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. 3, Issue 5, May 2015

File Inclusion Vulnerability Analysis using Hadoop and Navie Bayes Classifier

[1] Vidya Muraleedharan [2] Dr.KSatheesh Kumar [3] Ashok Babu

[1]M.Tech Student, School of Computer Sciences, Mahatma Gandhi University, Kerala, India

[2] Assistant Professor, Department of Futures Studies, University of Kerala, Kerala, India

[3]Assistant Professor, School of Computer Sciences, Mahatma Gandhi University, Kerala, India

ABSTRACT- Big data is an evolving data set that describes any voluminous amount of structured, semi-structured and unstructured data and is beyond the ability of a traditional database tool .Big data can be analysed for extracting valuable information. Hadoop rides the big data where the massive quantity of information is processed using clusters of commodity hardware. A web server log is automatically created by a server which maintains a history of vulnerability attacks. SQL Injection and Remote File Inclusion are the two most frequently used vulnerability attacks and hackers preferring easier rather than complicated attack techniques .RFI uses the weakness of PHP language which in today's world is the most widely used. Hadoop Technologies like Oozie, Hive,Pig and Sqoop can be used to analyze web log data to detect File inclusion vulnerabilities . Naïve Bayes algorithm is implemented in Hive user defined function to classify attack keywords in log file. Oozie is a job coordinator and work flow manager that supports several types of Hadoop jobs such as Java map-reduce, Streaming map- reduce, Pig, Hive, and Sqoop

KEYWORDS: Hadoop,Naïve Bayes Classifier,Pig,Hive,Oozie,Sqoop.

I. INTRODUCTION

A web log data is a collection of facts from the grids of web servers usually of unstructured forms of the digital universe. A large amount of the data available on the internet is generated either by individuals, groups or by the organization over a particular period of time. The volume of data becomes larger day by day as the usage of the web makes an inevitable part of human activities. The rise of these data leads to a new technology such as hadoop[1] that acts as a framework to process, manipulate and manage very large data sets along with the storage .web log data is a required high volume, high velocity and high variety information assets that demand cost-effective , innovative forms of information processing for enhanced insight and decision making. Log files contain information about user name, ip address, time stamp, access request, number of bytes transferred, result status, a url that referred and user agent. The log files are maintained by the web servers. By nalyzing these log files give a neat idea about the user. A statistical analysis of the server log may be used to examine file inclusion vulnerabilities.

These files are usually not accessible to general internet users, only to the web master or other administrative person.

To hackers, rfi/lfi [2] attacks are very attractive since they target php applications. With more than 77 percent of today's websites running php, rfi should be on every security practitioner's radar. Big data analytics applies advanced analytical techniques of large data sets to discover hidden patterns such as file inclusion attacks and other useful information. The growing number of technologies is used to aggregate, manipulate, manage and analyse big data. Some of the most prominent technologies in hadoop are hive, pig,sqoop,mongodb ,hbase,oozie etc. the proposed work analyses web log of an organization, to identify file inclusion vulnerabilities using hadoop technologies like oozie pig,hive and sqoop.oozie will coordinate and schedule the jobs of pig and sqoop.

II.HADOOP MAPREDUCE

mapreduce[3] a java based programming model consists of two phases: a "map" phase, followed by an aggregating "reduce" phase.. a map function processes a key/value pair (k1,v1,k2,v2) to generate a set of intermediate key/value pairs, and a reduce function merges all intermediate values [v2] associated with the same intermediate key (k2) as in (1). a map reduce job

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. 3, Issue 5, May 2015

usually splits the input data set into independent chunks which are processed by the map tasks. the framework sorts the output of the map, which are then input to the reduce tasks. both the input and the output of the processed job are stored in a file system. typically just zero or one output value is produced by the reducer. in map reduce, a mapper and reducer are identified by the following signature,

map (k1, v1) --->[(k2, v2)]reduce (k2, [v2]) ---->[(k3, v3)] (1)

the data stored in a file system namespace contributes to hdfs which allows master-slave architecture. the cluster consists of a single name node that manages the file system namespace. there can be a number of data nodes usually one per node in the cluster which periodically report to name node, the list of blocks it stores .it automatically re-replicates the data blocks and place them on multiple nodes for processing. using hdfs a file can be created, deleted, copied, but cannot be updated. the file system uses tcp/ip for communication between the clusters.

A) PIG

pig is a scripting language for analyzing large data sets. the development cycle of map reduce is very long. Writing the map phase and reduce phase, compiling and packaging the code, submitting the job(s), and retrieving the results is a time consuming. Pig's advantage is its ability to process Tera bytes of data simply by issuing a half-dozen lines of Pig Latin . It was created at Yahoo! to make it easier for researchers and engineers to mine the huge datasets there. Pig provides extensive support for user defined functions as a way to

specify easy processing.

