International Journal of Research in Engineering, Science and Management
Volume-3, Issue-3, March-2020
www.ijresm.com | ISSN (Online): 2581-5792

Spamdoop - A Privacy Preserving Big Data Platform for Collaborative Spam Detection

V. Vidya[1], C. K. Archana[2], M. Bindhu Shree[3], N. Meenakshi[4], R. Tanuja[5]
[1]*Assistant Professor, Department of Computer Science and Engineering, Nagarjuna College of Engineering and Technology, Bangalore, India*
[2,3,4,5]*Student, Department of Computer Science and Engineering, Nagarjuna College of Engineering and Technology, Bangalore, India*

***Abstract*: Spam has become the stage of decision utilized by digital law breakers to spread malignant payloads, for example, infection and Trojans. Spam is the platform used by cyber criminals to get information about the particular person and use that information for some malicious activities hiding his/her identity. Spam is the unwanted or fake message sent to e-mails and phone by the cyber criminals in order to gain information about the people like account number, credit card number, OTP, pin etc. After getting the information about the person the cybercriminal will use this information for their favor and that may lead the person life in to threat and may cause huge financial loss for the person. In order to prevent this, this project is useful where these kind of unwanted and fake messages sent to gain information about the user are identified and deleted without the user information within a 1-2 mins. By doing this the user will not able to see the message and give the information about the person to the criminals.**

***Keywords*: Big data, Spam detection.**

1. Introduction

The massive unsolicited emails sent is normally said to be spam. It is quite not possible to describe the word spam more accurately. A spam is one of the difficult challenges of the linked generation because of its impact and affects. Because of the effects of spam massive amount of research are opposing it. As the effort of many researches there is a gradual decline in the spamming activities. Now spam is no longer a threat but may cause damage financially. According to Kaspersky's spam and phishing statistics, about 66 percentage of e-mails that were sent in the first quarter of year 2014 worldwide were considered as the spam emails. As the effort of many researches and implementation of many project the Spam emails were reduced by about 77 percent of before by the span of 2 years. By implementing this project, the spam emails and messages will reduce by greater percentage. E-mails these days have become a popular and most favored means of communication. In case of emails get hacked by the criminals and if he gains the precious information about the person or the company which he is working then life will be under threat and loss for the company as well. Because of these reasons it would be better if this project would to implemented. The extent of email received and amount of spam is constantly growing. Spam mails are defined as electronic message posted to thousands of recipients usually for the purpose of advertisement or profit. Some of the spam emails modify as a phishing emails seeking user's confidential data and accessing their bank accounts for financial frauds. Phishing is a fraudulent attempt aimed at capturing user information such as username and passwords, account number, credit /debit card number, pin, OTP etc., by impersonating a trustworthy entity in an electronic communication. Apart from examining the attachments, the content alone may also be prone to word de-obfuscation technique to fool the spam filters.

2. Literature survey

Literature survey plays an initial and a major role in pursuing a project, publishing a paper or any research to be done in any field of interested topic. The major aim of literature survey is to gather information and stuff on the research topic to see what work has been done in the previous years by other scholars and to collect the future

[1] M. V. Gayoso, A. F. Hernandez,´ and E. L. Hernandez´. State of the art in similarity preserving hashing functions. In Proceedings of the International Conference on Security and Management (SAM), The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (World Comp), 2016.the best-known functions of this type are the context-triggered piecewise hashing functions, which create a signature formed by several hashes of the initial file. In this contribution, we present the state of the art of the most important similarity preserving hashing functions, analyzing their main features.

[2] J. Chen, R. Fontugne, A. Kato, and K. Fukuda. Clustering spam campaigns with fuzzy hashing. In Proceedings of the Asian Internet Engineering Conference (AINTEC), page 66. ACM, 2015.In this paper we propose a new method based on fuzzy hashing to cluster spam with common goals into the same spam campaign. Fuzzy hashing allows us to identify emails with similar content though usual identifiers are obfuscated. Using the proposed method, we process a three-year long dataset that consists of 540 thousand spam emails.

