EURA NOVA

FACULTIES OF SCIENCE AND ENGINEERING

Master Thesis Offers 2020-2021

For the latest information, please visit https://research.euranova.eu.
22nd April 2020

CONTENTS

EURA NOVA

INTRODUCTION

EURA NOVA is a data-driven Belgian company founded in September 2008 and located in Brussels, Marseille, and Tunis. Our mission is simple: bring life to our customers' great ideas, by offering best-in-class services in data science, software engineering, and data architecture. To do so, we invest significantly in in-house expertise and state-of-the-art knowledge. In line with this course of action, we offer academic programs in collaboration with universities. These offers include bootcamps, master theses topics, research internships, and PhDs topics. See below for details.

OUR MASTER THESIS OFFERS

This document presents master theses topics supervised by our research & development department. Each project is an opportunity to be actively involved in the development of solutions to address tomorrow's challenges in ICTs and to implement them today.

The students will work in a dedicated **international** team of engineers **with diverse expertise in machine learning, graph theory, artificial intelligence, high performance computing, etc.**

They will keep EURA NOVA informed of the project advancement and share their ideas and challenges using the in-house knowledge management tool. We value continuous learning and teamwork. We love to have a good time together. For more information on our R&D activities, please visit our website at https://research.euranova.eu.

HOW TO APPLY

When you have gone through our master thesis offers, pick your favourite three. Draft a short text for each one, stating why you find it interesting and what you would do about it. Send us this statement, along with your CV at career@euranova.eu.

MACHINE LEARNING & DATA SCIENCE

ONTOLOGY-BASED GDPR COMPLIANCE ANALYSIS

Context
With the activation of GDPR, multiple EU projects emerged to consolidate its concepts and rules, in order to help organizations and users better understand its definitions, requirements and facilitate compliance checking. Many projects have released ontologies/knowledge bases that cover specific parts of GDPR or Privacy Polcies. However, it is not clear if among all these solutions there exists one that is suitable to be used in automatic compliance checking [ACMM+19]. To be applicable in an actual data project, the ontology needs to satisfy a set properties, such as completeness and reasoning support. The examples of the ontologies that can be analyzed are the following:

- PrOnto/DAPRECO [RBP+19, PMRR18]
- GDPRtEXT [PFOL18]
- GDPRov [PL17]
- PrivOnto [OPS+17]
- ODRL and its extensions [DKPS, JGM+]

Business Opportunity
The analysis done in this project is going to provide practical and argumented comparison between existing Linked Data solutions for GDPR compliance and offer the optimal framework to be used in data-centric projects.

Contribution
The project includes the following objectives:

- Study state of the art in Data Privacy and GDPR ontologies and knowledge bases.
- Design an evaluation framework (e.g. reasoning queries) to compare selected ontologies in terms of industrial applicability to GDPR compliance.
- Evaluate selected ontologies on the test documents using the designed framework.
- Analyse the results and compare the completeness and usability of selected ontologies.
- (Optional) propose the formalization that consolidates the analyzed ontologies and completes them with missing elements for automatic GDPR compliance.

A TAXONOMY OF GRAPH EMBEDDING METHODS

Context

Graph embedding has been an intensive area of research in the last 3 years, with a significant number of contributions. This is probably because graph embedding has promising applications in retail, customer analytics, bioinformatics, medicine, etc. However, the high number of contributions makes the work of graph practitioners difficult since we get quickly lost in this jungle of algorithms. The objective of this thesis is to build a taxonomy of the existing approaches: can we cluster the algorithms in families ? Can we build a decision tree to select the best graph embedding technique, based on the characteristics of a problem? The focus of this project is more on the theoretical properties of existing work. However, there is a possibility to publish a paper highlighting this taxonomy.

https://github.com/benedekrozemberczki/awesome-graph-classification

Business Opportunity

EURA NOVA has been investing on graph for more than 9 years. This is an important topic for Machine Learning but also in Data management Governance. The embedding techniques represent a real opportunity to better model interaction between data. We have applications in retails, banking, insurance, telecoms and even aeronautics.

Contribution

The objective is to create a complete taxonomy of Graph Embedding approach, and then, recommend which family must be applied according to the usage and context.

