HSLU Lucerne University of Applied Sciences and Arts

`

Guidelines for Semester Projects
Data Warehouse and Data Lake

Lucerne University of Applied Sciences and Arts (HSLU)

Version: August 2024

This document provides guidelines for writing scientific work at Lucerne University of Applied Sciences and Arts.

Luis Terán, José Mancera & Jhonny Pincay

1. Table of Contents

2. Evaluation Components

Table 1: Evaluation Components

Evaluation Component	Content/Description	Evaluation Criteria	Weight
Project Proposal	A detailed proposal covering project objectives, background, business questions, methodology, table of contents, and references.	- Objectives and Background (25%) - Problem Statement (20%) - Business Questions (20%) - Methodology (20%) - Table of Contents and References (5%)	5%
Mid-term Presentation	Presentation of the business idea and current project progress, including data acquisition and data pipelines.	- Confidence of Speakers (Rehearsal recommended) (10%) - Organization (20%) - Quality of Information (60%) - Time Usage (10%)	10%
Final Presentation	Presentation of the final project and prototype with participation from all students.	- Confidence of Speakers (Rehearsal recommended) (10%) - Organization (20%) - Quality of Information (60%) - Time Usage (10%)	15%
Project Implementation	Delivery of a functional data lake prototype as a module outcome.	- Functionality - Design and Usability - Technical Complexity	20%
Project Report	A comprehensive and well-organized report (25 pages) synthesizing the project, excluding indexes and references.	- Clarity and Structure - Depth of Analysis - Relevance and Justification of Decisions - Writing Quality	50%

3. Selection of a Project

Students' suggestions for research topics are highly encouraged. However, the proposed topic must be approved by the examiners. The chosen project topic should align with the interests of both the student and the supervisor. An appropriate project topic is crucial for maintaining motivation and ensuring effective organization.

A. What Must Be Considered in the Proposal?

The following points must be considered in every **proposal.**

Table 2: Proposal Structure

Section	**Details/Requirements**
Title Page (Proposal and Report)	• Name of the university, department, and chair • Provisional title of the Project Report • Author details (name, family name, matriculation number, postal and email address) • Examiners • Place and date
Background and Motivation	Explain the relevance and importance of the topic, providing a clear rationale for the research.
Problem Statement and Research Questions	Clearly define the problem and formulate 3 to 5 research questions. Justify why each question is relevant and how it will be addressed, using appropriate research methods (e.g., literature review, case studies, or empirical study). (**Tip: Define research questions in Objective-Key-Results format with measurable metrics.**)
Data Sources	Briefly describe the data sources that will be used in the research.
Objectives and Scope	Outline the objectives and scope of the Project Report, specifying what is to be achieved within the work.
Audience	Identify the intended audience for the research and its outcomes.
Methods	Describe the research methods to be used, including steps and approaches (e.g., qualitative, quantitative, mixed methods).
Timeline	Provide a detailed timeline (approximately 5 months) that includes sub-goals, activities, and steps.
Table of Contents of Final Report	Present a provisional table of contents with at least two hierarchical levels.
Literature	Cite at **least ten relevant sources** that will be referenced in the research.

4. Data Sources Requirements

The project requires the use of at **least two dynamic data sources** as follows:

- **Dynamic Data Sources**: These are data sources accessed through an API or web scraping, where the data changes (e.g., Twitter, Twitch, Instagram, YouTube, etc.). **It is important to note that all endpoints from a single API vendor will count as one dynamic data source, regardless of the number of endpoints used.**

  Recommendation: https://github.com/public-api-lists/public-api-lists

- **Static Data Sources (optional):** These include static APIs, CSV files, HTML scraping, or data stored in S3 buckets that may consistently return the same data (e.g., World Bank, Stock Reports, HTML pages, Kaggle datasets).

5. Evaluation of the Project Report

B. Submission of the Project report

The Project Report must be submitted electronically as a PDF via ILIAS, and the corresponding GitHub repository should be cited in the references. Additionally, the Project Report must explicitly state if the results are intended to remain unpublished due to confidentiality or non-disclosure agreements.

C. General Evaluation Criteria

The key criteria elements (**Table 3**) are general guidelines for instructors when **awarding points** in the **Final Report Evaluation Grid (Table 5)**.