3. Proposed system

The spam message which were trying to get information from the people through fake and impressive message like telling that you have won a lottery of so and so amount and asking for account details. In this case if the person gives the information then the criminals will use that for their favors. In this case this project is very useful where it will delete the fake message without displaying the content to the user which helps in preventing information leak to the criminals. Now at present in gmail the spam messages are deleted in 15 days. During this period there is a chance of leaking information because of this the criminals can hack our account and it may cause financial loss for the person and the organization that the person is working for. To avoid this, we are trying to delete the message in 2-3 mins, without disclose the content of the message. Therefore, there is no chance of giving the information and our account will be safe form cybercriminals. If our account is hacked sometimes it may also put our life into threat so this project will be very useful us.

4. Methodology

In this section the fundamental methodologies are highlighted The Spamdoop Platform, Spamdoop is a stage enabling different elements to team up in early recognition of mass spam battles. Our stage additionally fulfils the security necessities of members. A review of Spamdoop is depicted in fig. 1 in the following segment which features the following key segments of the framework:

- *The Obfuscator:* The Obfuscator is mainly employed for the purpose of encoding the contents of the e-mails which will allow the parallel processing of the spam without the cost of revealing the contents of emails.
- *The Parallel Classifier:* The Parallel Classifier is used for the classification of the emails by utilizing the properties of encoding. This ensures routing messages similar to each other in the same buckets.
- *The Anomaly Detector:* The Anomaly Detector detects whether a certain email corresponds to a spam or not. The detection process depends on analyzing the size of the buckets along-with their rate of growth.

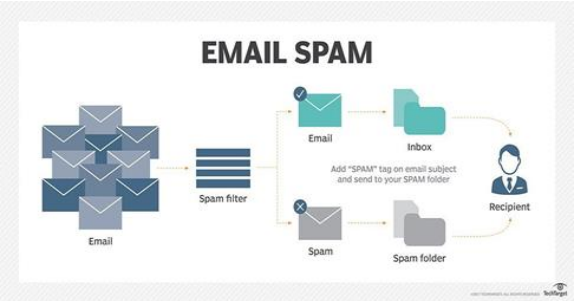

Fig. 1. Spam detection process

A. Map Reduce Technique

Applications frequently require more resources than what are available on a conventional inexpensive machine. Many organizations find themselves with business processes that no longer fit on a single cost effective computer. A simple but expensive solution is to buy specialty machines having a lot of memory and high CPU Power. This solution scales as far as what is supported by the fastest machines available, but usually the only limiting factor is the budget. An alternate solution is building a high availability cluster. Such a cluster typically endeavors to look like a single machine, and usually requires very specialized installation and administration services. Most of the high- availability clusters are proprietary and high priced. Thus an economical solution for acquiring the necessary computational resources is cloud computing. A common pattern is to have bulk data that is to be transformed by processing each data item which are essentially independent of one another; that is, using a single-instruction multiple-data (SIMD) algorithm. Hadoop provides an open source framework for cloud computing, as well as a distributed file system. Hadoop supports the Map Reduce model, introduced by Google as a method of solving a class of pet scale problems with large clusters of inexpensive machines. The model is based on two distinct steps for an application:

- *Map:* An initial ingestion and transformation step, in which individual input records can be processed in parallel.
- *Reduce:* an aggregation or summarization step, in which all associated records, must be processed together by a single entity. The core concept of Map Reduce in Hadoop is that the input may be split into logical chunks, and each chunk is initially processed independently, by a map task.

The results of these individual processing chunks can be physically divided into distinct sets and then sorted. Each sorted chunk is passed to a reduce task.

