Abschlussprüfung Winter 2024/25
Lösungshinweise

Fachinformatiker/Fachinformatikerin
Daten- und Prozessanalyse
1203

2 Sicherstellen der Datenqualität

Teil 2 der Abschlussprüfung

Allgemeine Korrekturhinweise

Die Lösungs- und Bewertungshinweise zu den einzelnen Handlungsschritten sind als Korrekturhilfen zu verstehen und erheben nicht in jedem Fall Anspruch auf Vollständigkeit und Ausschließlichkeit. Neben hier beispielhaft angeführten Lösungsmöglichkeiten sind auch andere sach- und fachgerechte Lösungsalternativen bzw. Darstellungsformen mit der vorgesehenen Punktzahl zu bewerten. Der Bewertungsspielraum des Korrektors (z. B. hinsichtlich der Berücksichtigung regionaler oder branchenspezifischer Gegebenheiten) bleibt unberührt.

Zu beachten ist die unterschiedliche Dimension der Aufgabenstellung (nennen – erklären – beschreiben – erläutern usw.).

Für die Bewertung gilt folgender Punkte-Noten-Schlüssel:

Note 1	=		100 – 92 Punkte	Note 2	=	unter	92 – 81 Punkte
Note 3	=	unter	81 – 67 Punkte	Note 4	=	unter	67 – 50 Punkte
Note 5	=	unter	50 – 30 Punkte	Note 6	=	unter	30 – 0 Punkte

1. Aufgabe (25 Punkte)

a) 6 Punkte

Extraktion
der relevanten Daten aus (verschiedenen) Quellen
Transformation
Umwandlung/Verarbeitung der Daten in das Zielformat für die weitere Verwendung
Laden
der Daten in die Zieldatenbank/Zielsystem/Zieldarstellung

b) 3 Punkte

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	**97**

- Ausreißerwerte, z. B. extrem hohe oder niedrige Werte
- Nicht plausible Werte
- Datenlücken in den Aufzeichnungen
- Duplikate von Sensordaten

c) 6 Punkte

Datenvalidierung:
Dies bezieht sich auf die Überprüfung, ob Daten den erwarteten Regeln und Kriterien entsprechen.
Beispiel: Überprüfen, ob die Temperaturen nur Zahlen sind und in festgelegten Bereichen liegen.

Datenverifizierung:
Dies bezieht sich auf die Überprüfung, ob die Daten tatsächlich existieren und korrekt sind.
Beispiel: Überprüfen, ob die Maschine existiert, indem die ID in der anderen Datenbank geprüft wird.

d) 10 Punkte

In dieser Aufgabe soll der Azubi seine Erkenntnisse zusammenführen. Die Aufgabe kann auf verschiedene Arten gelöst werden. Der Text hier stellt eine Beispielbeantwortung dar. Für die Bewertung ist nur relevant, dass es sinnvoll begründet wurde.
Betreff: Einschätzung zur Datenqualität
Text:
Sehr geehrte Frau Bellenbaum,
Hier ist meine Einschätzung zur Datenqualität: (zwei reichen)

- Die Genauigkeit der erfassten Daten ist in einigen Fällen unzureichend, da Inkonsistenzen zwischen den Datenquellen und den erfassten Werten festgestellt wurden.
- Die Vollständigkeit der Datensätze variiert, wobei einige Felder fehlen oder leere Werte enthalten.
- Die Konsistenz der Daten, insbesondere in Bezug auf Format und Einheiten, weist Unstimmigkeiten auf.

Maßnahmen zur Bereinigung der erkannten Datenprobleme: (zwei reichen)

- Validierung der Datenquellen: Es ist entscheidend, die Datenquellen zu validieren und sicherzustellen, dass die erfassten Informationen korrekt und aktuell sind. Dies kann durch regelmäßige Überprüfungen und Aktualisierungen erfolgen, um Inkonsistenzen zu vermeiden.
- Datenbereinigung: Wir sollten einen Prozess zur Bereinigung der erfassten Daten implementieren. Das beinhaltet das Auffinden und Korrigieren von fehlerhaften oder inkonsistenten Werten sowie das Hinzufügen von fehlenden Informationen. Dies kann manuell oder automatisiert durchgeführt werden, abhängig von der Art der Daten.
- Standardisierung und Dokumentation: Um die Konsistenz zu verbessern, sollten klare Standards und Richtlinien für die Datenerfassung und -speicherung entwickelt und dokumentiert werden. Dies umfasst einheitliche Formate, Einheiten und Feldbezeichnungen.