- Explore the state of the art of Graph Embedding methods
- Study the different approaches and potentially implement test bench to compare them
- Find key difference, pros and cons
- Propose a taxonomy and recommend their usage according to the context of the ML task objective and data set

EVALUATING PSO WITH THE STATE OF THE ART MINIMIZATION METHODS FOR AUTOML

Location:
Mont-Saint-Guibert, Belgium

Context
During the last 3 years we have seen a new generation of Machine Learning frameworks arriving in the state of the art. Their objective is to automate the selection of the set of algorithms to put together in order to maximize the results on the machine learning task objective (as classification or regression). These systems are called AUTOML. Given a data set and a downstream ML task, find the right set of feature engineering functions, algorithms and hyper parameter optimizations that will lead to the best results on the given data set. This is modeled as the CASH problem (Combined Algorithm Selection and Hyper-parameter optimization). We can clearly see two winning approaches emerging from the state of the art: sequential Bayesian optimizations (SMACK) and evolutionary approaches. However, we have seen recently the resurgence of well known optimization methods applied in ML tasks, as the Particle Swarm Optimization. The objective of this thesis is to develop a new AutoML technique based on Particle Swarm Optimization (PSO) and to compare this method to the state of the art, namely SMAC and Evolutionary methods.

Business Opportunity
EURA NOVA is exploring in its current research track the integration of multi-modal data sets in AUTOML tasks. As a result, we are interested in the impact of PSO or any other optimization methods on AUTOML pipeline.

Contribution
The objective is to evaluate the PSO compared with the state of the art methods:

- Explore the state of the art to find the most up to date optimization methods
- Implement a test bench where the student will be able to evaluate and compare existing approaches
- Implement a PSO approach to the CASH problem and investigate the implication of PSO on the CASH problem
- Propose to extend the approach to the recommendation problem

DISTRIBUTED DATA PROCESSING

EXPLORE INTERESTING PATTERNS IN STREAMS

Context

Streaming context where users have already defined a workload that consists of multiple CEP queries written in the language we are developing now in order to discover some interesting situations or system malfunctions. Or only a set of raw streams where events are flowing into a data lake, and analysing these streams in order to discover interesting situations, by correlating the events of the different streams, is needed online and offline (similar to Exploratory OLAPs).

Contribution

The objective of this master thesis is to discover some interesting patterns for the user based on their workload, and then to evaluate the effectiveness of these recommended patterns (represented in our pattern algebra).

COMPLEX EVENT PROCESSING BENCHMARK

Context

Despite the fact that there are many frameworks nowadays that can provide complex event processing features, there is no standard mechanism to assess and evaluate these frameworks quantitatively. With the recent advancements, especially on the distributed processing part, there is an increasing need for a solid benchmark that provides a biggest possible set of metrics that can help to:

- Evaluate newly implemented systems;
- Discover the best fit for a specific use-case given a set of requirements and the available resources.

Contribution

The objective of this master thesis is to design and implement a benchmark that is able to assess CEP frameworks. Some **but not all** of the metrics that must be provided are: throughput, latency, memory consumption, correctness of results.

Implementation

Build a complete pipeline that generates data and route it into a black box (CEP framework to be assessed), in order to collect statistics and summaries, and then generate a set of KPIs

CREATE A USER-FRIENDLY EVENT QUERY LANGUAGE (EQL) INTERFACE

Context

The language we are developing now in the LEAD track has the objective of covering as many operators needed in the market as possible. Thus, the focus for the time being is centered around covering and improving the current state of the art by providing unambiguous algebra and a physical execution plan that is mapped to a full stream processing engine's operators. However, we are still missing an important aspect in the work that is user friendliness. Having a user friendly language/interface is of a great importance for our work to get to industry. The goal of a user-friendly interface is to provide a good user experience by ascertaining the following attributes:

- Simple. A user-friendly interface is not overly complex, but instead is straightforward, providing quick access to common features or commands;
- Clean. A good user interface is well-organized, making it easy to locate different tools and options;
- Intuitive. In order to be user-friendly, an interface must be make sense to the average user and should require minimal explanation for how to use it;
- Reliable. An unreliable product is not user-friendly, since it will cause undue frustration for the user. A user-friendly product is reliable and does not malfunction or crash.

Contribution

The objective of this master thesis is to improve the language grammar or in other words make it sweeter for human use so things can be expressed more clearly, more concisely, or in an alternative style that some may prefer. This is supposed to be achieved by spreading some syntactic sugar here and there, and introducing a big set of defined functions that would make users' lives easier. Additionally, design and implement a user-interface that will help the user express their needs easily and more concisely.

RETHINKING MICROSERVICES AND OTHER SERVICES ORCHESTRATION WITH DIGAZU IN REACTIVE STREAMING

Context

digazu[1] is a batch and real-time data supply chain developed by EURA NOVA. It is designed to fulfil customers' big data needs, such as real-time reporting, data integration, data science models industrialization… digazu is built onto cutting-edge technologies such as Kafka, Flink, Kubernetes, HDFS, ensuring reliability, efficiency and scalability in production environments. While digazu was not initially designed to orchestrate microservices, its architecture is in some points similar to the architecture used in reactive streaming patterns used to orchestrate microservices[2].

