Bachelor- / Master-Thesis: Refactoring eines Semantischen Data-Lake-Systems (*SEDAR)*

Motivation

Data-Lake-Systeme werden seit einigen Jahren als Repositories vorgeschlagen, in denen heterogene Daten gespeichert und zusammengeführt werden können [1]. Für die Entwicklung solcher Systeme müssen verschiedene Technologien aus den Bereichen Big Data, Datenbanken und maschinelles Lernen kombiniert werden. In den letzten Jahren wurden in mehreren studentischen Projekten verschiedene Elemente eines Data-Lake-Systems namens *SEDAR* entwickelt, die nun in die Produktionsumgebung integriert werden sollen. Die verwendeten Programmiersprachen waren Python (Flask) im Backend und JavaScript (React) im Frontend. Interessierte können sich gerne dieses Paper [3] durchlesen und dieses Video anschauen.

Die Notwendigkeit für ein gründliches Refactoring unseres Systems ergibt sich aus der zunehmenden Komplexität und den vielfältigen Anforderungen, die an moderne Data-Lake-Systeme gestellt werden. In unserem Fall wurden bereits verschiedene Elemente für *SEDAR* entwickelt, die jetzt im Rahmen des BMBF-Projekts i^{2}DACH in die Produktionsumgebung überführt werden sollen. Dies eröffnet die Chance, nicht nur bestehende Codebasis zu optimieren, sondern auch die Systemarchitektur selbst zu überdenken. Das Refactoring des Data-Lake-Systems mit Hilfe von Large Language Models (LLMs) wie GPT-3 kann dazu beitragen, die Robustheit, Effizienz, Skalierbarkeit und Funktionalität des Systems zu verbessern, indem es umfassende Analysen, automatisierte Optimierungen und die Implementierung von Best Practices ermöglicht. Dieser Schritt wird sicherstellen, dass SEDAR den aktuellen und zukünftigen Anforderungen an Data-Lake-Systeme gerecht wird und gleichzeitig die Grundlage für innovative Entwicklungen und Forschung im Bereich der Datenverarbeitung und -analyse schafft.

Der Schwerpunkt der Arbeit kann von den Studierenden in Absprache mit den Betreuern gewählt werden. Mögliche Aufgaben sind im Folgenden beschrieben, sollen sich aber nicht strikt auf diese Ziele beschränken, sondern idealerweise den Prototyp als Ganzes weiterentwickeln:

Aufgaben:

- Einarbeitung in das System
- Recherche, Vergleich und Auswahl geeigneter Tools für Refactoring (z.B. ChatGPT, GitHub Copilot, …)
- Erstellung eines Konzepts für *SEDAR 2.0* (Metadatenmodell, APIs, Datenbanken, Docker-Setup …)
- Einrichtung einer Produktionsumgebung und Aufsetzen einer CI/CD-Pipeline
- Demonstration der Verwendbarkeit und Effektivität der implementierten Erweiterung an einem Use Case aus dem I^{2}DACH-Projekt

Ihr Profil:

- Nachweisbare Erfahrung in Softwareprogrammierung
- Frontend-Programmierung mit JavaScript, vorzugsweise mit den Frameworks React und Angular
- Python OO-Programmierung

Interessiert? Fragen? Kontaktieren Sie uns!

Sayed Hoseini, M.Sc. – sayed.hoseini@hs-niederrhein.de

Prof. Dr. Christoph Quix – Christoph.quix@hs-niederrhein.de

Literatur

[1] C. Quix, R. Hai: Data Lake. In S. Sakr, A.Y. Zomaya (Eds.): Encyclopedia of Big Data Technologies. Springer 2019. https://doi.org/10.1007/978-3-319-63962-8_7-1

[2] Karmaker, Shubhra Kanti, et al. "Automl to date and beyond: Challenges and opportunities." ACM Computing Surveys (CSUR) 54.8 (2021): 1-36.

[3] Hoseini, Sayed et al. SEDAR: A Semantic Data Reservoir for Heterogeneous Datasets. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), 2023. Association for Computing Machinery, New York, NY, USA, 5056–5060. https://doi.org/10.1145/3583780.3614753