Mit freundlichen Grüßen,
[Ihr Name]

Bewertungsvorschlag:	1 Punkt	Sinnvoller Betreff
	1 Punkt	E-Mail-Format, d. h. Anrede und Grußformel.
	4 Punkte	Einschätzung der Datenqualität
	4 Punkte	Maßnahmen zur Verbesserung

2. Aufgabe (28 Punkte)

aa) 6 Punkte

k-nearest neighbors:

Ein neuer Datenpunkt wird anhand einer bestimmten Anzahl der benachbarten Datenpunkte klassifiziert. Dabei werden die nächsten Nachbarn befragt, was sie sind, und die Mehrheit der Antworten entscheidet über die Klassifizierung des neuen Datensatzes.

id3:

Der id3-Algorithmus baut anhand bekannter Datenpunkte einen Entscheidungsbaum auf, wobei der Informationsgewinn der Kriterien über die nächste Ebene des Baumes entscheidet. Auf unterster Ebene entscheidet die Mehrheit der Antworten.

ab) 5 Punkte, 1 Punkt für die Entscheidung und je 2 Punkte für die Argumente

- Die Wahl muss auf k-nearest neighbors fallen.
- Ein Entscheidungsbaum ist unflexibel in Bezug auf neue Daten.
- Des Weiteren sind die Daten numerisch und schwer in klassifizierte Daten umzuwandeln.
- Entscheidungsbäume neigen zu Overfitting
- u. a.

ba) 1 Punkt

Ja, die meisten Nachbarn sind failures.

bb) 4 Punkte

k = 1 sagt failure
k = 5 sagt no failure
k = 10 sagt failure
Die Größe von k kann also die Klassifizierung eines neuen Datenpunktes beeinflussen.

bc) 4 Punkt

Zwei der folgenden:

- K sollte **ungerade** sein, damit es immer eine Entscheidung gibt.
- K darf **nicht zu klein** sein, da sonst die Einstufung eventuell von einem Ausreißer zu stark beeinflusst wird.
- K darf **nicht zu groß** sein, damit die Entscheidung nicht zu unspezifisch wird. Wie man an dem Kreis in der oberen Grafik sehen kann, muss man das k nur groß genug wählen, damit das Ergebnis immer no failure ist.

c) 4 Punkte

1. Sie müssen numerische Werte erhalten, damit mit ihnen gerechnet werden kann. z. B. L = 1, M = 2, H = 3
2. Diese Wert müssten skaliert werden, damit sie die Größenordnung der anderen Daten widerspiegeln. In diesem Fall z. B. mal 10 → L = 10, M = 20, H = 30

d) 4 Punkte

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

$$= \sqrt{(20-20)^2 + (312{,}3-305)^2 + (1454-1380)^2 + (54{,}8-52{,}5)^2 + (253-242)^2}$$

