Big Data Analytics for Real-Time Network Anomaly Detection

Abstract:

In today's world the cyber attacks are becoming more and more complicated and hard to identify. The attacks targeting large networks have become more and more frequent which use advanced methods. Because of this, Traditional cybersecurity systems can no longer useful in detecting these dynamic attacks in real-time due to the large amount of data generated by modern networks. Therefore, for my project I was planning to explore how different big data analytics and machine learning techniques can be used to detect network anomalies in real-time to provide stronger security for the organizations and their network. The main objective for this project will be trying to use the latest technologies to identify different malicious activities like Distributed Denial of Service (DDoS) attacks, botnets and other network intrusions that can severely damage both the performance of the network and it's security.

I plan to use NetFlow[6], Which is a protocol system that can be used to collect and display information about network traffic as it flows in or out of an interface details like source and destination addresses, packet counts, and data volumes. This data can be essential for understanding both normal and abnormal network patterns. As it's hard to collect any abnormal network over personal router. We can find publicly available datasets like CTU-13, which is a collection of botnet traffic data [5] and some other datasets from Kaggle related to network anomalies. We can use these readily available dataset for intrusion detection analysis. We can use platforms like Azure HDInsight, Apache Hadoop, spark and Splunk to process and analyze large-scale, real-time traffic data[1]. These datasets will allow us to study network anomalies which covers a wide range of scenarios like common and rare attack patterns[2].

Next, I will try to use K-means clustering machine learning algorithm to group similar network behaviors and then try to detect deviations that can help identify potential attacks. The reason for using clustering technique in particular because this technique it doesn't rely on labeled data. This can help when dealing with real-time environment data[2]. Additionally, I plan to use Principal Component Analysis for the reduction of the dimensionality of the data to help improve the accuracy and interpretability of the results if it's required. By combining different unsupervised machine learning techniques with different analytical tools which can help to continuously monitor network activity and identify threats in real-time

By using the machine learning algorithm can help us in the detection of zero-day attacks new attacks which will have no predefined signature that is used by traditional security systems to detect the attack. The machine learning model can continuously analyze the data so that the system to adapt to changes in network traffic to reduce the false positives which can ensure that regular traffic is not mistakenly flagged as suspicious[1] and able to detect new attacks and help the organizations to respond quickly to those security attacks, which can help in reducing the impact of attack and prevent sensitive data loss.

The machine learning models can achieve up to 88% accuracy in detecting specific types of malicious behavior like UDP-based DDoS attacks[4], when applied to network data. The project will achieve more or similar levels of accuracy. From the information we get from this will not only help us strengthen cyber security measures but also provide the organization with the tools that are needed to proactively monitor and respond to emerging threats.

In conclusion, by combing the machine learning techniques along with big data platforms to push forward the field of network anomaly detection and develop a smart and more adaptive cybersecurity solutions. This could be used in different industries such as finance, healthcare, telecommunications, and government as everything is online in today's world. So, protecting sensitive data and ensuring network availability have become more and more important.

References:

[1]. Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19-31. https://doi.org/10.1016/j.jnca.2015.11.016

[2]. Ariyaluran Habeeb, R. A., Nasaruddin, F., Gani, A., Hashem, I. A. T., Ahmed, E., & Imran, M. (2019). Real-time big data processing for anomaly detection: A Survey. International Journal of Information Management, 45, 289-307. https://doi.org/10.1016/j.ijinfomgt.2018.08.006

[3]. Parwez, M. S., Rawat, D. B., & Garuba, M. (2017). Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network. IEEE Transactions on Industrial Informatics, 13(4), 2058-2065. https://doi.org/10.1109/TII.2017.2650206

[4]. Terzi, D. S., Terzi, R., & Sagiroglu, S. (2017). Big data analytics for network anomaly detection from NetFlow data. 2017 International Conference on Computer Science and Engineering (UBMK), 592-597. https://doi.org/10.1109/UBMK.2017.8093473

[5]. Stratosphere Research Laboratory. (2011). The CTU-13 dataset: A labeled dataset with botnet, normal, and background traffic. Retrieved from https://mcfp.weebly.com/the-ctu-13-dataset-a-labeled-dataset-with-botnet-normal-and-background-traffic.html

[6]. Yuzzi, R. (2021, March 10). The importance of symmetrical vs asymmetrical internet connections [Video]. YouTube. https://www.youtube.com/watch?v=lebIEzZcAKo