Table 3: Key Criteria for Evaluating Project Report

Criteria	Description
Structure and Development	The report follows a logical structure, stays focused on the problem, and avoids redundancies or unnecessary content.
Focus on Topic and Questions	The research and business questions are clearly defined and consistently aligned with the central problem.
Depth and Breadth	The report strikes a balance between covering all essential aspects and delving into core topics with necessary depth.
Content Quality	The report provides a comprehensive overview of the topic, supported by diverse sources, reflecting deep engagement.
Technical Quality	Demonstrates a solid grasp of technical aspects and methods, with appropriate application and logical derivation of results.
Clarity and Readability	The report is well-written, clear, concise, and free of grammar or spelling errors, maintaining appropriate language throughout.
Formal Aspects and Accuracy	Adheres to academic standards, with accurate citations, clear distinction between original and sourced ideas.
Originality	The report features a high level of original work, including unique data, programming, and visualizations.
Complexity	The report addresses a challenging topic with complex content, methods, and technical aspects.
Outcome	Successfully answers the research questions and achieves the project's objectives.

D. Final Report Components

Table 4: Expected Elements in the Final Report

Criterion	Expectation
General expectations for technical requirements	• Verifiable online system: Wherever possible, all results are implemented on a database system, S3, etc., accessible via the Internet or intranet. Specify the URL, user, password, and, if necessary, other parameters in the report so that the implementation can be checked during the evaluation. • Description: The results and the ways to achieve them are described understandably in prose. So it is important not only what the result is, but also how they came to it. • Code: All results are backed up with complete executable code wherever possible and available on a GitHub repository. • Screenshots: The implementation of all results is backed up with screenshots in the report wherever possible.
1. Project idea and use case	• Topic introduction • Motivation: explain why you have chosen this project • Problem definition & Goals • Business/Research questions: present the questions that your data lake implementation will answer. • Business intelligence questions: present the questions that your data warehouse implementation will answer.
2. Data Lake	• Architecture: Describe the system setup and all elements needed to answer your business questions (data lake) • Data: Explain why the data sources support your use case (data lake). Describe the data sources. Analyze the data sources for potential limitations.
3. Data Ingestion	• Data Source Description: Describe the data sources, specifying whether they are dynamic or static, and include any relevant insights about the REST API service, such as API quota limitations. • ETL Process Explanation: Provide a detailed explanation of the ETL processes, ideally represented with flow diagrams that describe each step.
4. Data Storage	• Data Storage Services: Specify the services used for data storage and the formats employed (e.g., Parquet, CSV) in cases such as S3 Buckets. • Relational Databases: Provide a brief description of the tables, including any relationships via primary keys, and include a diagram if applicable. • NoSQL or Vector Databases: Briefly describe the type of data stored and the indexing methods used for data retrieval.
5. Data Transformation	• Explanation of the transformations needed in the data and ideally represent them in a flow diagram to understand the sequence of steps • Insights in case of data imputation techniques or cleaning and assumptions around data • Data Quality: any processes considered to validate the quality of the data (Data Warehouse Transformations)
6. Data Warehouse	• Architecture: Describe the system setup and all elements needed to answer your business questions (data warehouse) • Data preparation: Explain why the data sources support your use case (data warehouse). Describe the data sources. Describe the ETL/ELP process required. • Data warehouse modeling: present the model used and implementation of your data warehouse.
7. Data Visualization	• Theoretical model: Describe the theoretical framework for the implementation of visualizations. • Data Warehouse visualizations: present visualizations, based on the theoretical framework, how the data warehouse answers the business intelligence questions proposed)

8. Conclusions	• Proper discussion of the solution. Highlight the advantages, disadvantages, trade-offs in terms or technologies used etc. • Proper discussions of the project outcomes. • Future work: areas to improve the project. What would you do different?

Figure 1 illustrates the minimum expected blocks to complete your project.

