzunabhängiger **Taschenrechner** ohne Kommunikationsmöglichkeit mit Dritten verwendet werden.
9. Wenn Sie ein **gerundetes Ergebnis** eintragen und damit weiterrechnen müssen, rechnen Sie (auch im Taschenrechner) nur mit diesem gerundeten Ergebnis weiter.
10. Für **Hilfsaufzeichnungen** können Sie das in der Tasche beigelegte Konzeptpapier verwenden. Bewertet werden jedoch grundsätzlich nur Ihre Eintragungen in diesem Aufgabensatz.

**Wird vom Korrektor ausgefüllt!**

## Bewertung

Für die Bewertung gilt die Vorgabe der Punkte in den Lösungshinweisen.

1. Aufg. 

--	--

 Punkte    2. Aufg. 

--	--

 Punkte    3. Aufg. 

--	--

 Punkte    4. Aufg. 

--	--

 Punkte

15   16                      17   18                      19   20                      21   22

Prüfungszeit

23

Die entsprechende Ziffer (1, 2 oder 3) finden Sie in der Abfrage nach der Prüfungszeit im Anschluss an die letzte Aufgabe.

**Gesamtpunktzahl**

24	25	26

Prüfungsort, Datum

Unterschrift

Gemeinsame Prüfungsaufgaben der Industrie- und Handelskammern. Dieser Aufgabensatz wurde von einem überregionalen Ausschuss, der entsprechend § 40 Berufsbildungsgesetz zusammengesetzt ist, beschlossen. Hinweis: Im Interesse einer besseren Lesbarkeit wird in der Aufgabenstellung und in den Angaben zur Aufgabenstellung nur die männliche Form (generisches Maskulinum) verwendet. Die verkürzte Sprachform beinhaltet keine Wertung und die gewählten männlichen Formulierungen gelten uneingeschränkt auch für die weiteren Geschlechter. Die Vervielfältigung, Verbreitung und öffentliche Wiedergabe der Prüfungsaufgaben und Lösungen ist nicht gestattet. Zuwiderhandlungen werden zivil- und strafrechtlich (§§ 97 ff., 106 ff. UrhG) verfolgt. – © ZPA Nord-West 2024 – Alle Rechte vorbehalten!

**Die Aufgabe 1 bis 4 beziehen sich auf die folgende Ausgangssituation:**

Die AMAG Soft GmbH ist ein mittelständisches Unternehmen mit derzeit 130 Mitarbeitern.

Das Kerngeschäft basiert auf drei Standbeinen:

- Entwicklung, Produktion und Vertrieb von mobilen Kompressoren, drucklufttechnischen Anlagen und Zubehör
- Service und Support für drucklufttechnische Anlagen
- Handel mit Druckluft

Das zentrale Geschäftsmodell besteht darin, die drucklufttechnischen Anlagen beim Kunden aufzustellen, um vor Ort die benötigte Druckluft zu erzeugen. Über Service-Level-Agreements gewährleistet die AMAG Soft GmbH die Verfügbarkeit, Leistung und Qualität.

**1. Aufgabe (25 Punkte)**

E-Mail erhalten	
Von	Mira Bellenbaum <bellenbaum@amagsoftgmbh.de>
An	pruefung@amagsoftgmbh.de
Betreff	Daten der Druckluftanlagen fehlerhaft?
Inhalt	<p>Hallo,</p> <p>von den drucklufttechnischen Anlagen werden Daten zur Ermittlung der Ausfallwahrscheinlichkeit einzelner Maschinen erhoben und in einer CSV-Datei zusammengeführt. Die Daten sind aber häufig fehlerhaft. Bitte führen Sie eine Überprüfung der Datenqualität durch.</p> <p>Liebe Grüße Mira Bellenbaum</p>

Per E-Mail-Anhang hat sie Ihnen einen Auszug aus den erhobenen Daten geschickt:

Maschinen-ID	Qualität	Prozesstemperatur [K]	Drehgeschwindigkeit (rpm)	Drehmoment [Nm]	Einsatzzeit [min]	Druck [bar]
M15008	M	308,3	1379	48	181	9,9
L47329	L	35,3	1473	39,9	184	9,8
L47330	L	308,3	1422	42,7	186	9,9
L47330	L	308,3	1422	42,7	186	9,9
M15011	M	308,2	1463	37,6	188	9,7
M15012	M	308,2	1584	41	191	9,6
47333	L	308,2	1850	27	194	9,8
L47334	L	308,2	1528	36,2	null	9,8
H29569	H	308,2	1987	19,8	198	9,9
M15016	M	308,1	1495	46	203	9,7

**Erklärungen**

- 308 Kelvin ~ 35 °Celsius
- rpm = engl. revolutions per minute; „Umdrehungen pro Minute“
- Qualität der Maschine, klassifiziert als L = Low, M = Medium, H = High

a) Für den Ablauf der Datenaufbereitung wird der ETL-Prozess verwendet.

Korrekturrand

Beschreiben Sie kurz die drei Schritte mithilfe des gegebenen englischen Textes.:

6 Punkte

In computing, extract, transform, load (ETL) is a three-phase process where data is *extracted* from an input source, *transformed* (including cleaning), and *loaded* into an output data container. The data can be collated from one or more sources and it can also be output to one or more destinations. (From Wikipedia, the free encyclopedia)

Extraktion:

---

---

---

Transformation:

---

---

---

Laden:

---

---

---

---

---

b) Überprüfen Sie die Stichprobe auf mögliche Probleme in Bezug auf Datenqualität. Nennen Sie mindestens drei konkrete Probleme aus der Stichprobe.

3 Punkte

---

---

---

---

---

c) Beschreiben Sie Datenvalidierung und Datenverifizierung.

Nennen Sie jeweils ein Beispiel aus der Stichprobe für jede Art.

6 Punkte

Datenvalidierung:

---

---

---

---

---

Datenverifizierung:

---

---

---

---

---

### Fortsetzung 1. Aufgabe

Korrekturrand

- d) Schreiben Sie Frau Bellenbaum zwei Einschätzungen zur Datenqualität per E-Mail zu der Stichprobe. Gehen Sie dabei auf Ihre Erkenntnisse aus b) ein und beschreiben Sie zwei Maßnahmen, die ergriffen werden können, um die erkannten Datenprobleme zu bereinigen und die Datenqualität zu verbessern.

10 Punkte

Neue E-Mail schreiben	
Von	pruefung@amagsoftgmbh.de
An	Mira Bellenbaum <bellenbaum@amagsoftgmbh.de>
Betreff	
Inhalt	

## 2. Aufgabe (28 Punkte)

Korrekturrand

Sie bekommen den Auftrag, die Ausfälle der Kompressoren zu analysieren, damit in Zukunft auffällige Maschinen gewartet werden, bevor sie ausfallen. Die Daten liegen in einer CSV-Datei vor. Die 1 in der Ergebnisspalte sagt aus, dass diese Maschine ausgefallen ist. 0 bedeutet keinen Ausfall.

Maschinen-ID	Qualität	Prozesstemperatur [K]	Drehgeschwindigkeit (rpm)	Drehmoment [Nm]	Einsatzzeit [min]	Ergebnis
M15008	M	308,3	1379	48	181	0
L47329	L	308,3	1473	39,9	184	0
L47330	L	308,3	1422	42,7	186	0
M15011	M	308,2	1463	37,6	188	0
M15012	M	308,2	1584	41	191	0
L47333	L	308,2	1850	27	194	0
L47334	L	308,2	1528	36,2	196	0
H29569	H	308,2	1987	19,8	198	0
M15016	M	308,1	1495	46	203	0
1H29571	H	308,2	1478	43,5	206	0
M15018	M	308,2	1499	40	211	0
L47339	L	308,2	1471	47	214	0
147340	L	308,2	1282	60,7	216	1
L47341	L	308,1	1412	52,3	218	1
147342	L	308,1	1586	35,5	0	0
M15023	M	308,2	1468	43,8	2	0
M15024	M	308,3	1440	43,6	5	0
147345	L	308,4	1733	26	8	0
H29580	H	308,2	1612	32,6	10	0
H29581	H	308,3	1458	42,9	15	0
L47348	L	308,3	1433	62,3	20	1
L47349	L	308,3	1318	59,1	22	0

Ausschnitt aus der CSV-Datei mit 10.000 Datensätzen

a) Sie sollen sich zwischen den Algorithmen *k-nearest neighbors* (kNN) und einem Entscheidungsbaum wie dem id3-Algorithmus entscheiden.

aa) Erklären Sie kurz beide Algorithmen.

