ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

Ω.Π.Μ.Σ.
ΤΕΧΝΗΤΗ
ΝΟΗΜΟΣΥΝΗ

DEMOKRITOS

Προτεινόμενα Θέματα Διπλωματικής Εργασίας

Business Intelligence & Automation Department

Automated Data Categorization in Enterprise Data Warehouse for Privacy Compliance

Introduction:

This project aims to address the challenge of categorizing entities stored in an enterprise data warehouse according to privacy policies. The categorization will be performed at the table level, taking into account the type of customer (Retail customers, Telecom customers, Visitor, prospect...) and the contextual information within each table (e.g., orders information, billing and revenue). Furthermore, specific attributes or columns may require distinct categorizations from their respective entities. For instance, in the case of a Call Detail Record (CDR) table, the table as a whole is categorized as "traffic data," while the columns containing calling or called numbers are categorized as "generic personal data," and the column containing information about the antenna carrying the call is categorized as "location data."

Problem Statement:

The enterprise data warehouse comprises a vast number of tables, exceeding 50,000, each containing multiple columns. Manual categorization of this magnitude is infeasible. Therefore, an automated approach is necessary to achieve efficient and accurate data categorization while maintaining a high level of confidence in the results.

Project Objectives:

- Design and implement an automated data categorization process that assigns a suitable category to each entity (table and column) stored in the data warehouse.
- Ensure the categorization process complies with privacy policies and guidelines.
- Develop a confidence index for each entity/column to determine the level of certainty in the assigned categorization.
- Enable manual validation for entities/columns with low confidence levels, allowing for human intervention when required.
- Provide flexibility in adapting the categorization process to accommodate any manual modifications made to the assigned categories.

- Leverage lineage information from ETL tools (e.g., Oracle Warehouse Builder or Oracle Data Integrator) or database catalog information (e.g., PL/SQL source code, views) to support the categorization process.

- Explore the utilization of machine learning algorithms and sampled pre-categorized data for training purposes.
- Incorporate content profiling of the tables to enhance the accuracy of the categorization process.
- Implement a storage mechanism to record the categorization results per entity/column.
- Develop a mechanism to identify and categorize newly discovered entities within the data warehouse incrementally.

Constraints and Assumptions:

- The project will focus on privacy policy-related categorization and may not cover other data classification aspects.
- The implementation will consider the compatibility and availability of relevant tools and technologies within the existing infrastructure.
- The project assumes access to a sufficient amount of labeled training data for machine learning-based categorization.
- The scope does not include modifications to the data warehouse schema or structure.**Stakeholders:**
- The project involves collaboration between the data warehouse team, privacy and compliance team, and relevant stakeholders from the organization who can provide expertise and domain knowledge for accurate categorization.

Business Intelligence & Automation Department

Predictive Modeling for KPI values using ML

Abstract:

Machine Learning Prediction involve using machine learning models to make informed predictions based on patterns and trends identified in data. These methods analyze historical data to learn relationships between input variables and target outcomes, enabling the model to make predictions on unseen data. Machine learning models can capture complex relationships and nonlinear patterns in the data, allowing for accurate KPI predictions. They can handle large datasets, adapt to changing circumstances, and incorporate multiple variables simultaneously.

The aim of this thesis is to create a Predictive Analytics System that leverages ML techniques to enhance the handling of certain issues in the systemic flow of orders, thereby improving the overall order management process. (Additionally, it aims to develop predictive models that contribute to the timely resolution of order bottlenecks and facilitate the efficient unstacking.)

The proposed system will use 2 data sources:

- A traditional relational database which contains significant order data (e.g., order source, order type, root product etc.)
- Data can be extracted in the desired format by periodically running SQL queries and transforming the results in the desired format (e.g., csv, xls, XML document, JSON and so on).
- Data will be **pulled** by the system.
- A No-SQL database that contains data about the actions performed on a set of stuck orders such as action type (e.g., resubmission of a request to a 3rd party system, abortion of the order etc.), action status (e.g., success, failure etc.) and so on.

NoSQL databases provide a larger count of write operations per second as compared to SQL databases. This is especially useful for logging i.e., save the results of certain actions on a set of orders.

