Proposta de uma arquitetura de Data Warehouse para análise de SDN e aplicações de Aprendizado de Máquina

Fernando Luiz Moro, Rodrigo Nogueira, Alexandre Amaral, Ana Paula Amaral

Instituto Federal de Educação, Ciência e Tecnologia Catarinense – Campus Camboriú
Caixa Postal 2016 – 88.340-055 – Camboriú – SC – Brasil

fernandoluizmoro@gmail.com, {rodrigo.nogueira, alexandre.amaral, ana.amaral}@ifc.edu.br

Abstract. *This paper presents the proposal of a data warehouse architecture that has as data source and object of study the IP flows and attacks in SDN. The proposal aims to provide a consistent and clean dataset for machine learning applications and any application that wishes to consume data of this feature. For the proposed objectives, a multidimensional database was developed, which is fed by an ETL stage based on the collection of network flows. Among the obtained results is the architecture itself, in which a dataset can be explored through OLAP queries by machine learning applications.*

Resumo. *Este artigo apresenta a proposta de uma arquitetura de data warehouse que tem como fonte de dados e objeto de estudo os fluxos IP e ataques em SDN. A proposta tem como objetivo fornecer um conjunto de dados consistente e limpo para aplicações de aprendizado de máquina e qualquer aplicação que deseje consumir dados desta característica. Para atingir os objetivos propostos, foi desenvolvido um banco de dados multidimensional, que é alimentado por uma etapa de ETL baseada na coleta de fluxos de rede. Dentre resultados obtidos é a arquitetura em si, na qual um conjunto de dados pode ser explorado através de consultas OLAP pelas aplicações de aprendizado de máquina.*

1. Introdução

Uma previsão realizada pela Forrester estimou que 500.000 dispositivos de *IoT (Internet of Things)* seriam comprometidos em 2017 [Moro 2017 *apud* Francis 2017]. Com o objetivo de prevenir e combater os ataques gerados por tais vulnerabilidades, tem sido empregado a integração entre as redes definidas por software (*Software-Defined Networking – SDN*) e as técnicas de aprendizado de máquina.

As redes definidas por software permitem através de um controle centralizado e homogêneo da rede o gerenciamento, a execução de tarefas de detecção e bloqueio de ataques de forma simplificada [Moro 2017 *apud* Ahmad *et al.* 2015]. Dentre as abordagens atuais aplicadas para a detecção de ataques em SDN, se destaca o emprego do aprendizado de máquina que vão além dos tradicionais métodos entrópicos que detectam uma anomalia já em andamento.

Os métodos de aprendizado de máquina permitem descobrir o ataque em uma rede, antes mesmo que este aconteça [Huang 2017]. No entanto, o grande desafio no emprego do aprendizado de máquina é que 80% de todo o esforço computacional é gasto na etapa de pré-processamento de dados [Losarwar 2012]. Um ambiente de *data*

warehouse, por sua vez, permite com que as dimensões sejam exploradas, já com os dados coletados, consistentes e limpos [Nogueira 2017]. Deste modo, quando aplicado os métodos de aprendizado de máquina, estes apenas se designam as suas reais tarefas.

2. Trabalhos relacionados

Com o grande volume de informações geradas, uma das principais causas para o crescimento do número de ataques está nas vulnerabilidades presentes nas atuais tecnologias. Isto mostra que os mecanismos para detectar e bloquear os ataques de redes se fazem necessários. Todavia, as atuais redes de computadores se tornaram complexas e heterogêneas, contendo dispositivos e softwares de inúmeros fabricantes com diferentes tecnologias e interfaces de acesso [Costa 2013].

[Huang 2017] desenvolveu um sistema de identificação de aplicativos que pode ser integrado com um sistema de gerenciamento de *QoS (Quality of Service)* em uma SDN. Em experimentos que resultaram em uma f-medida de 93.48%, o conjunto de dados continha diversas aplicações de teste. [Lopez 2017], ilustra em seu trabalho a avaliação de diversos métodos de aprendizado de máquina e a aplicação do algoritmo *PCA (Principal Component Analysis)*. O principal objetivo é realizar a seleção de atributos em cenários de análise de tráfego, no qual, o melhor resultado foi com seis características em árvores de decisão (97.4%) e o pior resultado foi com sete características em SVM-RFE (80.2%).

Exemplo da integração entre técnicas de *data warehousing* e aprendizado de máquina, é o caso de [Mansmann 2014], que obteve um modelo multidimensional da rede social *Twitter* e desenvolveu um ambiente de *data warehouse* que permitiu a criação de um cubo de dados, bem como a análise de sentimentos. [Nogueira 2017], em uma abordagem similar, desenvolveu um ambiente de *data warehouse* que coleta notícias em tempo real com um algoritmo de aprendizado de máquina que realiza o enriquecimento semântico na etapa de ETL (*Extract, Transform, Load*). O mesmo *data warehouse*, serve como fonte de dados para diversas aplicações através de uma API REST (*Representational State Transfer Application Programming Interface*).

