ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Big Data Architecture for Security Data and Its Application to Phishing Characterization

Himanshu Joshi

Department of CE, PVG's College of Engineering and Technology Pune, SavitribaiPhule Pune University Pune, India

ABSTRACT: As the internet grows, cyber security problems also arise with it. Different malicious activities are being carried out by the attackers so that they will be able to get the information of the victim. Using this information the attackers performs their illegal activities. Enterprises usually collect terabytes of security-relevant data, including network traffic, and software application events, among others. However, well established techniques, most of the time, are not scalable and typically produce many false positives when dealing with large amounts of data, degrading their efficacy. To face these emerging problems, big data analytics has attracted the interest of the security community. The use of big data frameworks for security solutions presents several benefits, such as the possibility of storing and using large quantities of security data. In this paper we design an architecture in which being built on the top of the Big Data frameworks that aims to mitigate the cyber security problem like phishing. It is being designed such that we are able to detect the phishing emails in a large data set and the information collected by the honeypot.

KEYWORDS-architecture, cybersecurity, spam, phishing, Hadoop, spark

I. INTRODUCTION

Security issues become more critical due to factors such as the large volumes and variety of data that may be vulnerable, the diversity of data sources and formats, and the velocity in which data are generated, typically following a stream nature with a high volume. Enterprises usually collect terabytes of security-relevant data, including network traffic, and software application events, among others. However, well established techniques, most of the time, are not scalable and typically produce many false positives when dealing with large amounts of data, degrading their efficacy. To face these emerging problems, big data analytics has attracted the interest of the security community. The use of big data frameworks for security solutions presents several benefits, such as the possibility of storing and using large quantities of security data. Although analyzing logs, network flows, and system events has been used for several decades in security solutions, conventional technologies are not adequate to be applied on such long term, large-scale volumes. In general, the traditional infrastructure keeps the data only for a limited period. Besides that, traditional techniques are inefficient when performing analytics and complex queries on large, unstructured datasets, while big data platforms perform these operations efficiently. In this paper we present architecture for cyber security applications based on big data frameworks. Our architecture has the capability of collecting data from different sources, storing, combining, and processing them effectively. For example, sources like pcap files and other logs from a honeypot, data streams collected from black list sites can all be stored in our system.

Motivation:

To mitigate cyber security problems such as spam and phishing and we show how it is being used to study spam and phishing collected using a global honeynet.

II. LITERATURE SURVEY

1. "Considerations for Privacy Preserved Open Big Data Analytics Platform", Surendra .R, Dr. Mohan .R.S
In this paper, we have tried to providelist of different major aspects to be considered for building apractical open big data analytic platform. Different data accesstechniques, inference control methods and other importantconsiderations to be made are discussed.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

2. **"A hadoop extension to process mail folders and its application to a spam dataset", Pedro H. B. Las-Casas, Vinicius Santos Dias, Renato Ferreira**

In this paper we present a Hadoop extension used toprocess and analyze large sets of e-mail, organized in mailboxes.To evaluate it, we used gigabytes of real spam traffic data collectedaround the world and we showed that our approach is efficientto process large amounts of mail data.

3. **"MATATABI: Multi-layer Threat Analysis Platform with Hadoop", Hajime Tazaki, Kazuya Okada**

In this paper, we report our experience from operatingthe Hadoop platform, called MATATABI, for threat detections,and present the micro-benchmarks with four different backendsof data processing in typical use cases such as log data and packettrace analysis. The benchmarks demonstrate the advantages of
distributed computation in terms of performance. Our extensiveuse cases of analysis modules showcase the potential benefit ofdeploying our threat analysis platform.Index Terms—Cyber security, Multi-

4. **"LARX: Large-scale Anti-phishing by Retrospective Data-Exploring Based on a Cloud Computing Platform", Tianyang Li, Fuye Han, Shuai Din and Zhen Che**

In this paper, we propose an offline phishing detectionsystem named LARX (acronym for Large-scale Anti-phishing byRetrospective data-eXploration). LARX uses network traffic dataarchived at a vantage point and analyzes these data for phishingattacks. All of LARX's phishing filtering operations use cloud
computing platforms and work in parallel. As an offline solutionfor phishing attack detection, LARX can be effectively scaled upto analyze a large volume of trace data when enough computingpower and storage capacity are provided.

