

School of Informatics (IIT)

## WRITTEN EXAMINATION

Course **Data Warehousing - teknologier och metoder**

Sub-course

Course code **IT382G**

Credits for written examination **6**

Date **2025-06-09**

Examination time **8:15 – 12:30**

Examination responsible **Manfred Jeusfeld**

Teachers concerned **Manfred Jeusfeld**

Aid at the exam/appendices

Students are allowed to bring a Swedish-English dictionary to the exam

- Instructions
- ☒ Take a new sheet of paper for each exam part.
  - ☐ Take a new sheet of paper when starting a new question.
  - ☒ Write only on one side of the paper.
  - ☒ Write your name and personal ID No. on all pages you hand in.
  - ☒ Use page numbering.
  - ☒ Don't use a red pen.
  - ☐ Mark answered questions with a cross on the cover sheet.

Grade points: 90 (30 for each of the three parts)

Answer in Swedish or English. Answer all the sub-questions of a question. To pass the exam you must reach at least 50% of the possible points. The ECTS grades are calculated as follows from the percentage of points you achieved (90 points=100%):

<50%: F; 50-59%: E; 60-69%: D; 70-79%: C; 80-89%: B; 90-100%: A

**Examination results shall be made available according to university rules via LADOK**

*Good luck!*

Total number of pages excluding this page: 3

### Part 1: Central concepts (30 points)

- a) (10) The relational schema of a data warehouse can follow the so-called “star schema” or the “snowflake schema” approach. Create a **snowflake schema** for a company that sells products. Include at least three dimensions (time, customer, product, shop,...) , where at least one dimension is extended according to the snowflake approach. How can a snowflake schema be converted to a star schema?
- b) (10) Dimension tables sometimes need to be updated, e.g. when a new customer is added. They may also need to be updated, e.g. when the address of a customer changes. The course discussed **three strategies to make such updates**. Describe these strategies and give an example on how they work in the case of updating the address of a customer. What are the advantages and disadvantages of the strategies!
- c) (10) So-called **information package diagrams** are used to collect the requirements for a data warehouse. Explain what is represented in such diagrams. Give an example that has at least three dimensions and two metrics. Consider a supermarket chain that sells products of certain product type and from different brands. They have regular supermarket branches in cities and so-called mini-markets, Create an information package diagram (table) that has at least three dimensions and three attributes per dimension.

## Part 2: Data Warehousing and OLAP (30 points)

- a) (10) Explain what we understand by a “**data cube**”. Explain the data cube operations “**selection**” and “**drill-down**” (ca. 5-10 lines of text per operation). Use diagrams to visualize the effect of these operations.
- b) (10) Dimension tables need to be constructed in a way that the aggregation (roll-up) is done correctly, in particular for avoiding **double-counting** or **missing values** in the aggregation (SUM,AVG,...). Explain what has to be checked for dimension tables in order to avoid the two problems. Give an example of a malformed dimension tables where double-counting can occur.
- c) (10) Views in a data warehouse (“**aggregate fact table**”) can be materialized (i.e. the content of the view is precomputed). Give an example star schema that shows such an aggregate fact table and explain it. Why does it make sense to precompute such fact tables?

### Part 3: Data Mining (30 points)

- a) (10) Discuss the **main differences of OLAP versus “data mining”**. Which technique is more automated? What types of knowledge can be found in data mining that cannot be easily found with OLAP (or multi-dimensional queries)? Which are more suitable for predicting what happens in the future? List at least three main differences!
- b) (10) What are **decision trees**? Give an example of a small decision tree. What is the role of the “test nodes”? How are they related to attributes (of subjects). Give small example with 3-5 nodes.
- c) (10) Explain the idea and principal procedure of the **“memory-based reasoning”** method? What is the goal of the method? Give an example to illustrate the method.