B) HIVE

The Apache Hive data warehouse software facilitates querying and managing large data sets residing in distributed storage. Hive provides a mechanism to query the data using a SQL-like language called HiveQL. At the same time this language also allows programmers to express their functions in Java. Algorithm can be implemented through Hive user defined functions.

C) OOZIE

An Oozie work flow is a collection of actions and Hadoop jobs arranged in a Directed Acyclic Graph (DAG).Oozie can manage different type of hadoop jobs ,can control work flow jobs based on time and data triggers, can manage batch coordinator applications. Oozie has been designed to scale, and it can manage the timely execution of thousands of work flow in a Hadoop cluster, each composed of possibly different jobs such as Pig,Hive and Sqoop.

D) SQOOP

Sqoop is a Hadoop tool designed to transfer data between Hadoop and relational databases. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from the Hadoop file system to relational databases.

III. FILE INCLUSION ATTACK DETECTION

Hadoop framework access a large semi structure data in a parallel computation model. Log files usually generated from the web server comprise of large volumes of data that cannot be handled by a traditional database or other programming languages for computation. The proposed work aims on preprocessing the log file and keeps track of File inclusion attacks using Hadoop technologies like pig Hive Oozie and Sqoop. The unstructured log file is loaded into HDFS for storage and processing. Data cleaning is the first phase carried out in the proposed work as a preprocessing step in web server log files. The log file contains a number of records that corresponds to automatic requests originated by web clients,that includes a large amount of erroneous, misleading, and incomplete information. In the proposed work such unwanted information is removed using Pig Udf(User Defined Function).Pig Udf is written in java is also responsible for formatting time,date to store in a particular format in HDFS.After applying data cleaning step, applicable resources is stored in the HDFS as CSV file. The whole work is co-ordinated by Oozie,the work flow manager.The output of pig is a structured log file which is given to Hive to extract attack information.Hive allows to write Hive Query Language (HQL) statements that are similar to standard SQL statements. HQL statements are broken down by the Hive service into MapReduce jobs and executed across a Hadoop cluster.Hive uses Naive Bayes Classifiier to search for a particular keywords of File Inclusion attacks. Naive Bayes Classifiier [4] is implemented as Hive UDF which is co-ordinated by Hive Client.The Hive Thrift Client is much like any database client that gets installed on a user's client machine . It communicates with the Hive services running on the server. Hive Thrift Client can be used within applications written in C++, Java, PHP, Python, or Ruby.The attack detected

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. 3, Issue 5, May 2015

records are filtered and stored in a separate file which contains the details of RFI,LFI and SQL injection and the keywords of each attack.IP2C is a small library that provides IP to country resolution using a binary file.Ip2C is invoked in Hive UDF to convert IP address to country to know the attacks origin.Using Sqoop the attack detected records,which contains the cause of attack ,the origin of attack is loaded from HDFS to MYSQL .Oozie's[5] workflow which is built as XML file defines a set of actions to be performed as a sequence or in control dependency. Fig 1 shows the System Architecture.

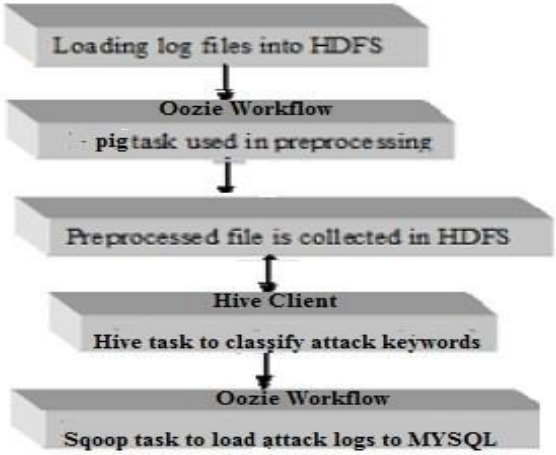

FIG 1.System Architecture

IV.RESULTS AND INTERPRETATION

The web server logs are analyzed for efficient vulnerability attack identification using the Hadoop Mapreduce. This framework is used to compute the log processing in pseudo distributed modes of the cluster. The web server log gathered from an organization's web server is used for processing in Hadoop environment.

The process is analyzed with Apache Hadoop-0.20.2
-cdh3u6. The log files have entries shown in Fig. 2.As can be above, each record in the file is identified by an IP,User ID, date and time, protocol, method,resource,browser,OS used,status,error code, number of bytes transferred Etc. The log file is loaded into HDFS for analyzing.

