# Mögliche Themen für Masterarbeiten

Die hier vorgestellten Themen können bei Interesse von Prof. Dr. Christoph Quix bzw. Sayed Hoseini, M. Sc., am Fachbereich Elektrotechnik und Informatik betreut werden.

Sofern Sie sich für ein Thema interessieren, schicken Sie bitte eine E-Mail an christoph.quix@hs-niederrhein.de und sayed.hoseini@hs-niederrhein.de.

Ihre Anfrage sollte folgende Dinge mindestens enthalten:

- Formale Angaben wie Name, Matrikelnummer, Studiengang, etc.
- Eine kurze Motivation, warum Sie sich für das Thema interessieren
- Eine kurze Erklärung, inwiefern Sie die entsprechend angegebenen Voraussetzungen des Themas erfüllen

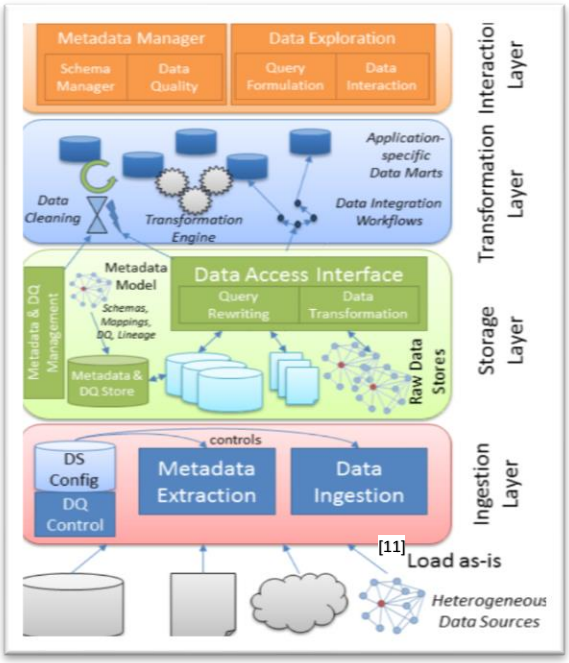
Hochschule Niederrhein
University of Applied Sciences
Elektrotechnik und Informatik
Faculty of Electrical Engineering and Computer Science

# I. Weiterentwicklung eines Data-Lake-Systems

## Motivation

Data-Lake-Systeme werden seit einigen Jahren als Repositories vorgeschlagen, in denen heterogene Daten abgelegt und zusammengeführt werden können. Zur Entwicklung solcher Systeme müssen diverse Technologien aus den Bereichen Big Data, Datenbanken und Machine Learning miteinander verknüpft werden. In den letzten Jahren wurde in mehreren studentischen Projekten verschiedene Elemente eines Data-Lake-Systems entwickelt, welche nun in die Produktivumgebung integriert werden sollen.

Als Programmiersprachen wurden *Python* (*Flask*) im Backend und *JavaScript* (*React*) im Frontend verwendet. Der Fokus der Arbeit kann in Absprache mit den Betreuern vom Studierenden ausgewählt werden. Mögliche Themen werden im Folgenden beschrieben, sie sind aber nicht strikt auf diese Zielsetzungen zu beschränken, sondern es soll idealerweise der Prototyp als Ganzes weiterentwickelt werden. Es ist geplant, das System im Projekt *I²DACH* an die Anlagen des HIT-Instituts anzubinden. Die Anlagen produzieren verschiedenste Sensor- u. Maschinendaten mit Bezug zu Versuchen aus der Lackchemie, welche über REST-APIs bezogen werden können. Im Folgenden werden einige generische Aufgaben aus verschiedenen Ebenen des Systems beschrieben, welche weiterentwickelt werden können, die nächste Seite enthält Beschreibungen zu konkreteren Themen:




**Interaction:**

- Weiterentwicklung von Benutzerschnittstellen zum Abfragen von Daten und Metadaten
- Visualisierung von Datensätzen und Ontologien
- Access Management, Data Governance, Authentifizierungsmöglichkeiten (z.B. OAuth mit *KeyCloak*)

