Dominik Kowald, Peter Muellner, Margarita Dubina
Graz University of Technology
Computer Science and Biomedical Engineering
Institute of Interactive Systems and Data Science

1 Databases WS24: Group Project

Published: Wed. 23.10.2024
Deadline Concept: Fri. 29.11.2024, 11.59pm
Deadline Full: Wed. 08.01.2025, 11.59pm

This group project aims to provide practical experience with databases in a freely chosen (data) domain. For this, build groups of 5 students in TeachCenter, and work on a data-driven problem of your choice. The group project consists of two parts (concept and full submission):

1.1 Concept (10/30 points)

Describe the concept of your project, which includes the research problem and the motivation:

- Choose a dataset, describe the research problem that you want to solve with it and why it is relevant (for you)
- A list of potential dataset repositories is provided in the course slides (lecture 2) and at the end of this document, but you are also free to use an own dataset from your field, or just a dataset that you like
- Important is the research problem or question you want to answer and why a relational database is the right way to store it (so it should have multiple tables that are meaningfully connected)
- You can also be creative - an example is given in the course slides (lecture 2)

Submit your concept in the form of a PowerPoint / LaTeX slide-set (any format is fine) with exactly 2 slides:

- Slide 1: title slide, which lists the group members along with their responsibilities in the group project
- Slide 2: your concept as described above with a link to the dataset used (= first objective of group project)

The expected result is a .pdf file named `DBConcept_<groupID>.pdf`, submitted in TeachCenter. You will receive feedback on this in case something needs to be changed.

1.2 Full Submission (20/30 points)

Besides the research problem and motivation, which was the first objective of this group project (see Section 1.1), 4 additional objectives should be addressed in the full submission. Each objective will be graded with additional 5 points maximum:

- Objective 2: Describe your dataset (e.g., Entity relationship diagram), how you have transferred the dataset to a database scheme and how you ingested the data (similar to Exercise 1 + first part of Exercise 2)
- Objective 3: Describe your database queries (should contribute to your research problem) and any data handling you performed (similar to second part of Exercise 2)
- Objective 4: For presenting/interpreting/discussing the results of your queries, you can use any tool or library that you want (Lecture 7 provides some examples, e.g., Pandas, Seaborn). You should interpret (process/analyze/visualize) the queried data.
- Objective 5: Describe reproducibility aspects related to your project, which should include a link to the dataset, and if additional appendices/code/schema files are submitted, how they need to be used to reproduce your results (a link to a GitHub repository is also fine)

Submit your full project in the form of a PowerPoint / LaTeX slide-set (any format is fine) with exactly 6 slides:

- Slide 1: title slide, which lists the group members along with their responsibilities in the group project (who has done what?)
- Slide 2: Objective 1 - research problem and motivation (as in the concept submission)
- Slide 3: Objective 2 - data modeling and data ingestion (see above)
- Slide 4: Objective 3 - database queries and data handling (see above)
- Slide 5: Objective 4 - presentation and discussion of results (see above)
- Slide 6: Objective 5 - reproducibility aspects (see above)

The expected result is a .pdf file named `DBProject_<groupID>.pdf`, submitted in TeachCenter. Additionally, you can submit any relevant appendices, code files, or data schema files (but no big datasets). The projects will be discussed in the first two lectures after the Christmas break (see course calender).

Some potential dataset sources are the following:

- Recommender Systems and Personalization Datasets (UC-San Diego): `https://cseweb.ucsd.edu/~jmcauley/datasets.html`
- Stanford Network Analysis Project (SNAP): `https://snap.stanford.edu/data/index.html`
- Data Science Challenges (Kaggle): `https://www.kaggle.com/datasets`
- Open Machine Learning Platform (OpenML): `https://www.openml.org/search?type=data&sort=runs&status=active`
- Open Data Platform Austria (data.gv.at): `https://www.data.gv.at/suche/?typeFilter%5B0%5D=dataset`
- Dataset Search Engine (Google): `https://datasetsearch.research.google.com/`
- GroupLens Research (MovieLens etc.): `https://grouplens.org/datasets/`