$$= \sqrt{0 + 7{,}3^2 + 74^2 + 2{,}3^2 + 11^2}$$

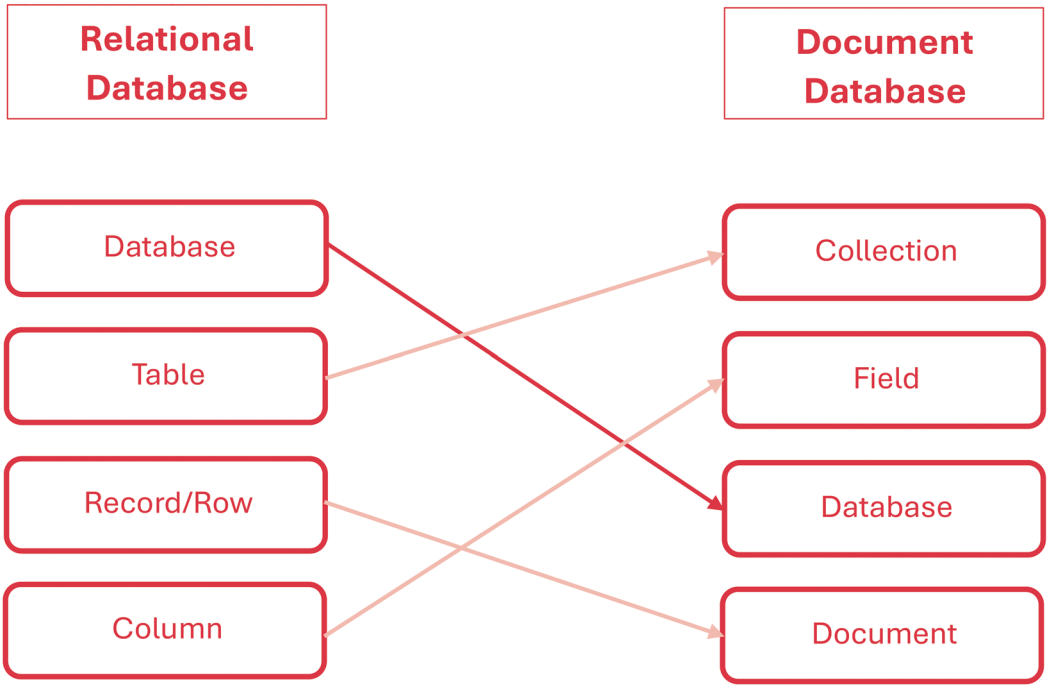
$$= \sqrt{53{,}29 + 5476 + 5{,}29 + 121} \approx 75{,}20$$

3. Aufgabe (25 Punkte)

a) 8 Punkte, je Erläuterung 2 Punkte

Atomicity (Abgeschlossenheit)	Eine Transaktion ist eine geschlossene unteilbare Einheit. Sie wird entweder ganz oder gar nicht ausgeführt. Tritt irgendwo in der Transaktion ein Fehler auf, so müssen bereits durchgeführte Änderungen wieder rückgängig gemacht werden.
Consistency (Konsistenz)	Die für das jeweilige Datenbanksystem festgelegten Konsistenzregeln müssen sowohl vor als auch nach der Transaktion erfüllt sein. Kann die Einhaltung der Regeln nicht überprüft werden oder tritt ein Fehler auf muss die gesamte Transaktion rückgängig gemacht werden.
Isolation (Abgrenzung)	Es muss gewährleistet sein, dass gleichzeitig durchgeführte Transaktionen voneinander abgegrenzt sind. D.h. das Ergebnis muss das gleiche sein, als ob die Transaktionen hintereinander durchgeführt würden. Realisiert wird das z. B. durch Sperrverfahren.
Durability (Dauerhaftigkeit)	Die dauerhafte Speicherung der Daten muss auch nach einem Systemfehler (Software-Fehler oder Hardware-Ausfall) garantiert sein. Dauerhaftigkeit kann durch das Schreiben von Transaktionslogs sichergestellt werden.

b) 3 Punkte, je Pfeil 1 Punkt



ca) 7 Punkte

```
// sensoren
[
  {
     "id": 1,
     "anlage_id": 1,
     "art": "pressure",
     "einheit": "bar"
  },
  {
     "id": 2,
     "anlage_id": 1,
     "art": "temperature",
     "einheit": "C"
  },
      …
```

