HSLU Lucerne University of Applied Sciences and Arts

Project assignment

Module Database Management for Data Scientists (DBM)

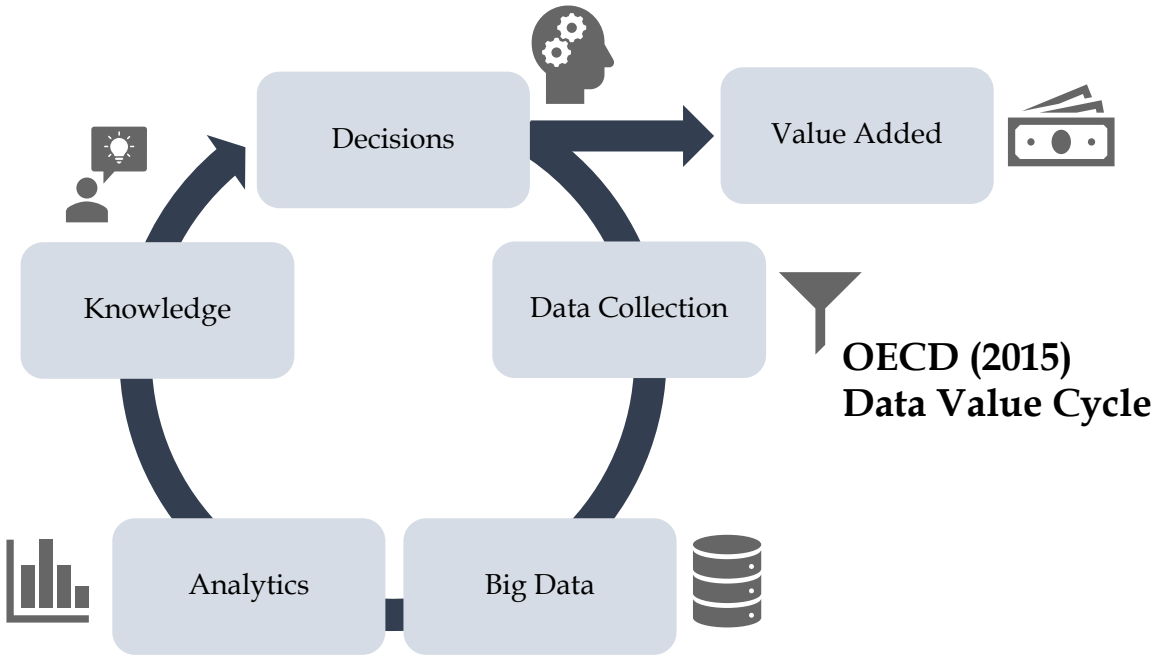
Objective: During the semester, you will work on a database project in teams of up to four. You will learn about SQL database technologies through their application in a use case. Thus, you can learn important skills as a data scientist. The project result will support decision-making with a visualization in a business intelligence (BI) tool by providing information obtained from a database query. We work with MySQL as an SQL database, and we use Metabase as a BI tool. You install these systems on the virtual machine (VM) server to simulate a realistic scenario.

Procedure:

1.) Plan the database application
2.) Define the database structure
3.) Load and transform data
4.) Analyze database
5.) Optimize database performance
6.) Visualize and apply results
7.) Write project report

1. Plan database application

- Develop ideas for your project to generate value from data.



- The data analysis in your project should support **decision-making** and thus create value with data, analogous to the Data Value Cycle (OECD, 2015).
- Specify which **key figures** can be determined based on which data.
- Specify the **decision rules** that show how data-based decision-making supports the use case.
- Select the data that can help decision-makers in the use case. Combine **at least two** different or formally independent data sources. For example, one Kaggle project counts as one data source, even if it contains different files. Two Kaggle projects count as two data sources.
- A selection of possible data sources can be found in the appendix.

2. Define the database structure

- Since you are working with existing data, you must reverse engineer the conceptual model based on the data already available.
- Analyze the structure and content of the source data using data examples
- Consider what the conceptual ER model might look like based on the source data and model it. In other words, you must extract the conceptual model of the existing business entities, attributes, and relationships from the source data and integrate them across the various sources.
- Create a conceptual, graphical model that maps the business entities, relationships, and attributes in the source data. We expect multiple entity types and relationships, each with qualitative attributes and detailed descriptions.
- The next step is to implement the data structure in the SQL database as a physical schema (SQL DDL). The effective schema in the DBMS may differ from the conceptual model.
- The database schema of the SQL DB must be in third normal form.
- Also show the relationship between the conceptual model and the database schema.
- In addition to the model and diagram, you should also describe the most important data that is essential for transformation, analysis, visualization, and decision-making in prose.

