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

Bachelor Thesis

An Evaluation of Apache Jena and Apache Hive for Data Meshes: Using Semantic Enrichment and LLMs for data products recommendations

Artemis Photiou

Limassol, May 2025

CYPRUS UNIVERSITY OF TECHNOLOGY

Faculty of Engineering and Technology

Department of Electrical Engineering, Computer Engineering, and Informatics

Bachelor Thesis

An Evaluation of Apache Jena and Apache Hive for Data Meshes: Using Semantic Enrichment and LLMs for data products recommendations

Artemis Photiou

Advisor: Dr. Andreas S. Andreou

Limassol, May 2025

Copyrights

Acknowledgements

I would like to express my sincere appreciation to everyone who supported and encouraged me throughout the course of this thesis. I am especially grateful to my professors, Dr. Andreas Andreou and Dr. Michalis Pingos for their ideas, guidance, and valuable feedback during the development of this work. Their support played an important role in shaping the quality of the thesis. I would also like to thank my collaborator Panagiotis Papageorgiou for working with me during the first part of the project. The foundation we built together formed a significant part of this research.

ABSTRACT

This thesis investigates two key challenges in the world of Data Lakes: The semantic organization of metadata and the automated generation of meaningful data products using an LLM. The first part of the thesis evaluates which system is most suitable for implementing a metadata enrichment mechanism, specifically comparing Apache Hive and Apache Jena in terms of scalability, query efficiency and storage performance. Experimental results demonstrate that Hive outperforms Jena in large scale environments. The second part of the thesis proposes a framework in which an LLM autonomously suggests data products by providing to the LLM a user-defined concept, metadata retrieved from Hive and sample records from HDFS. The purpose of this study is to examine whether the LLM can act as a domain expert by reasoning over structured and unstructured input and by performing external web searches. Specifically, the framework is evaluated across two domains and three complexity levels, measuring the precision and quality of the suggested data products. The results show that while the LLM performs well in simple scenarios, its effectiveness declines as concept complexity and dataset pool size increases.

Keywords: Semantic Enrichment, Apache Hive, Apache Jena, Data Lakes, Large Language Models