5. **"SpamCloud: A MapReduce based Anti-Spam Architecture", Godwin Caruana, Maozhen Li and Hao Qi**

From a computing perspective and context, Spam canbe described as an Internet scale problem. A potential approachfor tackling Spam is consequently via the application of Internetscale algorithms and techniques. A number of approaches existwhich are pitched to tackle this type of challenge in such fashion,including MapReduce. This research evaluates the degree of
feasibility and applicability of Hadoop's Map Reduce Frameworkwhen applied to spam filtering.

6. **"PhishStorm: Detecting Phishing With Streaming Analytics", Samuel Marchal, Jérôme François, Radu State, and Thomas Engel**

in this paper efficient implementationpatterns that allow real-time analytics using Big Dataarchitectures such as STORM and advanced data structures basedon the Bloom filter. In thispaper, we introduce PhishStorm, an automated phishing detectionsystem that can analyze in real time any URL in order to identifypotential phishing sites. PhishStorm can interface with any emailserver or HTTP proxy. We argue that phishing URLs usuallyhave few relationships between the part of the URL that mustbe registered (low-level domain) and the remaining part of theURL (upper-level domain, path, query). We show in this paperthat experimental evidence supports this observation and can beused to detect phishing sites.

III. PROPOSED SYSTEM APPROACH

Our architecture has the capability of collecting data from different sources, storing, combining, and processing them effectively. For example, sources like pcap files and other logs from a honeynet, data streams collected from black list sites and security-related search streams from social networks like Twitter, can all be stored in our system. Different algorithms, possibly implemented in different programming environments, can then be used to process combine data from different sources as needed. The architecture is depicted in Figure 1. It is composed by five parts: (i) data collection, (ii) storage, (iii) reader, (iv) processing and (v) visualization.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Proposed system architecture

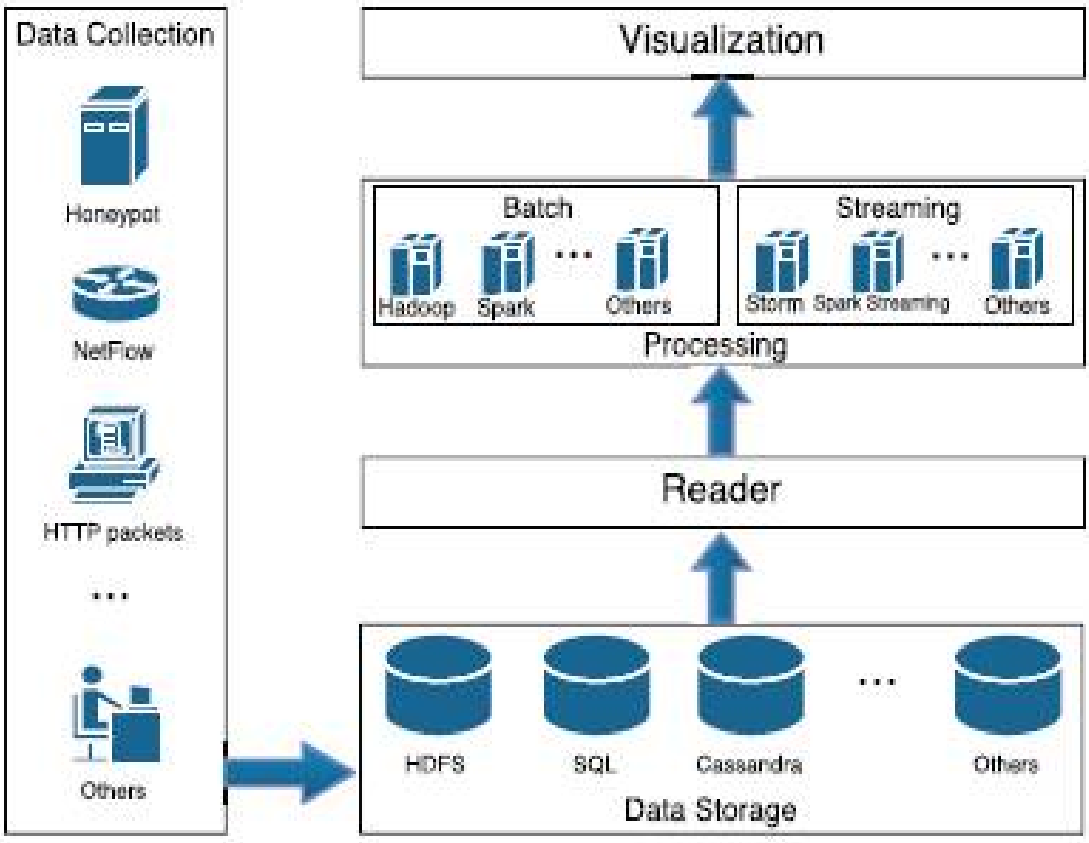

Figure 1. Architecture of our suggested solution.