3. Load and transform data

- Install a MySQL database server with client-server architecture (server in the cloud or VM on Lab Services HSLU). It is important that your database system is implemented online so that the evaluators can log in to review your work. You must therefore disclose the access data for the evaluation on ILIAS. There is a survey with text input in the Project folder for this purpose.
- Use the ELT approach: First import the data into your database. For MySQL, use the LOAD command, as it is very powerful.
- You should transform the data within the database using the functions of the database server according to the ELT principle. In SQL, you can transform data using the INSERT ... SELECT statement.
- Preprocessing is possible as long as the majority of the transformation takes place in the database server.

4. Database analysis

- Write at least one database query for your use case to gain insights for decision-making.
- The SQL query should contain at least 8 different keywords. E.g.:
    - Join tables (JOIN)
    - Processing value sets into a single value (aggregation, e.g., AVG)
    - Grouping aggregates by dimensions (GROUP BY)

    - Filtering data records (WHERE)
    - Selecting attributes (SELECT)
    - Etc.
- Describe in detail what your database query does and how it works. Show how you can calculate the key figures you defined in point 1 using this database query.
- Show the results of the queries and discuss how the results support the use case.

5. Optimize database performance

- Indicate which measures you are taking to optimize query speed
- Collect data to determine whether the optimization is working: Analyze execution plans before and after optimization. Measure runtimes before and after optimization.

6. Visualize and apply results

- Install the BI tool (Metabase) and connect it to both databases. Display the configuration you are using.
- Display the analysis results interactively in the BI tool using database queries, i.e., with parameters, in a graphical format so that they can be used optimally in decision-making.
- Show how a user can interactively work with the visualization of the SQL and NoSQL queries.
- Show that the output of the SQL and NoSQL queries and visualizations is the same.
- Demonstrate in a practical way how the visualization and the original use case are related: Use a user to demonstrate how the visualization improves a specific decision with data.
- Based on the key figures, make a recommendation for an exemplary data-driven decision for the use case.

7. Write a project report

- Write a technical report about your project.
- Make sure to include the **names and email addresses** of the authors, the **name of the team**, and the **title of your project** on the title page. The title page serves as binding confirmation that all team members have contributed to all parts of the project and therefore agree to a joint team grade.
- Maximum length: **40 pages** (excluding title page, table of contents, and appendix).
- Finally, include a reflection and lessons learned on the topics of database technology, project management, teamwork, and the use of artificial intelligence.
- All AI Tools that have been used must be clearly disclosed.
- Create a PDF file (no zip files, etc.) with your report. Only the content in the single PDF file of your project report will be considered for evaluation (excluding the appendix).
- Upload the PDF file to the submission folder on ILIAS. One submission per team is sufficient. If multiple submissions are made, the submission of the first person in alphabetical order will be evaluated.
- **Submission: Final Submission**. At the end of the semester, submit the final status of the project in good time, which will count towards the module grade.

8. Administrative

- The written report in PDF format, and only this, will be evaluated based on the criteria specified in the project assignment.
- As a rule, four students work together on all parts of a project (team assessment). If this is not possible, or if no agreement can be reached, teams of three, teams of two, or individuals may also work on a project. However, the scope of the project remains the same.
- In the event of team conflicts, a meeting between the entire team and the lecturer must be organized as soon as possible.
- If a team has to be split up, all former members may only use the results they have produced independently until the team is split up. All members may use researched public information (publications, data sets, tools, etc.) as long as it is correctly referenced.
- Repeat students must develop a completely new project with a completely new team.

9. Assessment criteria and expectations

Your project work will be assessed based on the following criteria.

Project idea:
- The use case is relevant and coherent.
- The decision support is based on a clear decision rule with a mathematically defined key figure.
- Describe 2+ independent but integrable data sources that enable the use case and are also necessary for the use case.

Data model:
- The source data analysis shows the structure and content using concrete examples.
- The conceptual model explains the essential entities, relationships, and attributes of the use case graphically.
- The database schema in DDL defines the structure of the database, including correct data types, primary and foreign keys in third normal form in SQL

Loading & transformation:
- The loading processes are traceable and performant with MySQL LOAD
- The transformations of the data into the target schema are meaningful and scalable in SQL

Analytics:
- Data analysis has an appropriate level of complexity with 8+ keywords in SQL.
- Data analysis relates directly to the use case in order to calculate the key figure for the decision rule.

Optimization:
- Performance optimization uses 3+ database approaches to increase execution speed
- Execution speed is measurably faster after optimization.

Visualization:
- The interactive BI dashboard in Metabase displays the key figure graphically and parameterized based on the calculation in the MySQL database
- The decision rule is applied to a (fictitious) user based on the key figure in order to positively influence a decision in the use case.

Lessons learned.
- One project-based insight will enable better work with databases in the future.
- One project-based insight will enable better project organization in the future.