Tomando conhecimento das abordagens da literatura este trabalho foi construído baseado na seguinte hipótese: “É possível desenvolver um *data warehouse* baseado em uma rede definida por software para alimentar as aplicações de aprendizado de máquina?”.

3. Arquitetura proposta de um data warehouse para análise de SDN

A arquitetura proposta neste trabalho está ilustrada na Figura 1. A fonte de dados é obtida através da coleta dos fluxos IP seguindo a metodologia proposta por [Amaral 2015]. Uma vez coletados, é realizada a etapa de limpeza e transformação dos dados, na qual destaca-se principalmente a transformação das datas para o padrão do modelo multidimensional. Posteriormente, é realizada a carga no banco de dados multidimensional, a partir do qual é possível a exploração do cubo de dados através de consultas OLAP (*Online Analytical Processing*). A implementação segue uma arquitetura HOLAP (Hybrid *Online Analytical Processing)* utilizando o servidor PostgreSQL 10.

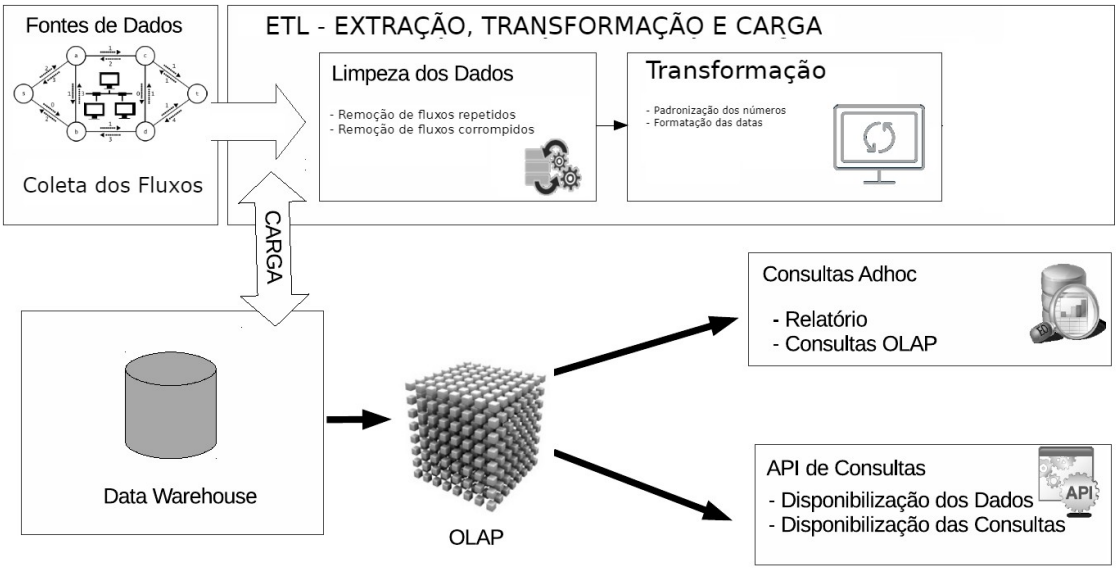

Figura 1. Arquitetura do data warehouse proposta

O banco de dados multidimensional é responsável por consolidar e armazenar os fluxos IP coletados e pré-processados. Para tal, foi utilizado o modelo de estrela proposto por [Kimball 2011] conforme é mostrado na Figura 2. No modelo desenvolvido, o objeto de análise é o fluxo IP, no qual as dimensões fornecem métricas para avaliar o comportamento da rede, principalmente, em cenários de ataque.