6 Punkte

k-nearest neighbour:

---

---

---

---

---

---

---

---

id3:

---

---

---

---

---

---

---

---

Fortsetzung 2. Aufgabe →

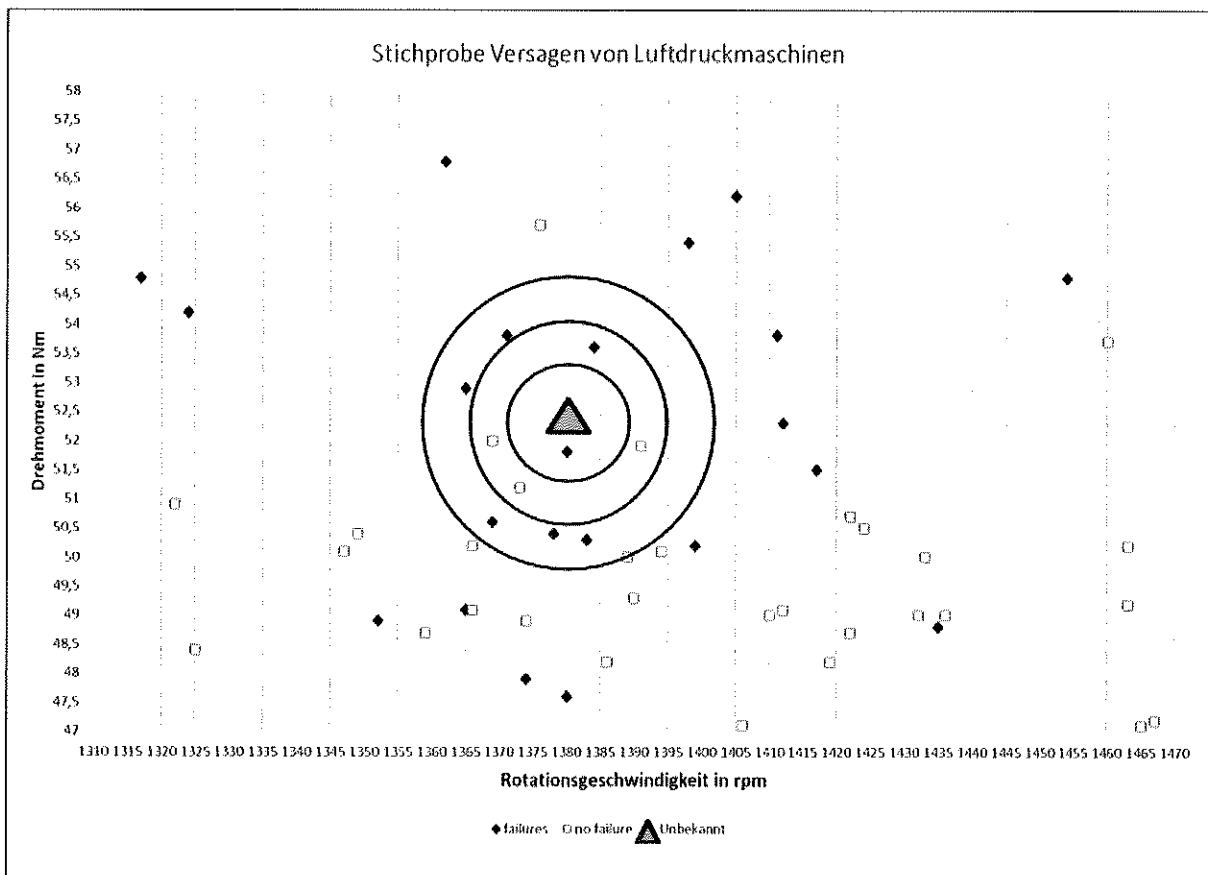
## Fortsetzung 2. Aufgabe

Korrekturrand

ab) Entscheiden Sie sich für einen der beiden Algorithmen und begründen Sie Ihre Entscheidung mit zwei Argumenten.