Business Opportunity

Many customers are moving gradually to microservices. Designing orchestration and implementation patterns with a tool like digazu would help customers accelerate their transformation and help them define good practices and standards.

Objectives

The objectives of this master thesis are the following:

- Study state of the art microservices orchestration patterns
- Analyse gaps and opportunities for such orchestrations using digazu
- Implement additional modules in digazu for microservices orchestration
- Define standards and good practices to orchestrate microservices using digazu

[1] https://digazu.eu/

[2] https://medium.com/capital-one-developers/microservices-when-to-react-vs-orchestrate-c6b18308a14c

IMPLEMENTING GDPR COMPLIANT SQL PROXY IN SQL STREAMING ON FLINK FOR DIGAZU

Context

digazu[3] is a batch and real-time data supply chain developed by EURA NOVA. It is designed to fulfil customers' big data needs, such as real-time reporting, data integration, data science models industrialization… digazu is built onto cutting-edge technologies such as Kafka, Flink, Kubernetes, HDFS, ensuring reliability, efficiency and scalability in production environments.

Business Opportunity

Integrating a GDPR compliant SQL proxy in SQL streaming on Flink in the transformation layer of digazu would help customers define processes around data that are GDPR compliant and help demonstrate compliance with GDPR requirements.

Objectives

The objectives of this master thesis are the following:

- Analyze how ABAC models can be used in the context of GDPR and metadata management
- Implement a GDPR compliant SQL proxy in SQL streaming on Flink in the transformation layer of digazu
- Define standards and good practices to define and maintain metadata for GDPR compliance

[3]https://digazu.eu/

REAL-TIME DATA CLEANING ON FLINK FOR DIGAZU

Context

digazu[4] is a batch and real-time data supply chain developed by EURA NOVA. It is designed to fulfil customers' big data needs, such as real-time reporting, data integration, data science models industrialization… digazu is built onto cutting-edge technologies such as Kafka, Flink, Kubernetes, HDFS, ensuring reliability, efficiency and scalability in production environments.

Business Opportunity

Data quality is an important factor to leverage data science models. Integrating real-time data cleaning capabilities in digazu would help customers get more insight out of their data.

Objectives

The objectives of this master thesis are the following:

- Analyze the state-of-the-art data cleaning techniques
- Assess techniques that can be executed in real time
- Implement a data cleaning module on Flink in the transformation layer of digazu

[4]https://digazu.eu/

PSEUDO-ANONYMISATION ON STREAMS FOR DIGAZU

Context

digazu[5] is a batch and real-time data supply chain developed by EURA NOVA. It is designed to fulfil customers' big data needs, such as real-time reporting, data integration, data science models industrialization… digazu is built onto cutting-edge technologies such as Kafka, Flink, Kubernetes, HDFS, ensuring reliability, efficiency and scalability in production environments.

Business Opportunity

Adding pseudo-anonymization capabilities to digazu would help customers be compliant with GDPR regulations and facilitate the data provisioning to data scientists.

Objectives

The objectives of this master thesis are the following:

- Analyse the state of the art pseudo-anonymisation techniques
- Assess how such techniques can be executed in real-time streaming
- Implement a pseudo-anonymization module in digazu

[5]https://digazu.eu/

BENCHMARKING OF THE DIGAZU TECHNOLOGIES

Context

digazu[6] is a batch and real-time data supply chain developed by EURA NOVA. It is designed to fulfil customers' big data needs, such as real-time reporting, data integration, data science models industrialization… digazu is built onto cutting-edge technologies such as Kafka, Flink, Kubernetes, HDFS, ensuring reliability, efficiency and scalability in production environments.

Business Opportunity

Benchmarking the technologies used in digazu would allow to better define the use cases that can be covered by the platform, as well as its limitations.

Objectives

The objective of this master thesis is to benchmark the technologies used in digazu, using different combinations of parameters such as:

- Volume of data
- Data throughput
- Number of topics
- Number of transformations
- Number of CPU's
- RAM
- …

[6] https://digazu.eu/

BENCHMARKING APACHE KAFKA, APACHE PULSAR, AND NATS

Context

Those last 10 years, we have seen emerge new kinds of pub-sub messaging system. They have been designed for architectures such as microservices or paradigms such as stream processing. Those messaging systems offer different guarantees, different features and capabilities (e.g. throughput, latency, schema).

In this master thesis we propose to study Apache Kafka, Apache Pulsar, and NATS.

The study will focus on the comparison of the underlying architecture, the development of a bench, the benchmarking, and the discussions on the results.