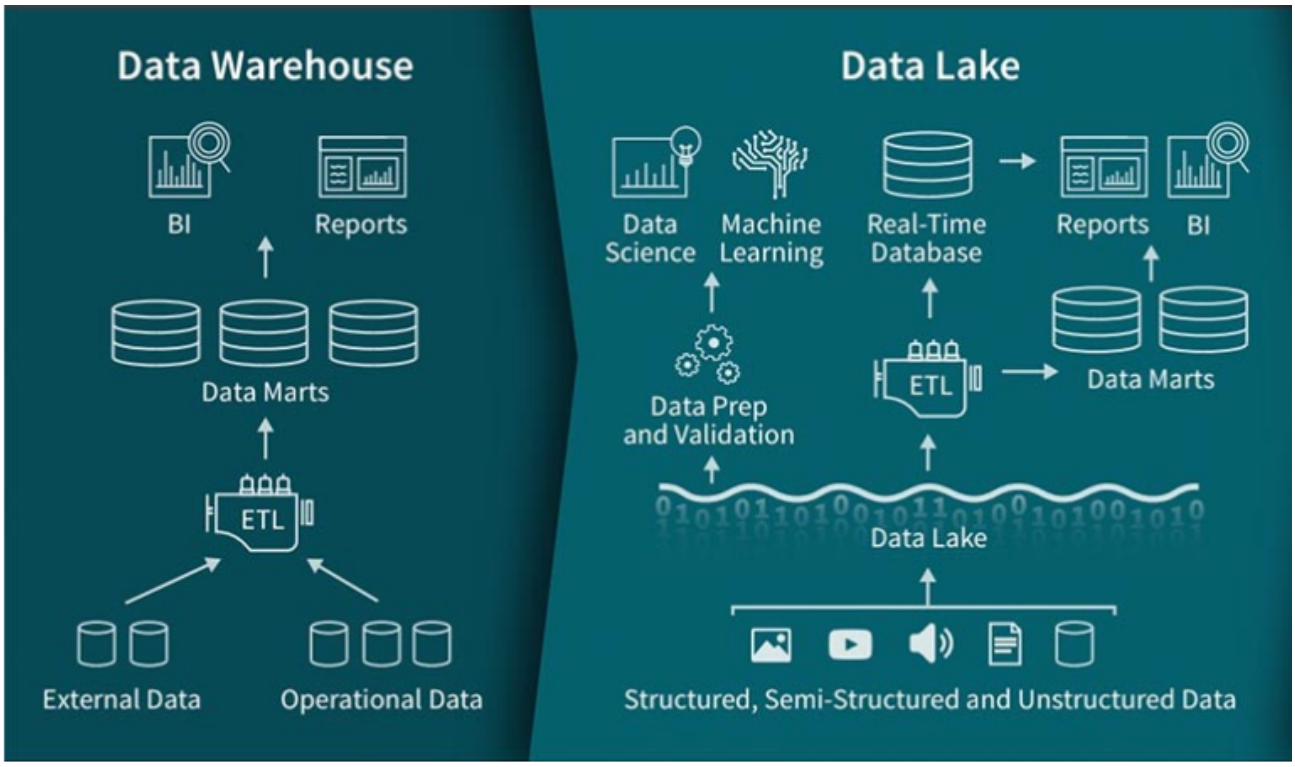

Fig 1. Data Lake & Data Warehouse Architecture

Please note that **Machine Learning Pipelines and real-time ETLs may be optional, depending on API quotas or hardware limitations that could affect the ability to run certain pipelines locally**.

E. Final Project Evaluation Grid

The following grid shows the detailed evaluation point system for Project Reports.

Table 5: Final Report Evaluation Grid

Evaluation Criteria	Maximum possible points	Student 1	Student 2	Student 3
1. Introduction	**5**			
a) Topic introduction b) Motivation c) Problem definition & Goals d) Business/Research questions e) Business intelligence questions			X	
2. Data Lake	**5**			
a) Architecture b) Data Assumptions		X		
3. Data Ingestion (Data Lake)	**5**			
a) ETL Pipelines			X	
4. Data Storage	**5**			
a) Databases and/or storage services		X		
5. Data Ingestion (Data Warehouse)	**5**			
a) ELT Pipelines.		X	X	
6. Data Warehouse	**5**			
a) Architecture b) Data Preparation c) Data Warehouse Modeling			X	
7. Data Visualization	**5**			
a) Theoretical model b) Data Warehouse Visualization			X	
8. Conclusions	**5**		X	
a) Proper discussion of the solution b) Proper discussions of the project outcomes. c) Future work	5	X	X	X

F. Evaluation of Prototypes (Code Review & Requirements)

The prototype is considered a complement to your report. It shows the result of the implementation. The following criteria are considered when evaluating the technical implementation.

Table 6: Technical Implementation Criteria

Criteria	Description
GitHub Repository	All code must be delivered through a referenced GitHub repository. Public repositories are accepted by default. For solutions involving confidential information, students must add instructors as collaborators for visibility.
Repository README File	The repository MUST include a README file written in markup language. This file should explain technical aspects such as how to run the code, file descriptions, goals, and authors.
Code Readability	Code should have an intuitive, easy-to-follow structure, with code separated into logical functions. Variable and function naming should adhere to PEP8 style guidelines.
Code Execution	The code should run without errors. Instructors reserve the right to perform random checks and execute the code if necessary.