Proposed system advantages

1. To implement a spam analysis application to identify phishing campaigns.
2. The possibility of storing and using large quantities of security data.
3. The spam emails are detected successfully.
4. Provides security from harmful emails.
5. To study spam is the direct interaction with the agent responsible for the abuse, making it possible to capture important information about it.

IV. MODULES

A. DATA COLLECTION

The first step of our architecture is the collection of data.The most important aspect of this step is to identify relevantdata sources. Due to the constant advances of techniquesused by attackers, it is increasingly necessary to use distinctsources of data in order to mitigate cybersecurity threats.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

B. *STORAGE*

Considering the heterogeneity of data that can be usedby an application, it is important to have a storage strategythat will maximize performance and facilitate the use of thedatasets. Every data source has different characteristics andis produced in a different volume and velocity, thus this isimportant to be considered in order to store them in thebest possible way.

C. *READER*

Partitioning the blocks into logical records is performedby the Input Format associated with a file and the RecordReader for it. The Input Format tasks are (i) validate theinput; (ii) split the input blocks and files into logical chunks;and (iii) create a Record Reader implementation to be usedto create key/value. A Record Reader uses the data within theboundaries created by the input split to generate key/valuepairs that will be used by the mappers.D. ProcessingOnce we have a scalable environment to store all types ofdata, from different ources, that may be of interest, we needa tool to process and analyze terabytes in a scalable manner,so we can achieve high performance.

E. VISUALIZING RESULTS

Finally, after executing all operations, the application mustbe able to present it in a human-friendly form. Different visualization platforms may be added to the architectureeasily using HDFS as a data repository, or by integratingthem directly into the applications.

V. CONCLUSION

In this paper we introduced an architecture that enables the implementation of Big Data applications to be used in the context of cybersecurity. As a case study, we developedan application that aims to process spam traffic using HDFS, Hadoop, Spark, and data collected from honey pots spread in different locations of the world. We were able to demonstrate the power of our application, from the implementation of simple operations to be used in the analysis of spam traffic, to more complex operations such as phishing detection.

The proliferation of data sources and data collecting structures has lead to a large increase in the data available for cyber security experts. To process such large volumes of data, scalable massive data processing solutions are needed. As mentioned in literature survey, the present work on the project uses the LSH algorithm which detects the spam emails but it gives accuracy of 98.1%. The system complexity is also high. Our system will reduce the complexity along with that the accuracy increases to 99.5% and the spam emails will be detected successfully in less time.

REFERENCES

[1]Kotsiantis, S.B., Zaharakis, I.D., Pintelas, P.E.: Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 2006

[2] Y. Yu, Y. Mu, and G. Ateniese, “Recent advances in security and privacy in big data,” j-jucs, Mar 2015.

[3] P. H. B. Las-Casas, V. Santos Dias, R. Ferreira, W. Meira, and D. Guedes, “A hadoop extension to process mail folders and its application to a spam dataset,” in International Symposiumon Computer Architecture and High Performance ComputingWorkshop (SBAC-PADW), Oct 2014, pp. 108–113.

[4] P. H. B. Las-Casas, V. Santos Dias, R. Ferreira, W. Meira, andD. Guedes, “A hadoop extension to process mail folders andits application to a spam dataset,” in International Symposiumon Computer Architecture and High Performance ComputingWorkshop (SBAC-PADW), Oct 2014, pp. 108–113.

[5] T. White, Hadoop: The Definitive Guide. O’Reilly Media,Inc., 2012.

[10] G. Caruana, M. Li, and H. Qi, “Spamcloud: A mapreducebased anti-spam architecture,” in Int’l Conference on FuzzySystems and Knowledge Discovery, 2010.

[11] S. Marchal, J. Francois, R. State, and T. Engel, “Phishstorm:Detecting phishing with streaming analytics,” in Network andService Management, IEEE Transactions on, Dec 2014.

[12] T. Li, F. Han, S. Ding, and Z. Chen, “Larx: Large-scale antiphishing
by retrospective data-exploring based on a cloudcomputing platform,” in ICCCN 2011, July 2011.

[13] H. Tazaki, K. Okada, Y. Sekiya, and Y. Kadobayashi,“Matatabi: Multi-layer threat analysis platform with hadoop,”in BADGERS, 2014.