Subject number : XXXXXXXXX

Sujet corrigé - 6 - Data Normalization, Denormalization, Analysis and Communication

This is the **question paper**. It is **NOT** the answer sheet.
Please check that the number on your question paper matches the number on your answer sheet. To complete the answer sheet correctly, you must:

- use a **black** ink pen
- shade in the boxes **completely without going over the edges**
- if you make a mistake, erase the box with a whiteout ("Tipp-Ex"), **but do not redraw it**
- every question has a **unique correct answer**.

Box correctly ticked

1 A **B** C D E

Box incorrectly ticked

1 A ~~B~~ C D E

1 What is the main goal of 2NF?
1 Point - Only one correct choice

- ☐ **A.** To ensure all attributes are mutually independent
- ☒ **B.** To eliminate partial functional dependencies on a primary key
- ☐ **C.** To ensure each non-key attribute is directly dependent on some candidate key
- ☐ **D.** To ensure that every table has a single-column primary key

> 2NF addresses partial dependencies, ensuring non-key attributes depend on the entire primary key, not just a part of it.

2 What is a transitive dependency?
1 Point - Only one correct choice

- ☐ **A.** An attribute that depends on the part of a primary key
- ☒ **B.** A non-key attribute depends on another non-key attribute
- ☐ **C.** A primary key that depends on a foreign key
- ☐ **D.** A candidate key depends on another candidate key

> A transitive dependency occurs when a non-key attribute depends on another non-key attribute, which itself depends on the primary key.

3 Which normal form specifically addresses multi-valued dependencies?
1 Point - Only one correct choice

- ☐ **A.** 2NF
- ☐ **B.** 3NF
- ☒ **C.** 4NF
- ☐ **D.** BCNF

> 4NF deals with multi-valued dependencies, where a table may have independent multi-valued facts about an entity.

4 A university database stores student clubs as follows:

StudentID	Clubs
101	Chess, Debate, Photography
102	Debate, Robotics

Which normal form does this violate, and why?
1 Point - Only one correct choice

- ☐ **A.** 2NF, because Clubs is not atomic
- ☒ **B.** 1NF, because Clubs contains repeating groups
- ☐ **C.** 3NF, because StudentID is not a candidate key
- ☐ **D.** BCNF, because Clubs depends on StudentID

> 1NF requires atomic values; "Chess, Debate, Photography" is a repeating group.

5 **What is a common trade-off when denormalizing a database?**
1 Point - Only one correct choice

- [ ] **A.** Improved performance when executing complex joins
- [x] **B.** Increased risk of data anomalies
- [ ] **C.** Reduced need for indexing
- [ ] **D.** Better alignment between the logical and physical schema

> Denormalization can lead to data anomalies (like update anomalies) because redundant data is introduced, which can become inconsistent.

6 **Why might a database designer choose to denormalize data from an external source?**
1 Point - Only one correct choice

- [ ] **A.** To reduce coupling between the external source and the internal schema
- [x] **B.** To integrate data that does not fit the existing schema
- [ ] **C.** To guarantee that future schema evolution will be easier
- [ ] **D.** To enforce integrity constraints across independently managed datasets

> Denormalization is often used for external data that is not easily mapped to the existing schema or changes frequently, making strict normalization impractical.

7 **When is denormalization NOT the first solution for performance issues?**
1 Point - Only one correct choice

- [ ] **A.** When the workload is dominated by expensive join operations
- [x] **B.** When performance issues can be addressed in the processing layer
- [ ] **C.** When integrating data from external, non-normalized sources
- [ ] **D.** When business rules evolve and invalidate a normalized design

> Performance issues should first be addressed in the processing layer, using tools like indexes or caching, before considering denormalization.

8 **What is the primary goal of Exploratory Data Analysis (EDA)?**
1 Point - Only one correct choice

- [x] **A.** To understand data patterns, spot anomalies, and generate hypotheses
- [ ] **B.** To identify causal relationships in the data
- [ ] **C.** To verify assumptions required for advanced statistical models
- [ ] **D.** To quantify how far the data is from the expected distributions

> EDA is used to explore and summarize data, identify patterns, detect anomalies, and form hypotheses for further analysis.

9 **Which of the following is a key principle of effective data storytelling?**
1 Point - Only one correct choice

- [ ] **A.** Presenting the data in the same format regardless of audience expertise
- [ ] **B.** Maximizing the number of visualizations to capture all nuances
- [x] **C.** Tailoring the narrative to the audience's needs and knowledge level
- [ ] **D.** Avoiding interpretation so that the audience draws its own conclusions

> Effective data storytelling focuses on clarity, relevance, and audience engagement.

10 **What does Simpson's Paradox illustrate about data analysis?**
1 Point - Only one correct choice

- [ ] **A.** Aggregating data typically reduces the influence of confounding variables
- [x] **B.** A trend observed in subgroups can reverse when the groups are combined.
- [ ] **C.** Data analysis should always focus on overall trends, not subgroups.
- [ ] **D.** Data analysis should always focus on subgroups, not overall trends.

> Simpson's Paradox demonstrates that trends in subgroups can reverse when the data is aggregated, highlighting the importance of stratification in analysis.