Fachbereich 4: Informatik
Forschungsgruppe E-Government

Adresse: Universitätsstr. 1; D-56070 Koblenz

Koblenz, <9 August 2024>

Vereinbarung zur Betreuung einer Masterarbeit / *Agreement to a Master thesis*

Name Student	Matr.Nr.	Studiengang / Study Programme
Md Naiem Siddique	219203013	Web and Data Science

Titel der Abschlussarbeit / Title of Thesis	Metadata Catalog for Data Assets in Lakehouse Architecture: A Literature review and prototype development.

Betreuerin/ Supervisor	Prof. Dr. Maria A. Wimmer
Zweitbetreuer*in/ Second Supervisor	Dr. Ulf Lotzmann

Ausgangspunkt, Problemstellung und Zielsetzungen der Abschlussarbeit /
Point of departure, problem scope and aim of thesis (ca ½-1 Seite / page):
Starting point of a discussion

In the field of information and technology, data has emerged as a critical asset. Big tech companies craft personal user data as techno-economic objects (i.e user metrics) which has the potential to add to their revenue stream [1]. Modern applications and services are increasingly data-centric and data-intensive. The big data value chain [2] includes wide range of domains such as customer profiling, identifying user needs, facilitating information exchange, enabling machine learning and predictive analysis, assessing and enhancing business value, monitoring system performance and integrity, ensuring security and many more. In short, the digital ecosystem is heavily dependent on data and on suitable data architectures [3]. Therefore, among the main purposes of this research is to investigate and analyze emerging big data architectural concepts such as data lake [4], data warehouse, data Lakehouse, data mesh etc. and how data assets in such concepts can be made findable and accessible to different stakeholders.

A crucial success factor [5] of digital ecosystems is the effective management of data assets and its value chain [6]. As data is a valuable currency in the digital world, this paper aims to investigate the development and usage of a comprehensive data catalog [7] for indexing, cataloging, and sharing data assets. While several proprietary managed services such as Azure purview, Collibra, Unity catalog, Google data catalog, AWS glue, IBM Watson address this capability [8], this research however focuses on providing guidelines for improving data interoperability and discover-ability with open-source cataloging and management technologies.

A case study using publicly available COVID-19 datasets is included to provide clarity on the concepts discussed. The state funded COVID and AI expert system [9] aims to build a knowledge base founded on the available data [10] related to COVID infection and assist policy makers to make effective decisions to contain the virus. This case study will help understanding the value proposition and feasibility of this research. A sample catalog for the expert system will be useful

validating the findings and results. In summary, this research follows a systematic and structured methodology that involves comprehensive data collection, rigorous analysis, validation through expert feedback, and the formulation of documented results. Each research question is addressed using a tailored approach to ensure that the outcomes contribute effectively to the development of a data catalog optimized for the COVID and AI system.

This research aims to address the problem domain of developing a discoverable and interoperable data catalog by studying existing literature, comparing reference models [11],and formulating requirements. The research focuses on scientific literature findings, industry standards, practices, and the use of a Lakehouse architecture as the foundation for the catalog. The paper provides structured recommendations, measurable processes, and steps for developing the catalog, supported by a case study. Additionally, it explores different data architectures, their features, functionalities, advantages, and disadvantages, justifying the selection of the Lakehouse architecture [12] and the COVID and AI system. The research formulates fundamental concepts, connects related studies, and considers the pros and cons of existing research to present implementation-ready guidelines. This master thesis also discusses relevant semantics and terminologies while demonstrating a proof of concept within a case study of COVID and AI system.