5 Punkte

b) Gegeben ist die folgende Grafik, welche zwei Features, Drehmoment und Drehgeschwindigkeit, für Luftdruckmaschinen visualisiert.



ba) Geben Sie eine Einschätzung darüber ab, ob die zu klassifizierende Maschine in nächster Zeit gewartet werden sollte.

1 Punkt

bb) Die konzentrischen Kreise um den zu klassifizierenden Datenpunkt sollen Sie bei der Entscheidung unterstützen.

Korrekturrand

Erläutern Sie, inwiefern die Größe des Parameters „k“ ausschlaggebend für die Entscheidung ist.

4 Punkte

bc) Nennen und begründen Sie zwei allgemein sinnvolle Kriterien für „k“.

4 Punkte

c) Die Daten für die Qualität sind bereits in low (L), medium (M) und high (H) quality klassifiziert.

Erklären Sie anhand dieses Beispiels, wie diese Daten aufbereitet werden müssen, damit sie für den kNN-Algorithmus verwendet werden können.

4 Punkte

## Fortsetzung 2. Aufgabe

Korrekturrand

- d) Die oberen Graphen beziehen sich nur auf zwei Features. Gegeben sind Daten für eine bereits klassifizierte Maschine und eine zu klassifizierende Maschine.

Rechnen Sie exemplarisch die euklidische Distanz für alle Features aus. Geben Sie den Rechenweg an.

4 Punkte

Koordinaten der bekannten Maschine:

Qualität	Prozesstemperatur [K]	Drehgeschwindigkeit [rpm]	Drehmoment [Nm]	Einsatzzeit [min]	Ergebnis
M	312,3	1.454	54,8	253	0

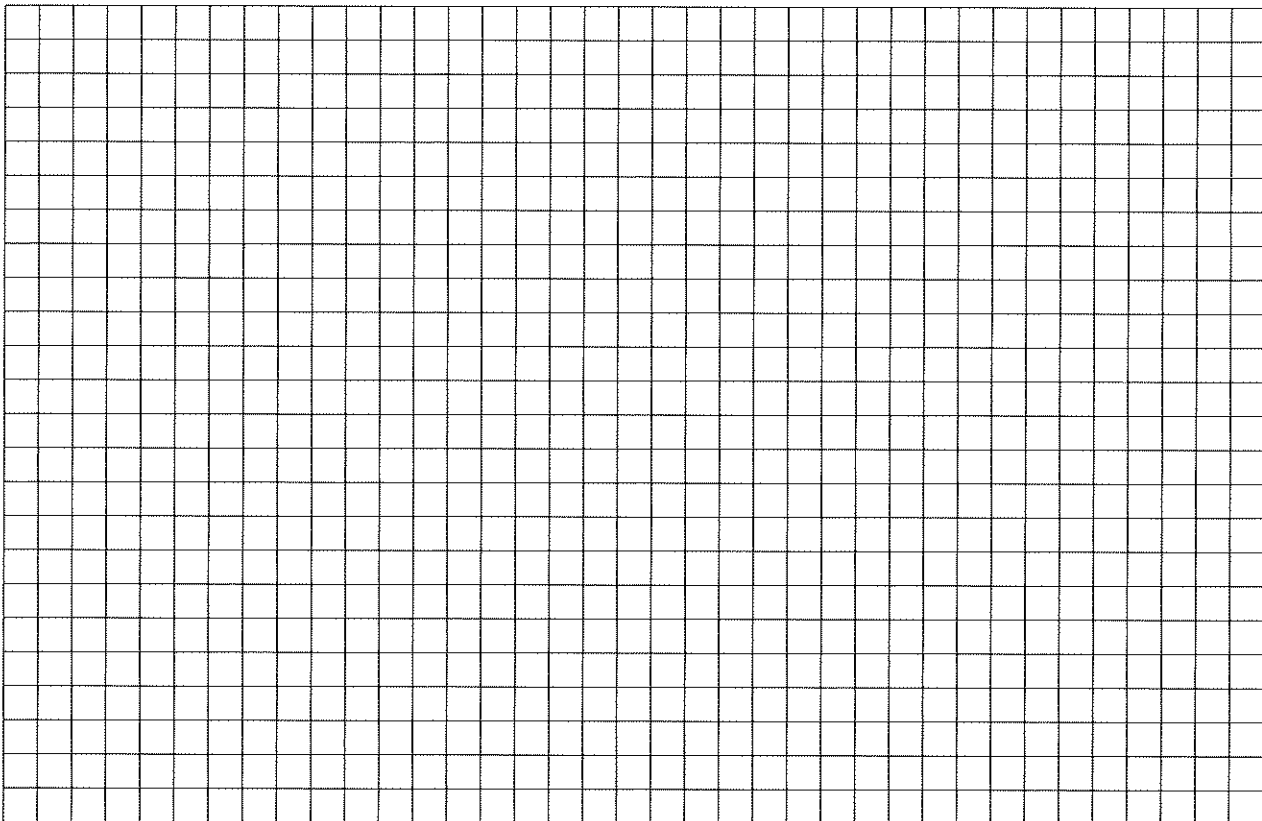
Koordinaten der zu klassifizierenden Maschine:

Qualität	Prozesstemperatur [K]	Drehgeschwindigkeit [rpm]	Drehmoment [Nm]	Einsatzzeit [min]
M	305	1.380	52,5	242

Euklidische Distanz:

Sind die Punkte  $p$  und  $q$  durch die Koordinaten  $p = (p_1, \dots, p_n)$  und  $q = (q_1, \dots, q_n)$  gegeben, so gilt:

$$\sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$





### 3. Aufgabe (25 Punkte)

Korrekturrand

Die Informationen über die Druckluftanlagen und deren Sensoren in den einzelnen Standorten sind zurzeit zentral in einer relationalen Datenbank gespeichert.

Zur Steigerung der Ausfallsicherheit sollen die Informationen dezentral gespeichert werden.

Für Datenbanktransaktionen in relationalen Datenbankmanagementsystemen sind folgende vier Eigenschaften (Akronym ACID) von besonderer Wichtigkeit:

a) Beschreiben Sie die jeweilige Bedeutung der vier Transaktionseigenschaften.

8 Punkte

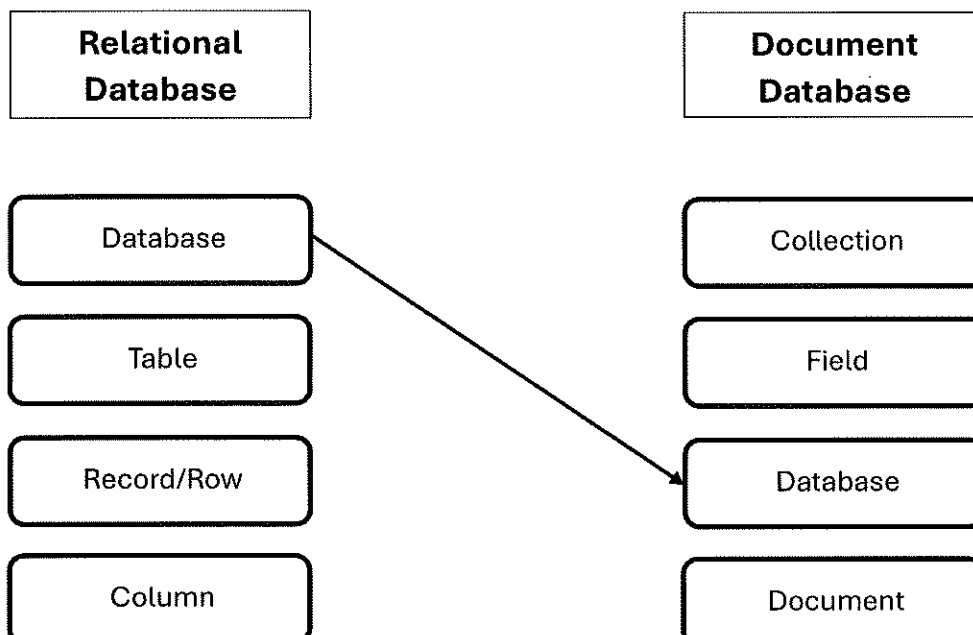
Atomicity (Abgeschlossenheit)	
Consistency (Konsistenz)	
Isolation (Abgrenzung)	
Durability (Dauerhaftigkeit)	

b) Um eine hohe Verfügbarkeit der Daten zu erreichen, bekommen alle Standorte mit Druckluftanlagen eine Teildatenbank in einem verteilten Datenbanksystem.

