ATLANTIS PRESS
Advances in Computer Science Research (ACRS), volume 54
International Conference on Computer Networks and Communication Technology (CNCT2016)

Hadoop-based Intrusion Detection Technology and Data Visualization for Website Security

Xiao-ming ZHANG*, Yu-xin WANG,Ge-tong ZHANG and Guang WANG

Department of Computer,Beijing Institute of Petrochemical Technology,Beijing,China

*Corresponding author

Keywords: Website, Intrusion detection, Hadoop, Visualization, MapReduce

Abstract. Network attack becomes more popular today. It is still difficult to detect the intrusion activity accurately for users websites through traditional approaches. A kind of integrated system is designed and established for the website security analysis based on Hadoop system. The core components of hardware subsystem is intrusion detection system (IDS) and Hadoop cluster. The data can be interacted between the intrusion detection system (IDS) controller, detection engine, website server, transferring server and Hadoop cluster. Based on the Map Reduce model, the Word Count algorithm is revised to obtain data statistical results such as IP address, intrusion level, intrusion type and intrusion time. These analyzed results are saved automatically into the MySQL database to form data tables, including statistical analysis, level distribution, week analysis and danger efficiency. Then, the resulted data are presented with visualization effect for the website managers.

Introduction

Network attack activities on the website become more popular and complicated. From the web browsing effects, some web pages may be changed with wrong information. Meanwhile, denial of service attack takes place to shut down the website server or network, making it inaccessible to its intended users. The manager need know the website operating situation online. However, most of the organizations lack of experience in preventing from such attacks. The traditional approaches with intrusion detection system (IDS) are always fail when facing to actual attacks.

With the network attacking technology, the higher prevention caution is required to avoid the single system detection problem. One of the advanced technologies is big-data technology. It based on cloud computing platform and data mining technology. For example, With Log analysis, a kind of Hadoop-based Intrusion Detection System is designed as the simulation data as Map-Reduce algorithm [1] . For the computational Algorithms, the traditional Apriori association algorithm was applied in Hadoop cloud platform for the intrusion detection system [2]. The results show that the optimized Hadoop cloud IDS could be better in detection effects. Similarly, the Hadoop system was presented for fussy association analysis [3]. After comparison with traditional Apriori algorithm, the results demonstrated the advantage of cloud computing. Because all these simulation data are from data-set in KDD-CUP99 [2-6], it lacks truly validation in network operation process.

Another test data-set is LLS DDOS 1.0. It was applied for the large-scale network security situation analysis to get the multidimensional association rule mining [7]. Besides, one kind of experimental system for intrusion detection was designed [8]. It was used to detect two types of intrusion of network sniffing and port scanning. It can

only for experiment teaching. In [9], a type of log detection system was established for website based on cloud platform. It put the detection results together with the alerting information to raise the accuracy for attack alerting.

Hadoop was an idea developed from Google's MapReduce, a software framework which breaks down an application into numerous small parts. These parts called fragment can run on any node in the cluster. The current Apache Hadoop ecosystem consists of the Hadoop Commonl, Hadoop Distributed File System (HDFS), MapReduce, YARN, and a number of related projects such as Apache Hive and HBase.

In order to solve the problems of large scale intrusion detection and on-line data analysis, a kind of website intrusion detection system is designed based on Hadoop system [10]. The core components of hardware subsystem is intrusion detection system (IDS) and Hadoop cluster. It aims to establish a kind of private cloud platform for users to monitor the website running situation. In this paper, the system is emphasized on data flow design and software development. Finally, the data are presented with visualization effects through web pages for users.

Proposed System

The monitoring system aims to truly website operation security in our university. In order to prevent from complicated network attach, the big-data technology based on cloud platform is required for deep analysis.

System Network Structure

The system is integrated with hardware subsystem and software subsystem. The former is mainly composed of website server, IDS device and Hadoop cluster. And, the software subsystem includes multilevel software components, HDFS, MySQL database system and visualization programs. The system network structure is design as shown in Figure 1.