Code Style (Extra)	The code will be reviewed for software engineering best practices, including the use of iterators, generators, classes, mutations, modularity, race conditions, exception management, code tests, and function documentation.
Code Project Leads Table	Students **MUST** include a table specifying which team members were responsible for each task, simulating real-world scenarios where code ownership is clearly defined. (see **Table 7**)

Table 7 is an example that **MUST** be added to the annex in your report and evaluation sheet (provided by the instructors)

Table 7.- Code Project Leads

Responsibilities	Student 1	Student 2	Student 3
Introduction			
■ API Data Source 1 (Ingestion)	X		
■ API Data Source 2 (Ingestion)		X	
■ Additional Data Sources (Ingestion)			X
■ Databases or S3 services	X	X	X
■ Any other scripts	X	X	
■ Jupyter Notebooks (Pipeline - Design)		X	
■ …			

G. Grade Criterion

Decisions for the evaluation are 1.) the requirements in the project task document, 2.) the technical quality and quantity of your technical solution, and 3.) the conceptual description of your solution. In detail, 0 to 5 points are awarded per criterion based on the following evaluation logic:

Table 8.- Grade Points Criterion

Points	Evaluation Criteria	Detailed Description
5	Expectations are exceeded	- All expectations are met, and additionally: - More solutions are delivered than required. - Technical results are exemplary for other students. - The description of the solution is publishable as is.
4	Expectations are met	- All solutions are delivered according to the project requirements. - Technical results are correct, with no errors. - The description of the solution is professionally written.
3	Expectations are not met, but results are sufficient	- Most solutions are delivered according to requirements. - Most technical results are correct, with only minor errors. - The description of the solutions is clear.
2	Results are insufficient	- Most required solutions are missing. - Most technical results are incorrect or contain errors. - The description of the solutions is unclear.
1	Results are useless, but one point is awarded for effort	- All required solutions are missing. - All technical results are incorrect. - The description of the solutions is incomprehensible.
0	Missing	- The criteria are not addressed at all in the project report.

6. Plagiarism

All submitted project reports will be rigorously checked for plagiarism using specialized software and web-based services. These tools scan the Internet, indexed books, and other sources to compare content with your report. Any report found to contain plagiarism will receive a failing grade of 1.0 (very poor). In severe cases, disciplinary action may be taken.

Using tools like **ChatGPT** for code or text structure inspiration is allowed, but remember that your reports will be carefully reviewed. AI hallucinations or text inconsistencies will be appropriately penalized even if the effort was made

"Non-quoting" and "copy-pasting" will not pay off!

7. Literature

- Searching, reading, and analyzing **existing literature** on the research topic are essential for any scientific research. Every author of a project report is expected to research the relevant literature *systematically* and *carefully*.
- It can be distinguished between the following categories of scientific **literature** or works:
    - textbooks
    - research papers in scientific journals
    - research papers of scientific conferences and congresses
    - habilitations and dissertations
    - Bachelor/Master Theses (Diploma Theses) and Seminar Thesis, and
    - scientific contributions to the World Wide Web.

Every category of literature has certain **advantages** and **disadvantages** (Table 9).

Table 9: Categories of Scientific Literature[5]

Category of literature	Advantages	Disadvantages
Textbooks	■ easy to find ■ comprehensive ■ comparative	■ often not up-to-date ■ often not specific ■ not available for new fields
Research papers (journals)	■ topic oriented ■ up to date	■ narrow (author is focused on the strengths of his own idea)
Research papers (conferences)	■ topic oriented ■ very up to date	■ narrow ■ not mature
Habilitation & **dissertation**	■ methodical founded, specific ■ established in research	■ too specific ■ not adequate for own research
Bachelor- & **Masterarbeit**	■ manifold & high quantity ■ discussion of the literature	■ quality not guaranteed ■ often low scientific contribution
Text in the World Wide Web	■ very easy & quick to find ■ easy to copy ■ very up to date	■ quality not guaranteed at all ■ information overload ■ reference not stable

8. Collaboration and “Project” Management

The following points should help to improve the **collaboration** between the student and the supervisors/lecturers.

- For questions, please email **ALL** supervisors clearly and concisely.
- For technical issues (e.g., configurations or code), share a GitHub repository or consider using a video to explain the problem (e.g., with Loom: https://www.loom.com).
- Depending on the complexity, supervisors may schedule a 1:1 live session if questions cannot be addressed during the Q&A lecture.