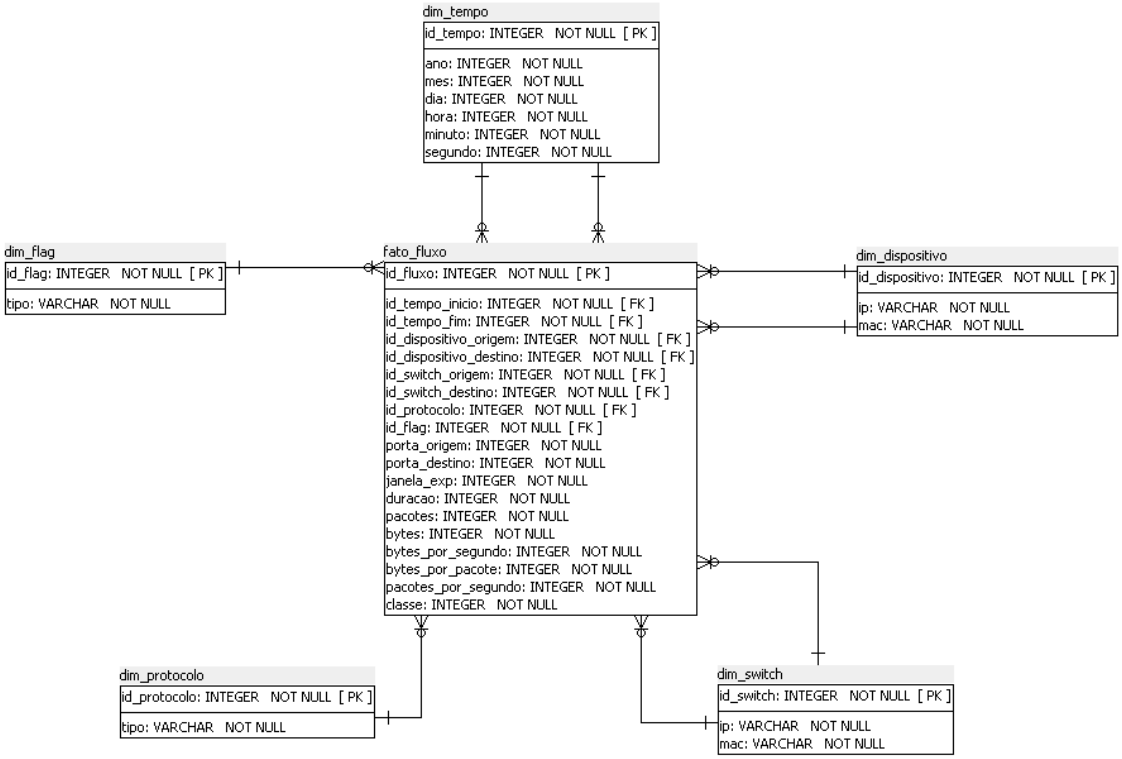

Figura 2. Modelo multidimensional da arquitetura proposta

4. Considerações finais, resultados obtidos e esperados

A partir da arquitetura desenvolvida é possível realizar a exploração do cubo de dados respondendo questões como: “Qual é o IP que mais atacou?”, “Qual a média de ataques?”, “Qual o período que mais houve um ataque?”. Um cubo de dados é amplo, e espera-se que através de sua exploração seja possível realizar experimentos e avaliar o desempenho da arquitetura desenvolvida no emprego de algoritmos de aprendizado de máquina.

Este artigo é fruto de uma pesquisa interdisciplinar em andamento, que integra as áreas de redes de computadores, segurança da informação, banco de dados e inteligência artificial. Deste modo, o que foi apresentado até o momento está em constante desenvolvimento e os experimentos aqui citados consistem em trabalhos futuros.

Referências

Amaral, A. A. (2015). Computação autonômica aplicada ao diagnóstico e solução de anomalias de redes de computadores. Universidade Estadual de Campinas (UNICAMP).

Costa, L. R. (2013). OpenFlow e o Paradigma de Redes Definidas por Software. Universidade de Brasília.

Huang, N.-F., Li, C.-C., Li, C.-H., et al. (2017). Application identification system for SDN QoS based on machine learning and DNS responses. In *2017 19th Asia-Pacific Network Operations and Management Symposium (APNOMS)*. IEEE.

Kimball, R. and Ross, M. (2011). *The data warehouse toolkit: the complete guide to dimensional modeling*. 2nd revised ed. Canada: John Wiley and Sons, Inc.

Lopez, M. A., Lobato, A. G. P., Mattos, D. M. F. and Alvarenga, I. D. (2017). Um Algoritmo Não Supervisionado e Rápido para Seleção de Características em Classificação de Tráfego. In *XXXV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. Sociedade Brasileira de Computação (SBC).

Losarwar, V. and Joshi, D. M. (2012). Data Preprocessing in Web Usage Mining. In *International Conference on Artificial Intelligence and Embedded Systems (ICAIES’2012)*.

Mansmann, S., Ur Rehman, N., Weiler, A. and Scholl, M. H. (2014). Discovering OLAP dimensions in semi-structured data. *Information Systems*, v. 44, p. 120–133.

Moro, F. L., Amaral, A., Amaral, A. P. and Nogueira, R. (nov 2017). Detecção e autorreparo de anomalias em redes definidas por software. In *XVII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais*. Sociedade Brasileira de Computação (SBC).

Nogueira, R. (2017). Newsminer: um sistema de data warehouse baseado em textos de notícias. Universidade Federal de São Carlos (UFSCar).