**Transformation:**

- Erweiterung der Transformationsworkflows, insbesondere für semistrukturierte Datensätze (JSON, XML)
- Integration von Operatoren für Datenqualität und Machine-Learning-Algorithmen

**Storage:**

- Verwaltung und Erweiterung der Metadaten, der während der Ingestion extrahierten Datensätze
- Verwaltung und Suche einer geeigneten Ontologie für die Chemie-Domäne

**Ingestion**

- Ingestion von Maschinendaten mitsamt ihren Metadaten

## Ablauf der Abschlussarbeit

Die Aufgabe wird zu Beginn im Detail mit den Betreuern ausgearbeitet und die Ziele der Arbeit genau definiert sowie ein Zeitplan aufgestellt, welcher in einem Exposé festgehalten wird. Es wird eine VM für die Entwicklung des Systems sowie eine GitLab-Instanz zum Verwalten des Programmcodes bereitgestellt. Die Struktur des bereits existierenden Programmcodes wird zu Beginn im Detail erklärt um den Einstieg zu vereinfachen. Während des Projekts finden regelmäßige Treffen zwischen den Betreuern und den Studierenden statt, bei denen Zwischenergebnisse präsentiert und diskutiert werden. Das System soll zunächst mit verschiedenen heterogenen Datensätzen getestet werden und dann idealerweise für einen Anwendungsfall am HIT-Institut eingesetzt werden.

# Konkrete Themen für die Weiterentwicklung des Data-Lake-Systems

## 1. Weiterentwicklung zu einem *Scientific Data Lake*

In [2] bezeichnen die Autoren ein Data-Lake-System, welches für Data-Science-Aktivitäten eingesetzt wird, als einen *Scientific Data Lake.* Weitere Literatur wie zum Beispiel [3], beschreibt wie man das Metadatenmodell eines Data-Lake-Systems erweitern kann, um Analyseverfahren effektiv erfassen und beschreiben zu können. Ziel einer Abschlussarbeit in diesem Bereich wäre die Implementierung dieser Ideen um die Auswertung Analyseverfahren und das Trainieren von Modellen innerhalb des Systems zunächst zu ermöglichen, zu vereinfachen und dokumentieren zu können. Die Plattform „*Data Robot AI*“ [4] könnte hier als Inspiration dienen. Das System sollte idealerweise alle Phasen eines gängigen Machine-Learning-Projekts unterstützen (in [4] Prepare, Build, Deploy, Predict, Monitor, Optimize). Basis für diese Arbeit könnte die Open-Source-Bibliothek von Hopsworks [5] oder MLFLow [6] sein. Weitere (möglicherweise Open-Source-)Plattformen für MLOPs sollten zunächst evaluiert werden.

**Aufgaben**:

- Metadatenmodell für Datenanalyse erweitern, Training, Prediction und Deployment von ML-Modellen ermöglichen, Metadaten erfassen, Ingestion von externen ML-Projekten
- TensorFlow/PyTorch-on-Spark Lösung in das System integrieren, z.B. [12]
- Empfehlungssystem für Machine-Learning-Modelle implementieren.:
    - Feature Engineering in das System integrieren (z.B. Data Augmentation, Train/Test Split, Labelling, Spark-PCA, …)
    - Algorithmen vorschlagen basierend auf Datentypen (Kategorisch/numerisch, Bild, Text, CSV, etc.), Problemtyp (Un-/Supervised, Regression/Klassifikation), Metrik, …
    - Training des Modells auf automatisch generiertem Dockerimage
    - Automatische Evaluation des Modells nach benutzerdefinierter Metrik

## 2. Management eines Data-Lake-Systems mittels Semantic Web Technologien

In den letzten Jahren wurden einige Bespiele für den Einsatz von Technologien aus dem Bereich des Semantic Web für Metadatenmanagement eines Data-Lake-Systems veröffentlicht [7,8]. Die Idee ist es, Metadaten (von z.B. Datensätzen) mit Instanzen aus Knowledge Graphs zu annotieren um den Datenkatalog des Systems um Semantik anzureichern. In aktuellen Projekten wird bereits erste Arbeit in diese Richtung geleistet. Ziel der Abschlussarbeit wäre es, den aktuellen Stand der Forschung in diesem Bereich zu ermitteln und die Nutzbarmachung von Ontologien innerhalb des Systems zu erhöhen. Zum Beispiel könnte durch Werkzeuge zur automatisierten oder besonders benutzerfreundlichen Annotation von Datensätzen entwickelt werden. Insbesondere sollte erforscht werden, inwiefern man die innere Struktur des Graphen verwenden kann, um zum Beispiel weitere Daten für das Trainieren eines ML-Modells zu finden. Als Inspiration kann das Datenintegrations-Tool *Karma* [10] betrachtet werden.

