UNIVERSITY
OF SKÖVDE

School of Informatics (IIT)

WRITTEN EXAMINATION

Course **Data Warehousing - teknologier och metoder**

Sub-course

Course code **IT382G** Credits for written examination **6**

Date **2025-03-28** 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 questions. To pass the exam, all three parts must be passed. Each question has an equal weight of 10 points. The final grade is calculated from these points. All three part solutions must have at least 50% of the possible points.

Examination results should be made public within 18 working days

Good luck!

Total number of pages excluding this page: 3

UNIVERSITY
OF SKÖVDE

Part 1: Central concepts (30 points)

a) (10 points) What are the functions of the ETL process? Provide a diagram that shows the role of the ETL process. Your description of the functions should be comprehensive, i.e. at least ½ of a page. What are possible implementation strategies for ETL? For example, how to handle data that is in spreadsheets? How does the ETL process make sure that it does not interfere with operational systems?

b) (10 points) Consider a data warehouse on the sales of train tickets. The ticket is sold at a certain **time**, to a certain **customer**, for a certain **train**, and for a **connection** between a **start station** and an **end station**. Customers have attributes like address, gender, etc. Trains have a train number, belong to a certain train category (regional, high speed, ...). Stations belong to cities, which are part of regions, which are part of countries. Create a star schema for the data warehouse using UML class diagrams. The measurement attribute shall be the **ticket price** and the **quantity** sold to the customer.

c) (10 points) **Information packages** are a tool to capture requirements to a data warehouse. Consider a logistics company that transports containers on trucks between harbors, warehouses and customers. Among others, the bus company is interested in the time it takes to transport certain container types from/to certain location types. Create an information package diagram (table) that has at least three dimensions and three attributes per dimension.

UNIVERSITY
OF SKÖVDE

Part 2: Data Warehousing and OLAP (30 points)

a) (10 points) Create a two-dimensional PIVOT table for the dimensions **time** and **product** out of the following fact table. Note that the price attribute is the measurement attribute here.

Time	Location	Product	Customer	Price
9:12	Skara	Cykel-B	Fred	5050
10:31	Skövde	Cykel-A	Fred	10 500
12:30	Malmö	Cykel-A	Mary	15 800
12:47	Malmö	Bil-B	Paul	118000

b) (10 points) Certain measurement attributes are not **summarizable**, i.e. the application of the SUM operation makes no sense. Explain what we understand by "**flow observations**", "**stock observations**" and "**value-per-unit observations**". Provide **two** examples for **each** type of observation and specify whether the values can be summed up or not.

c) (10 points) **MOLAP** and **ROLAP** are two strategies to provide data cubes to data warehouse clients (such as OLAP tools). Explain what we understand by MOLAP and ROLAP! When should MOLAP be used and when is ROLAP a better choice?

UNIVERSITY
OF SKÖVDE

Part 3: Data Mining (30 points)

a) (10 point) 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 is more suitable for predicting what happens in the future? List at least three main differences!

b) (10 points) What are **neural networks**? Give an example diagram of a neural network. What is the role of the weights? What is the purpose of "training" the neural network? Answer all parts! What is the meaning of hidden layer?

c) (10 points) What are **association rules**? Explain the notion of "support factor" and "confidence factor".