Contribution

The objectives of this master thesis are to:

- Study the architecture of Apache Kafka, Apache Pulsar, and NATS.
- Propose benchmarks to highlight the differences.
- Develop an open source bench to run them.
- Running the benchmarks and analyse the results.
- Discuss the results.

Ideally, the master thesis will deliver an open source bench

This master thesis could be presented to the community as a talk, a workshop, a paper, or a blog post.

DISTRIBUTED STORAGE NEW RELATIONAL APPROACHES & NOSQL

Context

These last years, we have witnessed the emergence of a new category of distributed storages. They have the same features as traditional RDBMS while providing the distributed approach of the NoSQL DB. However, their architecture are still different and they offer different tradeoffs.

In this master thesis we propose to study (1) The distributed PostreSQL, (2) VoltDB, (3) NuoDB and to compare it to (1) Cassandra and (2) Druid.

The study will focus on the comparison of the underlying architecture, the performance tests and the discussions on the results.

Contribution

The objectives of this master thesis are to:

- To study the architecture of (1) the distributed PostreSQL, (2) VoltDB, (3) NuoDB
- To study the architecture of (1) Cassandra and (2) Druid
- To propose a benchmark for highlighting the differences
- To propose a test bench architecture and implementation
- To deploy the 5 analyzed framework and to apply the performance test
- To discuss the results and to interpret them

This MT could be proposed as workshop or conference paper.

DATA GOVERNANCE

GDPR-COMPLIANT ACCESS POLICY MANAGEMENT

Context
GDPR is now active in EU and worldwide as soon as a company deals with EU citizen data. Among other things, the GDRP enforce the right to access a data according to a context, a purpose and a legal ground. For instance, a marketeer, can access the table customer profile for marketing purpose, because the user signed a contract. The legal ground can grouped in:

- Contracts
- Explicit opt in
- Legitimate interest
- Regulatory obligation

In a privacy by design how to enforce automatically this legal ground within the access security management ?

In this master thesis, we propose to investigate how a RBAC (Role based Access Control) or an ABAC (Attribute Based Access Control) can be extended in order to support this kind of requirements. The student will study the existing offer in terms of meta data management, security and access policy systems in order to propose a solution.

Contribution
The objectives of this master thesis are to:

- Study the state of the art in meta data management, RBAC and ABAC
- Study the meta data management in Hadoop
- Propose a formalism to represent the concept of legal ground
- Propose an approach to integrate the legal ground within an access security policy

The project will be supported by our legal expert and by an expert in Hadoop. This MT could be proposed as workshop or conference paper.

REFERENCES

[ACMM+19] Maribel Acosta, Philippe Cudré-Mauroux, Maria Maleshkova, Tassilo Pellegrini, Harald Sack, and York Sure-Vetter, editors. Test-Driven Approach Towards GDPR Compliance, volume 11702 of Lecture Notes in Computer Science. Springer International Publishing, Cham, 2019.

[DKPS] Marina De Vos, Sabrina Kirrane, Julian Padget, and Ken Satoh. ODRL policy modelling and compliance checking. Technical report.

[JGM+] Karuna P Joshi, Aditi Gupta, Sudip Mittal, Claudia Pearce, Anupam Joshi, and Tim Finin. Semantic Approach to Automating Management of Big Data Privacy Policies. Technical report.

[OPS+17] Alessandro Oltramari, Dhivya Piraviperumal, Florian Schaub, Shomir Wilson, Sushain Cherivirala, Thomas B Norton, N. Cameron Russell, Peter Story, Joel Reidenberg, and Norman Sadeh. Privonto: A semantic framework for the analysis of privacy policies. Semantic Web, 9:1–19, 05 2017.

[PFOL18] Harshvardhan J. Pandit, Kaniz Fatema, Declan O'Sullivan, and Dave Lewis. GDPRtEXT - GDPR as a Linked Data Resource. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 10843 LNCS, pages 481–495. Springer Verlag, 2018.

[PL17] Harshvardhan J Pandit and Dave Lewis. Modelling provenance for gdpr compliance using linked open data vocabularies. In PrivOn@ ISWC, 2017.

[PMRR18] Monica Palmirani, Michele Martoni, Cesare Rossi, Arianna Bartolini and, and Livio Robaldo. Pronto: Privacy ontology for legal reasoning. In Electronic Government and the Information Systems Perspective, pages 139–152, Cham, 2018. Springer International Publishing.

[RBP+19] Livio Robaldo, Cesare Bartolini, Monica Palmirani, Arianna Rossi, Michele Martoni, and Gabriele Lenzini. Formalizing GDPR Provisions in Reified I/O Logic: The DAPRECO Knowledge Base. Journal of Logic, Language and Information, 2019.