A map task may run on any compute node in the cluster, and multiple map tasks may be running in parallel across the cluster. The map task is accountable for the transforming of the input records into key/value pairs. The output of all the maps is partitioned, and each partition is sorted. There'll be a partition for each of the reduced tasks. Each partition's sorted keys and the values associated with the keys is then processed by the reduce task. There may be multiple reduce tasks running in parallel on the cluster. The application developer needs to provide the following four items to the Hadoop framework: the class that will read the input records and transform them into one key/value pair per record, a map method, a reduce method, and a class that will transform the key/value pairs that the reduce method outputs into output records. The Hadoop Map Reduce framework needs a shared file system. This shared file system isn't required to be a system-level file system, as long as there is a distributed file system plug-in available to the framework. When HDFS is used as the shared file system, Hadoop is able to take advantage of knowledge about which node hosts a physical copy of input data, and will attempt to schedule the task that is to read that data, to run on that machine.

The Hadoop Distributed File System (HDFS) Map Reduce environment provides the user with a sophisticated framework to manage the execution of map and reduce tasks across a cluster of machines.

5. Conclusion

The collaborative spam detection platform begins referred to in this paper offers multiple benefits in the terms of safeguarding the privacy of all the stakeholders involved and the amount of data begin used. The encoding technique employed is effectively scalable on MapReduce platforms outdoing various distance-preserving hashing techniques. The technique used for bucketing simplified the process of offering easy classification and grouping of objects along with the anomaly detection based on histogram efficiently distinguished spam from ham.

Recent tests conducted have shown that the grouping time of digests is reduced by 53% when the work is distributed across four nodes. The computation time is decreased by 57% after using CRC32 on a single node and by 46% on for nodes. Also the processing was of six batches across four nodes was 52% faster. The above mentioned experimental results clearly shows that the Spam detection technique using Big Data are the need of the hour and should be employed to deal with the multitudes of problem related to spams.

References

[1] Kaspersky Lab, “Spam and Phishing Statistics for 2016”. https://www.kaspersky.com/about/press-releases/2016kaspersky-lab-reports- significant-increase-in-malicious-spam-emails-in-q1-2016.
[2] AlMahmoud, Abdelrahman, et al. "Spamdoop: A privacy-preserving Big Data platform for collaborative spam detection," IEEE Transactions on Big Data, 2017.
[3] Chen, Long, and Guoyin Wang. "An efficient piecewise hashing method for computer forensics." Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on. IEEE 2008.
[4] Wang, Min, et al. "A General Framework for Linear Distance Preserving Hashing." IEEE Transactions on Image Processing, vol. 27, no. 2, pp. 907-922, 2018.
[5] J. Francois, S. Wang, W. Bronzi, R. State, and T. Engel. “Bot-cloud: Detecting botnets using mapreduce,” in IEEE International Workshop on Information Forensics and Security (WIFS).
[6] F. Breitinger and H. Baier, “Similarity preserving hashing: Eligible properties and a new algorithm mrsh-v2,” in International Conference on Digital Forensics and Cyber Crime (ICDF2C), pages 167–182. Springer, 2012.
[7] Z. Zhong, L. Ramaswamy, and K. Li. Alpacas, “A large-scale privacy-aware collaborative anti-spam system,” in the 27th Conference on Computer Communications (INFOCOM). IEEE 2008.
[8] C. Tseng, P. Sung, and M. Chen, “Cosdes: A collaborative spam detection system with a novel e-mail abstraction scheme,” IEEE Transactions on Knowledge and Data Engineering, 2011.
[9] Apache Mahout: Scalable machine learning and data mining. https://mahout.apache.org/.
[10] Apache Hadoop. “Welcome to Apache Hadoop”. http://hadoop.apache.org/.
[11] Nagiwale, Amin Nazir, and Manish R. Umale. "Design of self- adjusting algorithm for data-intensive MapReduce applications." Energy Systems and Applications, 2015 International Conference on. IEEE 2015.
[12] Zhong, Zhenyu, Lakshmish Ramaswamy, and Kang Li. "ALPACAS: A large-scale privacy-aware collaborative anti-spam system," INFOCOM 2008. The 27th Conference on Computer Communications. IEEE 2008.