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 **2024-05-27** 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) Describe the general **architecture** of a data warehouse, i.e. the components that belong to a data warehouse system (data sources, data staging area, ...) and how they are connected to each other. Give the diagrams for at least **two architectures** ("hub and spoke" vs one other) and indicate the advantages and disadvantages of these two architectures. What are the implications to **efficiency** of the ETL process and to the efficiency of OLAP queries?

b) 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) Low **data quality** in data warehouses leads to wrong results. Which types of errors in data can occur? Explain at least **three types of "data corruption"** and give examples for these types. How can errors be corrected? Which component of the data warehouse is responsible for correcting errors? How can we be sure that the data in a data warehouse is complete?

UNIVERSITY
OF SKÖVDE

Part 2: Data Warehousing and OLAP (30 points)

a) Explain what we understand by a "**data cube**". Explain the data cube operations "**projection**", "**roll-up**", "**selection**" (ca. 5-10 lines of text per operation). Use diagrams to visualize the effect of these operations.

b) 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**". Give an example of a measurement attribute that cannot be summarized. Why is summarization over the **time dimension** a problem for stock observations?

c) A data warehouse can use **materialized views** to store aggregate facts such as sales summarized over days. Rather than computing the aggregate facts it can return the result from the materialized view. Describe **a star schema** that has two fact tables for product sales, one for the facts at the lowest level of granularity (each sale is a fact), and one for sales over a whole day. The dimensions are "product" (product name, type, category), "time" (second, day, month, year) and "location" (shop, city, region)

UNIVERSITY
OF SKÖVDE

Part 3: Data Mining (30 points)

a) 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) What are **association rules**? Explain the notion of "support factor" and "confidence factor".

c) Explain the idea and principal procedure of the "**k-Means**" method? For what does the "k" stand for? What is a centroid in the context of k-Means?