УДК 004.65:004.738.5

SCALABLE NETWORK TRAFFIC CLASSIFICATION USING BIG DATA FRAMEWORKS

S. Han[1]*, O. Mnushka*[2]

[1]*Master's Student, CEP Dept., NTU «KhPI», Kharkiv, Ukraine*
[2]*Senior Lecturer, CEP Dept., NTU «KhPI», Kharkiv, Ukraine*
sikai.han@cs.khpi.edu.ua

Network traffic classification has become an increasingly critical component in modern network management and security systems. With the explosive growth of internet traffic, the emergence of new applications, and the increasing sophistication of cyber threats, traditional methods of traffic classification have proven inadequate for today's complex networking environment. The challenge of accurately and efficiently classifying network traffic is further complicated by the growing adoption of encryption, dynamic port allocation, and the rapid evolution of application protocols.

The objective is to analyze, develop and test an scalable network traffic classification using big data frameworks.

In recent years, the convergence of big data technologies and artificial intelligence has opened new possibilities for network traffic classification. Big data frameworks such as Apache Hadoop and Apache Spark provide the necessary infrastructure to process and analyze massive volumes of network traffic data in real-time or near real-time. When combined with advanced machine learning algorithms, these frameworks enable more sophisticated and adaptive approaches to traffic classification [1].

Network traffic classification is an essential component of network management and security, enabling the identification of applications, protocols, and potential threats. Early techniques like port-based classification rely on well-known port numbers, but their effectiveness has diminished due to dynamic port allocations, port masquerading, and the rise of applications that use random ports. Deep Packet Inspection (DPI) offers higher accuracy by analyzing packet contents, although it faces limitations with encrypted traffic, computational demands, and privacy concerns.

Statistical-based classification provides insights by analyzing traffic patterns through metrics such as packet size, flow duration, and byte counts based on using various Python tools to simulate and model flow-based, packet-based, and protocol-specific metrics, offering flexibility in analyzing both individual packet and overall flow characteristics.

Modern approaches have incorporated machine learning techniques, significantly improving classification accuracy and scalability. Supervised learning methods, including Support Vector Machines (SVM), Decision Trees, and Neural Networks like CNNs and RNNs, are employed to create models that classify traffic based on labeled data. Unsupervised methods, such as K-Means and Hierarchical Clustering, enable the classification of unlabeled traffic by identifying patterns and clusters within the data.

The integration of machine learning techniques, especially deep learning and clustering algorithms, has enhanced traffic classification capabilities, providing higher adaptability in dynamic network environments. This combination of techniques allows for efficient handling of large datasets, crucial for modern, complex network infrastructures.

References:

1. Najm, I. A. Enhanced Network Traffic Classification with Machine Learning Algorithms / [I. A. Najm, A. H. Saeed, B. A. Ahmad, S. R. Ahmed et al.] // Proceedings of the Cognitive Models and Artificial Intelligence Conference. – ACM, 2024. – AICCONF '24. – С. 322–327. DOI: 10.1145/3660853.3660935.