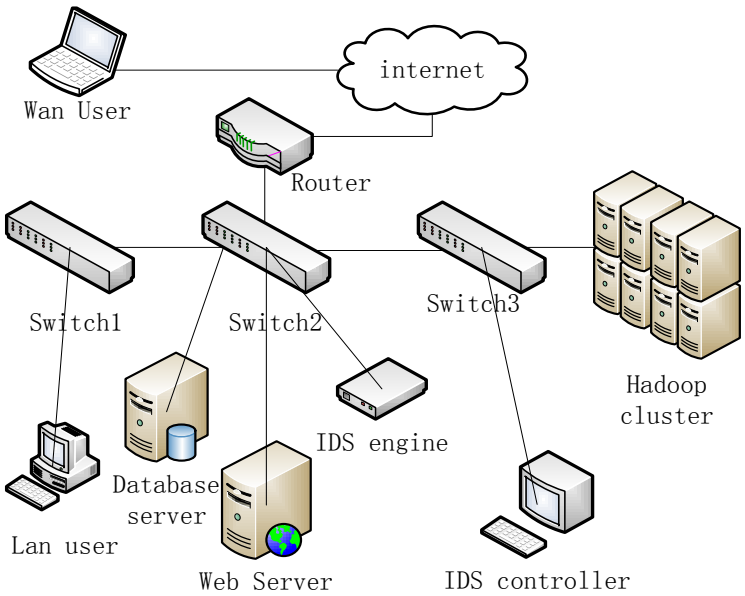

Figure 1. Network structure of Website monitoring system

The Hadoop cluster is designed for large-scale data processing and analysis. It consists of four servers which are configured with CentOS 6.5 and Hadoop 2.6 system. Three of them are served as computing nodes. All these servers are also equipped with

Mahout Algorithm libraries [11]. Besides, another server is adopted as data storage with MySQL database management system.

System Functions

The system is composed of four modules of source data collection, distributed computing, user interface and result storage.

There are two kinds of data source, includes log files from the website server directly and data captured by the IDS device. Then, these files will be transmitted through network to the MySQL database system. After data abstraction, cleaning and conversion, the data is formatted as text file. Furthermore, the text file is input into the Hadoop system for data mining.

In the system, the user can upload the reference parameters together with the source data into the HDFS. The file system provides reliable data storage and access across all the nodes in a Hadoop cluster. It links together the file systems on many local nodes to create a single file system.

After finishing cluster computing, the resulted data are transformed and stored in the relational management system. Next, the data can be accessed with visual effects from web browser.

Software Design

The software subsystem is designed as three layers of data storage, business computing and application service, as shown in Figure 2. There are two kinds systems for data storage. One is HDFS, which is responsible for data-set input into the Hadoop cluster. The other is the relational database management system of MySQL, which is used to save the data analysis results for users browsing remotely.

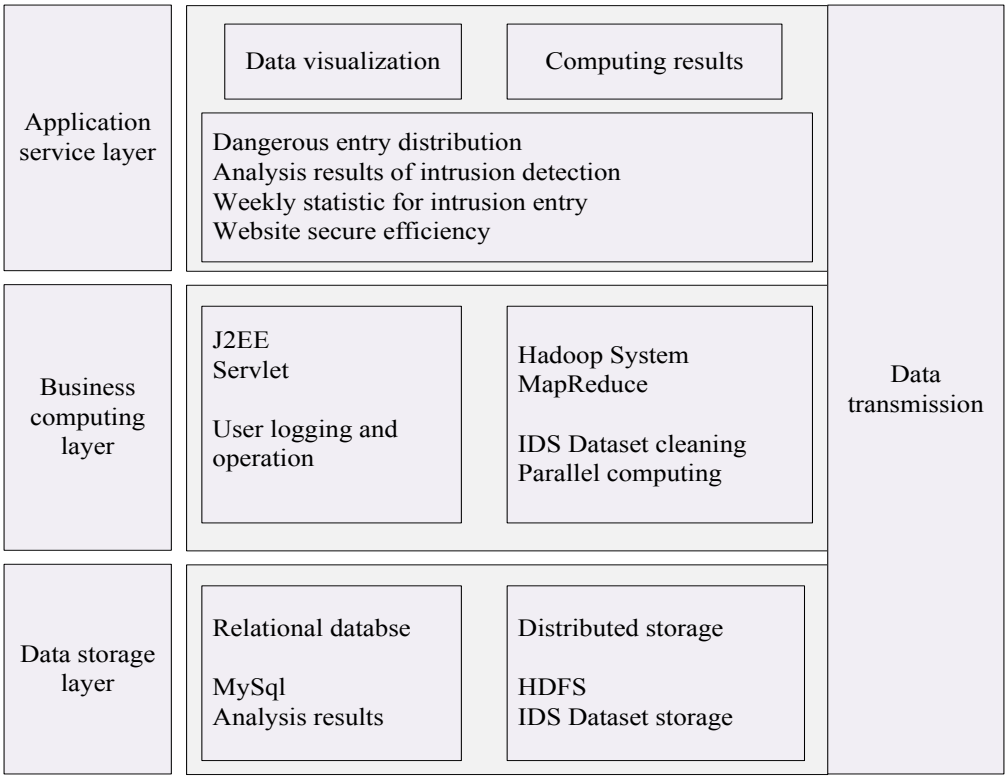

Figure 2. Design of software system structure

Core Module Design

Here, several software modules are stated as follows.

