Cyprus University of Technology
Department of Electrical Engineering and Computer Engineering and Informatics

Bachelor Thesis

Data Lake Semantic Enrichment via Traditional Systems and LLMs

Panagiotis Papageorgiou

Limassol, May 2025

CYPRUS UNIVERSITY OF TECHNOLOGY

Faculty of Engineering and Technology

Department of Electrical Engineering, Computer Engineering, and Informatics

Bachelor Thesis

Data Lake Semantic Enrichment via Traditional Systems and LLMs

Panagiotis Papageorgiou

Advisor: Dr. Andreas S. Andreou

Limassol, May 2025

Copyrights

Acknowledgements

I would like to thank everyone who supported me throughout this journey as an undergraduate student and during the preparation of this thesis. Firstly, I want to express my gratitude to my supervisors, Dr. Michalis Pingos and Dr. Andreas Andreou, for giving me the opportunity to work on such an interesting research topic. I also thank them for their insightful ideas and guidance throughout this research. Special thanks to Artemis Photiou for collaborating in the first stage of research and experiments and for making the journey more enjoyable.

ABSTRACT

In todays era, Big Data is often referred to as the ”new oil”. Businesses heavily rely on Data Lakes to store massive amounts of heterogeneous data, however without proper metadata mechanisms in place, these repositories turn into data swamps. This thesis explores alternative systems for traditional semantic enrichment of data sources by adapting a pre-existing semantic blueprint model in Apache Hive and comparing it in terms of insertion time, query performance and storage efficiency against a well established system, Apache Jena. Experimental results show that Hive offers significant scalability and storage efficiency benefits over Jena, however Jena is more suitable for small to medium-size Data Lakes that require dynamic schema evolution, complex relationships between data sources and perform infrequent queries for metadata retrieval. Additionally, this thesis explores the feasibility of LLM-driven approaches for semantic enrichment by proposing two novel pipelines and evaluating four different configurations. The results demonstrate that LLMs can be used as an alternative solution but often rely on high quality metadata to produce maximum accuracy. Expert-curated metadata produced the highest accuracy and low response times, while LLM-generated metadata offered a promising, semi-automated alternative with important trade offs. Finally, FAISS-based retrieval excelled in reducing operational costs as well as response times.

Keywords: Data Lakes, Semantic Enrichment, Large Language Models, FAISS.