Wesentliche Forschungsfragen oder Hypothesen / *Key Research Questions or Hypo*theses
(Beschreiben Sie kurz die wesentlichen (etwa 2-3) Forschungsfragen (oder Kernhypothese(n)), die Ihre Arbeit treiben und die Sie in Ihrer Arbeit behandeln werden. / *Briefly describe your key (ca. 2-3) research questions (or key hypotheses) driving your research work.*

1. What standards, reference data models [13] and best practices exist for metadata catalogs supporting data management?
2. How can reference metadata models be leveraged in metadata management and data governance in Lakehouse ?
3. What are the interoperability and discover-ability requirements for data assets – with a focus on the use case in COVID and AI?
4. What are the steps, procedures and guidelines to build a comprehensive metadata catalog?

Forschungsdesign und erwartete Ergebnisse / *Research design of thesis and expected results*
(Darstellung des Forschungsdesigns: was ist das grundlegende Forschungsparadigma? Mit welchen Methoden werden Sie Ihre Forschungsfragen / Hypothesen bearbeiten? Welche Ergebnisse sollen dabei erarbeitet werden? Ebenso sollen Zusammenhänge zwischen Methoden und/oder Ergebnissen dargestellt werden / *Description of the research design: – what will be the methods you will apply to answer your research questions and what will be the results you expect to generate along each of the methods?* – 1-1½ Seiten/pages)**:**

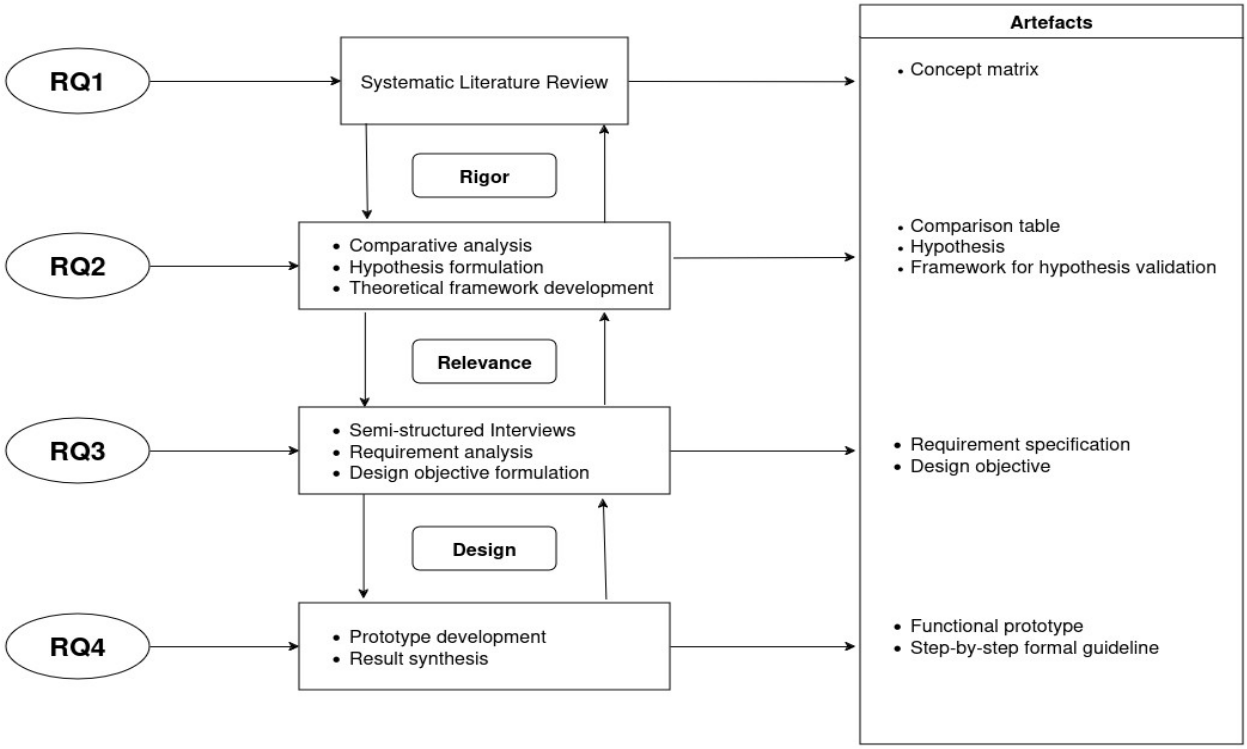

Figure 1: Research design landscape

This thesis is grounded in Design Science Research (DSR) [14] as the foundational methodology, given our objective to design a system for a comprehensive data catalog within a Lakehouse architecture. In addition to DSR, various other research methodologies will be employed to address the specific research questions outlined below:

Research Question 1 (RQ1): Data catalog standards, reference metadata models and best practices, and how these support data management and find-ability

Method: The primary method employed in this research is an systematic literature review [15] conducted within the PICOC (Population-Intervention-Comparison-Outcome-Context) [16] framework. This review will encompass a thorough examination of industry standards, requirements, recommendations, use cases, and best practices relevant to data-intensive systems.

Result: The anticipated outcome is a concept matrix that offers a comprehensive analysis and synthesis of pertinent scientific literature, existing reference metadata models, standards, use cases, and best practices within the domain. Additionally, this analysis will distill and highlight relevant data management and findability features inherent in these concepts.

Research Question 2 (RQ2): Metadata management in Lakehouse

Method: Building upon the literature review, this research question will be addressed by systematically comparing key concepts, techniques, technologies, and frameworks relevant to the study. Following this comparative analysis, research hypotheses will be formulated. Subsequently, a theoretical framework will be developed to validate these hypotheses.

Result: The outcomes of the comparison will be presented in a tabular format. The final deliverables for this research question will include the formulated research hypotheses and the developed validation frameworks.

Research Question 3 (RQ3): Requirement Analysis

Method: This question will be addressed by initiating semi-structured interviews [17] to identify the requirements for our case study and for data-intensive systems in general. Once the requirements have been finalized, design objectives will be formulated based on the specifications derived from the interviews.

Result: Based on the insights gained from the interviews, requirement specification documents will be created. Subsequently, concrete design objectives will be established to address the following research question.

Research Question 4 (RQ4): Prototype Development and recommendations

Method: A prototype of a data catalog will be developed for the COVID and AI case studies, guided by the design objectives outlined in Research Question 3. Additionally, to consolidate and formalize the research findings, a step-by-step guideline for developing comprehensive data catalogs will be created. These guidelines will be based on the evaluation of both theoretical insights and practical knowledge acquired through the research.

Result: The final deliverables will be twofold: 1. a functional prototype of the data catalog for COVID and AI system, and 2. a set of formalized guidelines for constructing data catalog for data-intensive systems.

Schlüsselliteratur / *Key literature* (Auflistung der Schlüsselliteratur (wissenschaftliche Literatur), auf der die Abschlussarbeit basieren wird (mind. 5 Angaben) / *List of key academic literature that forms the basis for your thesis (minimum of 5 entries*). Bitte vollständige Angaben und die Quellen im Text oben verweisen) / Please ensure to document the full citation data and to cite / reference the entries in the above elaborations**:**

[1] Birch, K., Cochrane, D., & Ward, C. (2021). Data as asset? The measurement, governance, and valuation of digital personal data by Big Tech. Big Data & Society. https://doi.org/10.1177/20539517211017308

[2] Becker, T. (2016). Big Data Usage. In: Cavanillas, J., Curry, E., Wahlster, W. (eds) New Horizons for a Data-Driven Economy. Springer, Cham. https://doi.org/10.1007/978-3-319-21569-3_8

[3] Avci, C., Tekinerdogan, B. & Athanasiadis, I.N. Software architectures for big data: a systematic literature review. Big Data Anal 5, 5 (2020). https://doi.org/10.1186/s41044-020-00045-1

[4] Nolte, H., & Wieder, P. (2022). Realising Data-Centric Scientific Workflows with Provenance-Capturing on Data Lakes. MIT Press Direct, 4(2), 426–438. https://doi.org/10.1162/dint_a_00141

[5] C.W Holsapple, M Singh, The knowledge chain model: activities for competitiveness, Expert Systems with Applications, Volume 20, Issue 1, 2001, Pages 77-98, ISSN 0957-4174, https://doi.org/10.1016/S0957-4174(00)00050-6.