9. References and Bibliography

The sources used (literature) in a Project Report must be **quoted using any format; nevertheless, it is highly recommended to use an engineering format (e.g., IEEE).** An example of quotation format is presented in the following way:

Author, *year of publication* and *pages* in brackets.

THREE EXAMPLES:

- “Eine Aktivität ist eine betriebliche Tätigkeit mit einem definierten Ergebnis. Sie wird von Menschen und/oder Maschinen durchgeführt” **[Österle 1993, p. 13]**.
- In Anlehnung an **[Herbst & Knolmayer 1994]** machen Geschäftregeln Aussagen über dieArt und Weise der Geschäftsabwicklung.
- Ähnliche Definitionen finden sich in **[Bauer et al. 1994, p. 101]**.

Following, you find examples how to quote different sources in reference lists.

THE QUOTED **SOURCE** IS A **BOOK**:

[Booch 1994] Booch, Grady: *Object-Oriented Analysis and Design with Applications*, 2nd edition, Benjamin/Cummings, Redwood City CA, 1994.

[Scheer 1994] Scheer, August: *Wirtschaftsinformatik – Referenzmodelle für industrielle Geschäftsprozesse*, 5. Auflage, Springer, Berlin, 1994.

ARTICLES IN A **BOOK**:

[Huckvale/Ould 1993] Huckvale, Tim; Ould, Martyn: Process Modelling – Why, What and How. In: Spurr, Kathy et al. (Eds.): *Software Assistance for Business Process Re-Engineering*. John Wi-ley & Sons, New York, 1993, pp. 81-97.

[Keller 1995] Keller, Gerhard: Eine einheitliche betriebswirtschaftliche Grundlage für das Business Reengineering. In: Brenner, Walter; Keller, Gerhard (Hrsg.): *Business Reengineering mit Standardsoftware*. Campus Verlag, Frankfurt 1995, S. 45-66.

ARTICLES IN A **JOURNAL**:

[Bunger/Heß 1995] Bungert, Winfried; Heß, Helge: Objektorientierte Geschäftsprozeßmodellierung. *IM – Information Management*, Jg. 10, Heft 1 (Februar 1995), S. 52-63.

[Wastel et al. 1994] Wastel, David; White, Phil; Kawalek, Peter: A methodology for business processre-design – experiences and issues. *Journal of Strategic Information Systems*, Vol. 3, No. 1 (1994), pp. 23-40.

Articles in a **Conference Proceedings** publsihed by a publisher:

[Herbst 1995] Herbst, Holger: A Meta-Model for Specifying Business Rules in Systems Analysis.In: Iivari, J.; Lyytinen, K.; Rossi, M. (Eds.): *Advanced Information Systems Engineering*; Proceedings, 7th International Conference, CAiSE '95, Jyväskylä, 12-16 June 1995. Sprin-ger-Verlag, LNCS 932, Berlin 1995, pp.186-199.

[Kueng 1995] Kueng, Peter: Ein Vorgehensmodell zur Einführung von Workflow-Systemen. In: Schweiggert, Franz; Stickel, Eberhard (Hrsg.): *Informationstechnik und Organisation – Planung, Wirtschaftlichkeit und Qualität; Proceedings*, Ulm, 28./29. September 1995. Berichte des German Chapter of the ACM, Band 47, Teubner-Verlag, Stuttgart 1995, p. 185-203.

Articles in a **Conference Proceedings** not publsihed by a publisher:

[Kohl 1994] Kohl, Claudia: Die Anwendbarkeit von OO-Konzepten in der Unternehmensmodellier-ung. In: *Proceedings des EMISA/MobIS-Fachgruppentreffen*, Universität Münster, 13./14. Okt. 1994, S. 51-53.

[Yu/Mylopoulos 1994] Yu, Eric; Mylopoulos, John: Using Goals, Rules, and Methods To Support Rea-soning in Business Process Reengineering. In: *Proceedings of the 27th Hawaii International Conference on Systems Sciences*, HICSS'94, Vol. IV, pp. 234-243.

Articles from the **World Wide Web**

[Snowdon 1997] Snowdon, R.A.: *Overview of Process Modelling*. available: http://www.cs.man.ac.uk/ipg/Docs/pmover.html, accessed 13th March 1997.

[WfMC 1996] WfMC: *Workflow Management Coalition – Terminology & Glossary*, version 2, June 1996. available: http://www.aiim.org/wfmc/standards/index.htm, zugegriffen am 26. Ok-tober 1998.