Data Acquisition

Owing the IDS controller is equipped in the local are network with necessary security feature, the FTP capturing mode is adopted for data acquisition. The pointer of FTP site is the position of IDS data from the IDS engine. Then, the response IDS logs will be copied to the exchanged server as initial database. The data in the IDS controller is collected remotely by Java programming with FTP type and submitted into the HDFS.

Data Cleaning

Data cleaning of IDS logs aims to extract useful key content from the logs records. Through the designed Map-Reduce program, many items such as entry IP address, entry time, entry type and level can be obtained.

From the view of parallel computing, the Map function is adopted to get these data. Each data are processed line by line as several data cleaning operations such as filling, deleting and formatting.

Data Analysis

Map Reduce is a framework for writing applications that process large amount of structured and unstructured data in parallel across a cluster of thousands of machines, in a reliable, fault-tolerant manner.

For the website cleared data, the same algorithm can be translated to MapReduce algorithms for running them on the Hadoop clusters by translating their data analytic logic to the MapReduce job. Figure 3 shows the data flow from the results of data cleaning to the database storage.

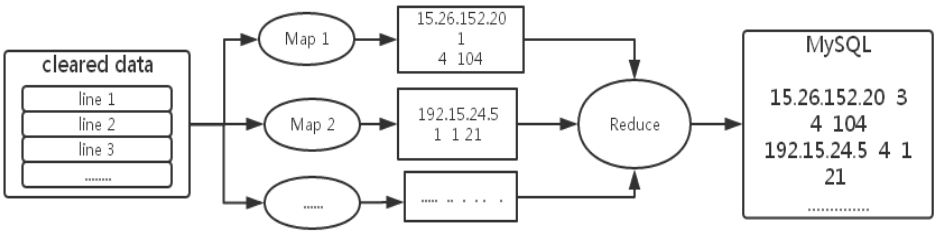

Figure 3. Map-reduce algorithm for data analysis

Data Presentation

The final stage of the process consists of visualization of the results of data analysis through the monitoring website. Visualization is an interactive way to represent the data insights. This can be done with various data visualization software's like Echart etc. By means of Echart components from the Baidu Company, the data are formatted as Jason from the database to the front end of website.

Experimental Analysis

The system is implemented under Hadoop-based IDS test environment, as shown in Figure 4. In the Figure 4, it shows the relation between entry date, IP address, entry frequency, the intrusion type and the danger level.

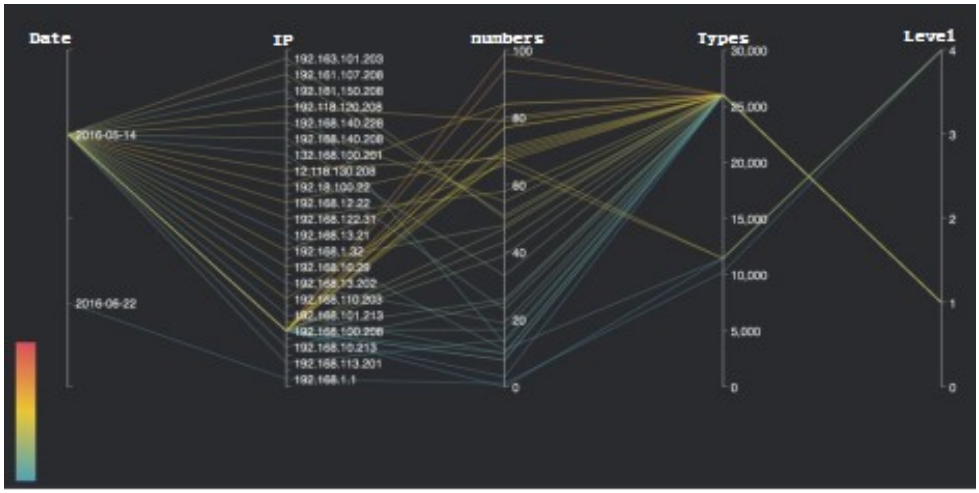

Figure 4. Relationship among intrusion parameters

For the intrusion time as each hour, the analysis result is shown in Figure 5.