[6] Faroukhi, A.Z., El Alaoui, I., Gahi, Y. et al. Big data monetization throughout Big Data Value Chain: a comprehensive review. J Big Data 7, 3 (2020). https://doi.org/10.1186/s40537-019-0281-5

[7] Subramaniam, P. (2021, March 12). Comprehensive and Comprehensible Data Catalogs: The What, Who, Where, When, Why, and How of Metadata Management. arXiv.org. https://arxiv.org/abs/2103.07532

[8] Jahnke, N., Spiekermann, M., Ramouzeh, B., (2022) Data Catalogs, Fraunhofer Insitute for Software and Systems Engineering ISST, <https://www.isst.fraunhofer.de/content/dam/isst-neu/documents/Publikationen/Datenwirtschaft/Fraunhofer-ISST_DataCatalogs_Report-kl.pdf>

[9] AI and COVID, Department of Computer Science 2023, University of Koblenz, accessed 9 August 2024, <https://www.uni-koblenz.de/de/informatik/iwvi/wimmer/projekte/ki-und-covid>

[10] Wahltinez, O., Cheung, A., Alcantara, R., Cheung, D., Daswani, M., Erlinger, A., Lee, M. K., Yawalkar, P., Lê, P., Navarro, O. P., Brenner, M., & Murphy, K. (2022b). COVID-19 Open-Data a global-scale spatially granular meta-dataset for coronavirus disease. Scientific Data, 9(1). https://doi.org/10.1038/s41597-022-01263-z

[11] Assaf, A., Troncy, R., & Senart, A. (2015). HDL – Towards a Harmonized Dataset Model for Open Data Portals. In Proceedings of the 5th International Workshop on Using the Web in the Age of Data (USE-WOD'15) and the 2nd International Workshop on Dataset PROFIling and Federated Search for Linked Data (PROFILES '15) @ ESWC'15 (pp. 62–74).

[12] Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., Torres, J. P., Van Hovell, H., Ionescu, A. M., Łuszczak, A., Świtakowski, M., Szafrański, M., Li, X., Ueshin, T., Mokhtar, M. K., Boncz, P., Ghodsi, A., Paranjpye, S., Senster, P., . . . Zaharia, M. (2020c). Delta Lake: high-performance ACID table storage over cloud object stores. Proceedings of the VLDB Endowment, 13(12), 3411–3424. https://doi.org/10.14778/3415478.3415560

[13] Musen, M. A. (2022, August 4). Modeling community standards for metadata as templates makes data FAIR. arXiv.org. https://arxiv.org/abs/2208.02836

[14] Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105. https://doi.org/10.2307/25148625

[15] Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical report, EBSE Technical Report EBSE-2007-01.< https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf/promis_misc/525444systematicreviewsguide.pdf>

[16] Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd Edition. Chichester (UK): John Wiley & Sons, 2019.

[17] Blandford, A. (2014). Semi-structured qualitative studies. Interaction Design Foundation - IxDF. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/semi-structured-qualitative-studies

Ich bin ich damit einverstanden, dass mein Name mit dem Titel der Abschlussarbeit sowie einer Kurzdarstellung des Exposés auf den Webseiten der Forschungsgruppe unter Qualifikationsarbeiten veröffentlicht wird (während der Ausarbeitung wie auch nach dem Abschluss).

I herewith agree that my name and the title of my thesis as well as an abstract of the thesis will be published at the web page of the research group under the folder qualification works (both, along the elaboration as well as when I will have finished my thesis).

Unterschrift der/s Studierenden
Signature of Student

Unterschrift der betreuenden Professorin
Signature of supervising professor