NoSQL databases provide change events that watch collections, databases, or deployments for changes, so they can act as **Event Publishers**.

Data will be **pushed** to the system.

Typically, **NoSQL** databases provide data in **JSON** format.

The system output will be KPI Values Prediction. (KPIs to be defined)

**Required background and skills: ML/AI, Relational & No SQL Databases*

Other requirements: All documents must be written in English.

Customer Systems Department

Leveraging Graph Databases for Enhanced Customer Relationship Management (CRM) Systems

Abstract:

Introduction: Discuss the importance and role of CRM systems in modern businesses and introduce the concept of graph databases. Briefly describe the purpose and structure of the thesis.

Background and Literature Review: Review existing literature on CRM systems, graph databases, and where they intersect. Discuss traditional RDBMS CRM systems and their limitations.

Comparative Analysis of RDBMS and Graph Databases for CRM Systems: Provide an in-depth comparison between RDBMS and graph databases in terms of performance, scalability, data modeling capabilities, and ease of use within the context of CRM.

Graph Databases for Advanced CRM Capabilities:

- Customer Segmentation: Discuss how graph databases can be used to implement advanced customer segmentation strategies in a CRM system, particularly focusing on real-time segmentation.
- Improving Recommendations: Examine the role of graph databases in enhancing recommendation engines within CRM systems.
- Social CRM: Explore the application of graph databases for Social CRM, particularly in analyzing social network relationships among customers.
- Data Integration and Scalability with Graph Databases in CRM Systems.

Scalability Issues:

Examine how graph databases handle scaling issues in large CRM systems, both in terms of data volume and query performance.

Advanced Analytics in CRM using Graph Databases:

- Real-time Analytics: Discuss how graph databases can enable real-time customer analytics.
- Enhanced Customer 360 Views: Study the role of graph databases in providing a holistic, 360-degree view of customers in a CRM system.
- Predictive Analytics: Investigate how graph databases can be used to enhance predictive analytics capabilities in CRM systems.

Privacy and Security Implications in Graph-based CRM Systems: Delve into the privacy and security implications of using graph databases for CRM systems.

Case Studies: Provide real-world case studies where graph databases have been implemented successfully in CRM systems. (e.g., COSMOTE CRM)

Conclusion and Future Work: Summarize the findings and propose future research directions based on the work done in the thesis.

IT SECURITY

Optimizing WAF with ML scenarios

Abstract:

A web application firewall (WAF) protects web applications from a variety of application layer attacks such as cross-site scripting (XSS), SQL injection (SQLi), and cookie poisoning, among others. Attacks to apps are the leading cause of breaches and in fact are the gateway to a company's valuable data. With a WAF in place, a company can block the array of attacks that aim to exfiltrate that data by compromising its internet facing systems.

A WAF operates through a set of rules often called policies. These policies aim to protect against vulnerabilities in the application by filtering out malicious traffic. The value of a WAF comes in part from the speed and ease with which policy modification can be implemented, allowing for faster response to varying attack vectors; during a DDoS attack, rate limiting can be quickly implemented by modifying WAF policies. These managed rules work extremely well for patterns of established attack vectors, as they have been extensively tested to minimize both false negatives (missing an attack) and false positives (finding an attack when there isn't one).

However, managed rules often miss attack variations, also known as bypasses, as static regex-based rules are intrinsically sensitive to signature variations introduced, for example, by fuzzing techniques. All these attack variations are difficult to be detected by the traditional managed ruleset without involving human intervention to monitor and deploy custom rules per each case.

This thesis aims at optimizing the Imperva WAF service with Machine Learning models that focus on web traffic anomaly detection or discovering attack patterns.

Specifically, this thesis scope is to:

- Identify bypasses and malicious payloads without human involvement.
- Develop a machine learning model trained on the good/bad traffic as classified by managed rules and augmented data to provide better detection of attack patterns.
- Provide definition of new policies ruleset based on the ML model feedback.

The deliverables should include:

- Well-documented analysis of the WAF data/logs analysis
- Well-documented implementation of the ML model

- Well-documented development of the ML model’s training process
- Final MSc Thesis document

**Required background and skills: ML/AI models/algorithms, Web Application Attacks (OWASP Top 10)*

Other requirements: All documents must be written in English.