Dazu muss vom relationalen auf ein dokumentenorientiertes Datenbanksystem umgestellt werden.

Ordnen Sie die jeweiligen relationalen und dokumentenorientierten Datenbankfachbegriffe richtig zu.

3 Punkte



Fortsetzung 3. Aufgabe →

ca) Das neue Datenbanksystem verwendet JSON-Dokumente.

Wandeln Sie auf der Folgeseite die gegebenen Entitäten der Tabelle Sensoren in die JavaScript Objekt Notation (JSON) um.

Nutzen Sie dafür die gegebene JSON-Syntax.

7 Punkte

```
// JSON-Syntaxbeschreibung
// Wert kann folgende Typen haben: Text, Zahl, Boolean, null, Objekt, Array
// Texte in Anführungszeichen
{
    "attributname": wert,
    ...,
    "arrayname":
    [
        {
            "attributname": wert,
            "objektname":
            {
                "attributname": wert,
                ...
            },
            ...
        },
        ...
    ],
    ...
}
```

Sensoren

id	anlage_id	art	einheit
1	1	pressure	bar
2	1	temperature	C
...	...	...	...

cb) Auf die Anlagen- und Sensordaten wird meistens gemeinsam zugegriffen. Außerdem ist die Anzahl der Sensoren begrenzt. In solchen Fällen werden die Daten in einer Dokumenten-Datenbank auch gemeinsam gespeichert.

Fügen Sie die beiden Sensordokumente in ein Array ein, das Teil des Anlagendokumentes ist.