cb) 7 Punkte

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

// auch richtig mit Sensor-IDs
```

4. Aufgabe (22 Punkte)

a) 10 Punkte
Es müssen nur 2 ausgewählt werden. Je 1 Punkt für die Nennung, 2 Punkte für die Beschreibung und je 1 Punkt je Vor-/Nachteil

RAID 0 (Striping):
- Erläuterung: RAID 0 verwendet Striping, bei dem Daten auf zwei oder mehr Festplatten aufgeteilt werden. Dies erhöht die Schreib- und Lesegeschwindigkeit, da Daten gleichzeitig auf mehreren Laufwerken verarbeitet werden.
- Vorteile: verbesserte Leistung durch Striping
- Nachteile: Keine Datenredundanz, ein Festplattenausfall führt zum Datenverlust.

RAID 1 (Mirroring):
- Erläuterung: RAID 1 verwendet Mirroring, wobei Daten gleichzeitig auf zwei Festplatten geschrieben werden. Dies erhöht die Datenintegrität, da eine Festplatte als Spiegel der anderen fungiert.
- Vorteile: hohe Datenintegrität, schnelles Lesen
- Nachteile: höhere Kosten aufgrund der doppelten Speicherkapazität

RAID 5 (Block-Level Striping mit Parität):
- Erläuterung: RAID 5 verwendet Striping und Parität. Daten werden in Blöcken auf Festplatten verteilt, und Paritätsinformationen ermöglichen die Wiederherstellung von Daten im Falle eines Festplattenausfalls.
- Vorteile: gute Kombination aus Leistung und Redundanz, effiziente Speichernutzung
- Nachteile: komplexe Wiederherstellung bei Festplattenausfall

RAID 10 (Kombination von RAID 1 und RAID 0):
- Erläuterung: RAID 10 kombiniert Spiegelung (RAID 1) und Striping (RAID 0). Daten werden auf zwei oder mehr Festplatten gespiegelt und dann gestreift.
- Vorteile: hohe Leistung und Datenintegrität, schnelle Wiederherstellung nach einem Festplattenausfall
- Nachteile: höhere Kosten aufgrund der doppelten Speicherkapazität

RAID 6 (Block-Level Striping mit doppelter Parität):
- Erläuterung: RAID 6 ist ähnlich wie RAID 5, verwendet aber doppelte Parität. Dies bietet Schutz vor Datenverlust bei Ausfall von zwei Festplatten.
- Vorteile: hohe Datenredundanz und Schutz vor Datenverlust bei Ausfall von zwei Festplatten
- Nachteile: Schreibgeschwindigkeit ist langsamer als RAID 5

b) 4 Punkte
Es empfiehlt sich für Datenintegrität und Leistung RAID 10.
RAID 10 kombiniert RAID 1 (Spiegelung) und RAID 0 (Streifen), was hohe Leistung und hohe Redundanz bietet. Es ermöglicht schnelles Lesen und Schreiben von Daten sowie den Schutz vor Datenverlust.
Andere RAID-Level mit sinnvoller Begründung sind auch möglich, wie z. B. RAID 6.

c) 5 Punkte
Pro Bereich reicht eine Maßnahme (je 1 Punkt). Weitere sinnvolle Maßnahmen sind möglich.

Bereich	**Maßnahme**
Anwendung	Authentifizierung Ausführbar nur in geschützten Umgebungen
Transport zwischen Anwendung und Datenbank	Transportverschlüsselung wie SSL/TLS
Datenbankservice	Whitelist für Zugriffe DB-User haben nur notwendige Rechte DB-Porteinschränkung
Server der Datenbank	Firewall geschützt Whitelist zum Einloggen User haben nur notwendige Rechte Updates
Festplatte im Server	Verschlüsselung/Bitlocker/AES u. a.

d) 3 Punkte, 1 Punkt Nennen der Strategie, 2 Punkte warum gewählt

Es gibt mehrere sinnvolle Strategien. Wichtig ist eine sinnvolle Begründung.

Vollständige Backups
Begründung: Diese Backups erfassen alle Daten und Objekte in der Datenbank und können direkt wiederhergestellt werden. Der Backup-Server hat genug Speicher dafür.

Inkrementelle Backups
Begründung: Diese Backups sind sehr klein und können sehr schnell erstellt werden.

Generationenprinzip als Kombination aus differenziellen, inkrementellen und vollständigen Backups

Transaktionsprotokoll-Backups
Begründung: Sind sofort erstellbar, da die Transaktionen mitgeführt werden. Jede Änderung ist nachvollziehbar. Datenbank kann weiterhin alle Daten schreiben.