EDR (MS Defender) Analysis and Optimization with the use of AI/ML

Abstract:

Endpoint Detection and Response (EDR) is an endpoint security solution that continuously monitors end-user devices to detect and respond to cyber threats like ransomware and malware. EDR security solutions record the activities and events taking place on endpoints and all workloads, providing security teams with the visibility they need to uncover incidents that would otherwise remain invisible. An EDR solution needs to provide continuous and comprehensive visibility into what is happening on endpoints in real time. EDR uses real-time analytics and AI-driven automation to protect against cyberthreats that get past antivirus software and other traditional endpoint security technologies.

This thesis aims at further utilizing MS Defender capabilities. Besides alerts and blocking actions, that help the security team to better detect, investigate, and respond to threats across all company's endpoints, there is also the need to implement appropriate "aggressive" responses to improve SOAR (security orchestration, automation and response), based on false/true alerts and multiple-subsequent alert patterns, etc.

Specifically, this thesis scope is to:

- Analyse and understand MS Defender's capabilities.
- Use AI/ML techniques to take advantage of MS Defender's capabilities and improve SOAR within the organisation.

The deliverables should include:

- Well-documented analysis of how EDR (i.e. MS Defender) works in terms of alerting, detecting and blocking.
- Well-documented analysis of how AI/ML can use MS Defender's augmented data for optimizing SOAR.
- Final MSc Thesis document

**Required background and skills: ML/AI, EDR Basics*

Other requirements: All documents must be written in English.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

Δ.Π.Μ.Σ.
ΤΕΧΝΗΤΗ
ΝΟΗΜΟΣΥΝΗ

DEMOKRITOS

Combine the use of Elastic X-Pack with OTE Group’s log collection systems and build use cases with the MITRE ATT&CK knowledge base

Abstract:

Elasticsearch is a distributed, free and open search and analytics engine for all types of data, including textual, numerical, geospatial, structured, and unstructured. X-Pack is an Elastic Stack extension that provides security, alerting, monitoring, reporting, machine learning, and many other capabilities.

MITRE ATT&CK is a globally accessible knowledge base of adversary tactics and techniques based on real-world observations. The ATT&CK knowledge base is used as a foundation for the development of specific threat models and methodologies in the private sector, in government, and in the cybersecurity product and service community.

This thesis aims at using X-Pack’s AI/ML capabilities for analysing ingested logs from OTE Group’s log collection mechanisms to better prevent, detect and respond along with improving threathunting techniques.

Specifically, this thesis scope is to:

- Analyse ingested logs with X-Pack.
- Build use cases based on MITRE ATT&CK knowledge base.
- Utilize these use cases for threathunting as a prevention mechanism.

The deliverables should include:

- Well-documented analysis of using X-Pack for exploring and analysing logs.
- Well-documented analysis of the developed use cases and how they can be used for threathunting.
- Final MSc Thesis document

**Required background and skills: Elasticsearch/X-Pack, ML/AI, MITRE ATT&CK, Threat Modelling*

Other requirements: All documents must be written in English.

Penetration Testing with AI

Abstract:

A penetration test (pen test) is an authorized simulated attack performed on a computer system to evaluate its security. Penetration testers use the same tools, techniques, and processes as attackers to find and demonstrate the business impacts of weaknesses in a system. Penetration tests usually simulate a variety of attacks that could threaten a business. They can examine whether a system is robust enough to withstand attacks from authenticated and unauthenticated positions, as well as a range of system roles. With the right scope, a pen test can dive into any aspect of a system.

This thesis aims at taking advantage of AI for penetration testing purposes.

More specifically, this thesis scope is to:

- Use AI to find vulnerabilities and attack paths in Windows domains and internal networks.
- Try to automatically exploit vulnerabilities with any technique possible.
- Escalate privileges to domain administrator or SYSTEM according to shortest or weakest path possible.

The deliverables should include:

- Well-documented analysis of using AI for detecting vulnerabilities and attack paths.
- Well-documented analysis of the exploitation techniques used for privilege escalation.
- Final MSc Thesis document

**Required background and skills: Pentest basics, AI/ML*

Other requirements: All documents must be written in English.