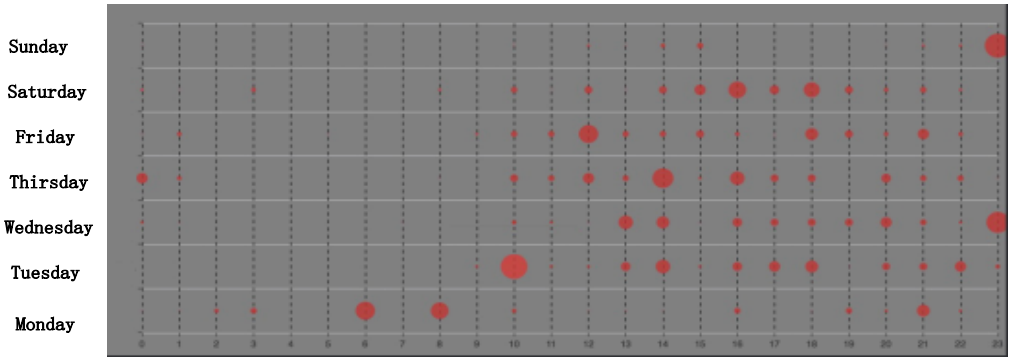

Figure 5. Distribution of intrusion time in one week

It can be seen that the intrusion danger becomes more serious at midnight on Wednesday and Saturday.

Conclusions

A kind of cloud platform for the website intrusion detection is established as large-scale data collecting and analysis system. The IDS information can be transformed automatically to the cloud server cluster. Based on the Hadoop system and Map-Reduce algorithm, the intrusion information from the website logs by IDS system can be processed automatically through data acquisition, data cleaning and data analysis. Experimental results under the analysis system show that the analysis results can be presented for the end-users with friendly interface and reliable effects.

Acknowledgment

The work is financially supported by Beijing Institute of Petrochemical Technology with Projects of BIPT-POPME-2015.

References

[1] Manish Kumar, M. Hanumanthappa. Scalable Intrusion Detection Systems Log Analysis using Cloud Computing Infrastructure. 2013 Proc.of Int. Conference on Computational Intelligence and Computing Research.

[2] Chen Zhen. Optimal Design of Intrusion Detection System in Hadoop Cloud Platform. Journal of Xi'an Technological University. 2012,32(9):716-722

[3] WEI De-zhi, WU Xu, LIN Li-na, WANG Qi-guang. Fuzzy Mining Algorithm for Rules Based on Cloud Computing in Intrusion Detection. Journal of Jilin Normal University (Natural Science Edition). 2012, (1):115-118

[4] Mohammed Nazim Feroz,Susan Mengel.Examination of Data, Rule Generation and Detection of Phishing URLs using Online Logistic Regression. 2014 IEEE International Conference on Big Data.

[5] Jakrarin Therdphapiyanak,Krerk Piromsopa. An analysis of suitable parameters for efficiently applying K-means clustering to large TCPdump data set using Hadoop framework.2013 10th Int.Conf. On Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology(ECTI-CON)

[6] ACM KDD CUP [Online]. http://www.sigkdd.org/ kdd-cup-1999-computer-network-intrusion-detection.

[7] Sun Yi-jun, Zhang Hong-li, He Hui. Mining Alert Association Rules in Large scale Network Security Situation Analysis[C]. 2007 National Network and Information Security Technology Conference.

[8] SUN Guo-zi, YU Chao, CHEN Dan-wei. An Implementation of Intrusion Detection Experiment System.Computer Education, 2010,(6):154-157

[9] Liu Jianjun,Hu Ying ,Dai Fangfang. The WEB Intrusion Detection Based on Hadoop[C]. 2014 19th National younger communication Conference.

[10] Zhang Xiao-ming, Hadoop-based System Design for Website Intrusion Detection and Analysis. 2015 IEEE Int. Conf. On Smart City/SocialCom/SustainCom (SmartCity).

[11]Overview - Apache mahout - Apache software foundation [Online]. https://cwiki.apache.org/confluence/display/MAHOUT/Overview.