**Aufgaben**:

- Annotation von Metadaten im Frontend erleichtern und Möglichkeiten erweitern
- Integration der Metadatenstruktur mit Data Catalog Vocabulary (DCAT)
- Integration der Metadatenstruktur mit Online-Ontologien (z.B. DBPedia, WikiData)
- Definition einer Schnittstelle, mit der man externe Annotationen mitsamt ihrer Datenquelle in das System aufnehmen kann
- Datenquellen mit Semantik (über SPARQL) ausfindig machen
- Modelle zur automatischen Annotation von Tabellen in das System integrieren (siehe [8])

# II. Machine Learning Projekte am HIT-Institut

Das HIT Institut hat im letzten Jahr eine hochmoderne High-Throughput-Formulation-Screening-Anlage (HTFS) in Betrieb genommen, welche Pionierarbeit auf dem Weg zur digitalen Chemie (Chemie 4.0) leisten wird. Hier wird massiv auf Algorithmen aus dem Maschinellem Lernen gesetzt, welche die Anlage automatisiert steuern, selbstständig Versuche durchführen und Ergebnisse zusammentragen sollen.

Die genaue Problemformulierung wird dem Studierenden überlassen, in enger Diskussion mit den Betreuern und den Chemikern aus dem HIT. Verschiedene Szenarien für die Analyse und Bewertung von Farb- u. Lackproben mittels Neuronaler Netze können geprüft und ein vielversprechender Ansatz implementiert werden.

Als Referenz können folgende Arbeiten betrachtet werden:

1. Gaoyuan Zhang, Sayed Hoseini, Christian Schmitz, Matthias Fimmer, Christoph Quix, *Deep Learning based Automated Characterization of Cross-cut Tests for Coatings via Image Segmentation*, https://link.springer.com/article/10.1007/s11998-021-00557-y

2. Erkennung von Defekten auf Lackoberflächen mit maschinellem Lernen, Bachelorarbeit, Benjamin Danker, September 2021, Prüfer: Prof. Dr. Christoph Quix

   Die Arbeit wird auf Anfrage herausgegeben.

# 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] R. Hai, C. Quix, and M. Jarke. "Data lake concept and systems: a survey." arXiv preprint arXiv:2106.09592 (2021).

[3] Zhao, Yan, et al. "Analysis-oriented Metadata for Data Lakes." 25th International Database Engineering & Applications Symposium. 2021.

[4] Data Robot AI Plattform, https://www.datarobot.com/

[5] https://www.hopsworks.ai/ , https://github.com/logicalclocks/hopsworks#what

[6] https://www.mlflow.org/

[7] Schmid, Stefan, Cory Henson, and Tuan Tran. "Using knowledge graphs to search an enterprise data lake." European Semantic Web Conference. Springer, Cham, 2019.

[8] Dibowski, Henrik, et al. "Using Semantic Technologies to Manage a Data Lake: Data Catalog, Provenance and Access Control." *SSWS@ ISWC*. 2020.

[9] Nguyen, Phuc, Ikuya Yamada, and Hideaki Takeda. "MTabES: Entity Search with Keyword Search, Fuzzy Search, and Entity Popularities", 2021.

[10] Yun, Hongyan, et al. "Research on Multi-Source Data Integration Based on Ontology and Karma Modeling." International Journal of Intelligent Information Technologies (IJIIT) 15.2 (2019): 69-87.

[11] Chihoub, Houssem, et al. "Architecture of Data Lakes." Data Lakes 2 (2020): 21-39.

[12] https://github.com/yahoo/TensorFlowOnSpark