7 Punkte

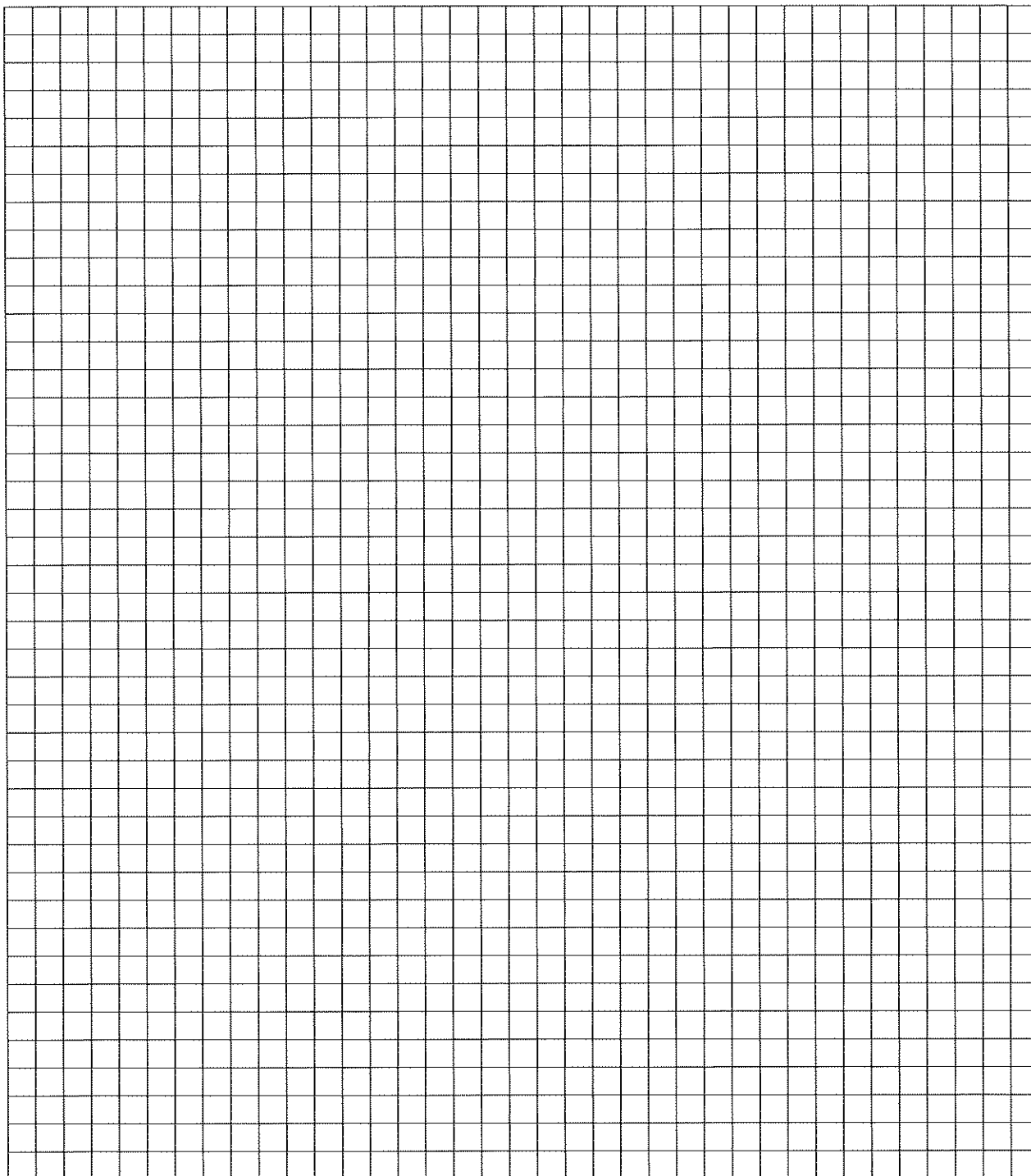
Anlagen

id	typ	baujahr
1	BASIC-G2	2023
...	...	...

Sensoren

id	anlage_id	art	einheit
1	1	pressure	bar
2	1	temperature	C

```
// anlagen
[
  {
    „id“: 1,
    „typ“: „BASIC-G2“,
    „baujahr“: 2023,
    „sensoren“: [
      {
        „id“: 1,
        „anlage_id“: 1,
        „art“: „pressure“,
        „einheit“: „bar“
      },
      {
        „id“: 2,
        „anlage_id“: 1,
        „art“: „temperature“,
        „einheit“: „C“
      }
    ]
  },
  ...
]
```



#### 4. Aufgabe (22 Punkte)

Korrekturrand

Sie werden gebeten, die Datenbankinfrastruktur zu optimieren und sicherzustellen, dass die Datenverfügbarkeit und -integrität gewährleistet sind. Sie sollen außerdem die Datenspeicherung verbessern und sicherstellen, dass bei Festplattenausfällen keine Daten verloren gehen.

a) Nennen und beschreiben Sie zwei allgemein gängige RAID-Level.

Nennen Sie zu den zwei RAID-Leveln je ein Vor- und Nachteil.

10 Punkte

RAID-Level:

Beschreibung:

Vorteil:

Nachteil:

RAID-Level:

Beschreibung:

Vorteil:

Nachteil:

b) Entscheiden Sie sich, welches RAID-Level Sie unter Berücksichtigung der Datenintegrität und Performance wählen, und begründen Sie Ihre Entscheidung.

4 Punkte

- c) Nennen Sie je eine geeignete Maßnahme, um die Informationen der Datenbank vor fremdem Zugriff zu schützen. 5 Punkte

Korrekturrand

Bereich	Maßnahme
Anwendung	
Transport zwischen Anwendung und Datenbank	
Datenbankservice	
Server der Datenbank	
Festplatte im Server	

- d) Um Änderungen rückgängig machen zu können, empfiehlt sich eine Backupstrategie für den zentralen Datenbankserver. Sensordaten werden rund um die Uhr erhoben und in der Datenbank gespeichert. Die Backups werden auf einem eigenen Server gespeichert, auf dem ausreichend Speicher ist.

Nennen Sie eine geeignete Backupstrategie und begründen Sie, warum Sie diese gewählt haben.

3 Punkte

---

---

---

---

---

---

### PRÜFUNGSZEIT – NICHT BESTANDTEIL DER PRÜFUNG!

Wie beurteilen Sie nach der Bearbeitung der Aufgaben die zur Verfügung stehende Prüfungszeit?

- ☐ 1 Sie hätte kürzer sein können.  
☐ 2 Sie war angemessen.  
☐ 3 Sie hätte länger sein müssen.

☐





