A novel approach to the storage of legal and medical data for machine learning

Degree programme: BSc in Computer Science
Thesis advisor: Prof. Dr. Erik Graf
Industrial partner: legal-i AG, Bern

This thesis investigates the implementation of data storage solutions for unique legal and medical data intended for machine learning applications within the real-world startup industry context.

Introduction

Big data and machine learning demand increasingly complex data storage solutions. Data Lakes, Lakehouses, and Meshs attempt to accommodate large enterprises' data needs by maintaining sophisticated data storage and management systems. These technologies have become so prevalent that developing a data storage and management solution without them seems antiquated. However, this assumption presents a pitfall for smaller enterprises or those with unique and novel data requirements.
Thus, this thesis aims to assess and implement a data storage solution in a real-world industry context, provided by the Swiss-based insure-tech start-up legal-i. Legal-i has a unique collection of sensitive medical and legal data, which must be stored in strict compliance with GDPR. The startup collects, stores, and maintains this data as training material for multiple machine learning models, powering the customer-facing services the company offers.

Methodology

This thesis and project leveraged a „fail-fast" and „iterate quickly" approach. The primary goal was to create a storage solution supporting data storage, and secondly, to facilitate the maintenance and management of training data. With this main target in mind, an MVP (Minimum Viable Product) was developed. With the MVP defined, the team worked towards its implementation using a defined tech stack. The challenges encountered during the implementation sprints were discussed, noted as lessons learned, and subsequently informed the final solution's requirements.

Results

Guided by the literature review carried out in the Project 2 preceding this bachelor thesis, the initial solution legal-i sought to implement was the well-documented and widespread data lake approach. Implementing such a solution using Apache Airflow and AWS's Data Lake-focused services caused a radical reassessment of legal-i's requirements for its data storage solution. The evaluation revealed that the company needed a more integrated and straightforward solution within its application. The use-case was inverted, with the storage solution no longer acting as an agent siphoning off data from the application, but the application serving and siphoning data to and from the storage solution. Additionally, the data modeling method was changed. The schema used in the data storage now emphasizes a holistic entity modeling a single information unit, rather than being agnostic to the data's relations.

Outlook

The implementation of the requirement specification obtained through lessons learned during the initial solution's implementation is ongoing and intended to conclude in the first quarter of 2024. Beyond the initial project's scope of addressing storage and machine learning use-cases, the current solution implementation is anticipated to offer additional user-facing features facilitated by the reorientation of the project.

Aimé Ehi
Data Engineering

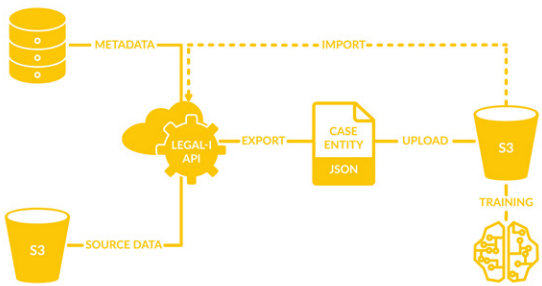

Diagram of the holistic case entity's data flow, showing its export, import, and training set generation.