- One project-based insight will enable better teamwork in the future.
- One project-based insight will enable better use of artificial intelligence in future project work. Also, all used AI tools are clearly disclosed.

Project report:
- The formal criteria (<40 pages, valid source references, file format, submission deadline) are met.
- The project report is well structured and clearly written.

10. Possible points per criterion

For each criterion, from the project idea to the project report above, including the general expectations, 0 to 3 points are awarded based on the following evaluation logic:
- 3 points: Expectations were met
- 2 points: Expectations were mostly met
- 1 point: Expectations were partially met
- 0 points: Expectations were not met

Important note: The expectations per criterion can only be met if the following evidence of performance is provided where possible:
- *Verifiable online system:* Wherever possible, all results must be implemented in a database system accessible via the Internet or intranet. In the report, provide the URL, user name, password, and any other parameters in the section on system architecture so that the implementation can be verified during the evaluation.
- *Description:* All results, together with the solutions used to achieve them, must be described in a clear and understandable manner. It is therefore not only important what the result is, but also how you arrived at it.
- *Code:* Wherever possible, all results are documented with complete, executable database code!
- *Screenshots:* Wherever possible, the implementation in the database is documented with screenshots of the results of the code execution in the report!

Anyone interested in the detailed evaluation can request access by email after the grades have been published.

11. Possible data sources (validated)

7k Books with Metadata
https://www.kaggle.com/datasets/dylanjcastillo/7k-books-with-metadata

Amazon Books Reviews
https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews

Anime Dataset 2023
https://www.kaggle.com/datasets/dbdmobile/myanimelist-dataset

Book-Crossing: User review ratings
https://www.kaggle.com/datasets/ruchi798/bookcrossing-dataset

Breached_dataset
https://www.kaggle.com/datasets/dannyduncan/breached-dataset

Common Password List (rockyou.txt)
https://www.kaggle.com/datasets/wjburns/common-password-list-rockyoutxt

Data from traffic counts for motorized private transport (hourly values), since 2012
https://data.stadt-zuerich.ch/dataset/sid_dav_verkehrszaehlung_miv_od2031

Food.com - Recipes and Reviews
https://www.kaggle.com/datasets/irkaal/foodcom-recipes-and-reviews

Game Recommendations on Steam
https://www.kaggle.com/datasets/antonkozyriev/game-recommendations-on-steam

Gold spot prices
https://www.gold.org/goldhub/data/gold-prices

Number of crimes by type of offense for selected municipalities and urban districts of the city of Zurich, since 2009
https://data.stadt-zuerich.ch/dataset/ktzh_pks_straftaten_tatbestandgruppe_gemeinden_stadtkreise

Nutritional values for common foods and products
https://www.kaggle.com/datasets/trolukovich/nutritional-values-for-common-foods-and-products

Manga & Anime dataset 2024
https://www.kaggle.com/datasets/duongtruongbinh/manga-and-anime-dataset

MeteoSwiss Open Data:
https://www.meteoschweiz.admin.ch/service-und-publikationen/service/open-data.html

Music & Mental Health Survey Results
https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results

Open Transport Data Swiss Actual Data Archive:
https://archive.opentransportdata.swiss/actual_data_archive.htm

PEN America's Index of School Book Bans
https://docs.google.com/spreadsheets/d/1slCpqLprPXHM-Wyt-WYJR30-NvbGLialVNR8qTsZFG8/edit?gid=0#gid=0

Population by month, city district, gender, age group, and origin
https://data.stadt-zuerich.ch/dataset/bev_monat_bestand_quartier_geschl_ag_herkunft_od3250

RecipeNLG (cooking recipes dataset)
https://www.kaggle.com/datasets/paultimothymooney/recipenlg

Spotify One Million Playlists
https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge

Spotify One Million Tracks
https://www.kaggle.com/datasets/amitanshjoshi/spotify-1million-tracks

Spotify Playlists

https://www.kaggle.com/datasets/andrewmvd/spotify-playlists

Steam Games Dataset
https://www.kaggle.com/datasets/fronkongames/steam-games-dataset

Streaming Service Price History
https://www.kaggle.com/datasets/webdevbadger/streaming-service-prices/data

The GDELT Project
https://www.gdeltproject.org/data.html

Top games on Twitch 2016 - 2023
https://www.kaggle.com/datasets/rankirsh/evolution-of-top-games-on-twitch

Top Video Games 1995-2021 Metacritic
https://www.kaggle.com/datasets/deepcontractor/top-video-games-19952021-metacritic

Video game sales with ratings
https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings

yFinance – Download market data from Yahoo! Finance's API
https://github.com/ranaroussi/yfinance

This list is based on past student projects that have been successfully completed. It is not exhaustive or binding. You are welcome to research your own data sources.