A. *Pseudo Distributed Mode*

Hadoop framework consist of five daemons namely Namenode, Datanode, Jobtracker, Tasktracker, Secondary namenode. In pseudo distributed mode all the daemons run on local machine simulating a cluster. After applying data cleaning step using Pig UDF applicable resources are stored in the HDFS as CSV file.Oozie will coordinate this process and feed back the CSV file to Hive as input file.

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. 3, Issue 5, May 2015

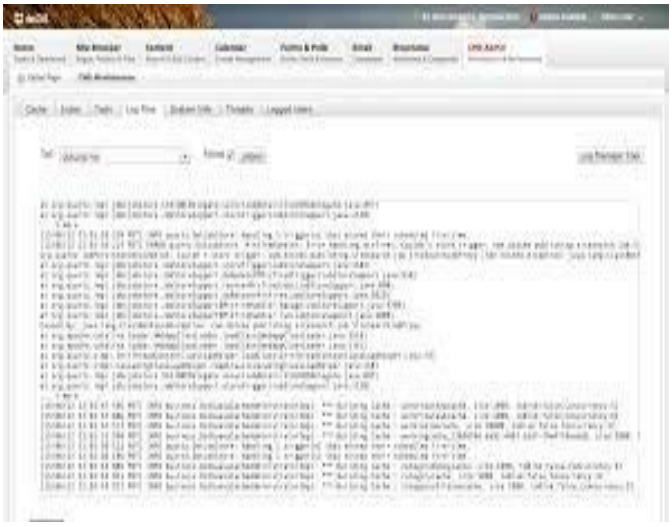

Fig 2.Log File

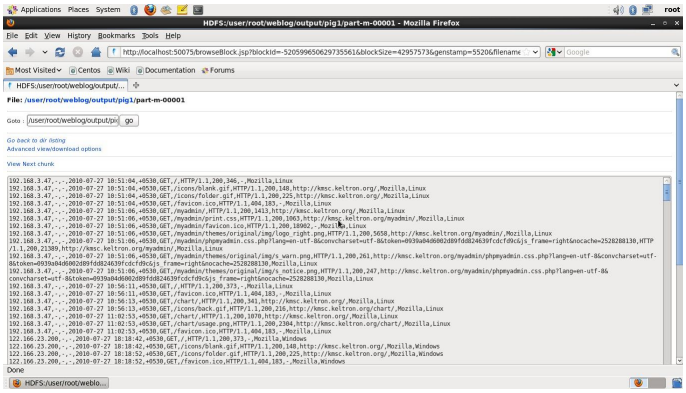

Fig 3. Pig Output

The cleaned web log data as shown in Fig.3 is used to analyze RFI,LFI and SQL injection .Hive is implemented using Hive Client.Hive output is also stored in HDFS which contains additional informations regarding the attack. It is shown in Fig.4.Sqoop loads Hive output file from HDFS to MYsql.It is managed by Oozie workflow.The JobTracker is the service within Hadoop that farms out MapReduce tasks to specific nodes in the cluster.The Oozie web console is enabled the next time the Oozie service is started or restarted.

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. 3, Issue 5, May 2015

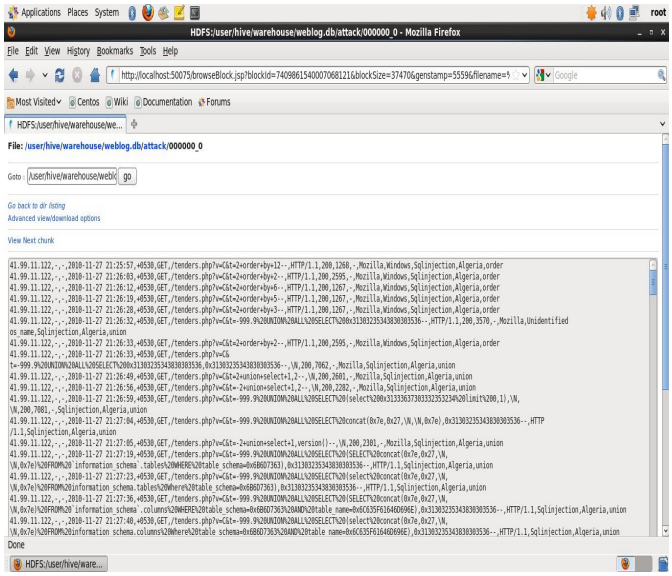

Fig 4. Hive Output

V.FUTURE ENHANCEMENT

Vulnerability analysis can be performed at Real time using Big Data processing framework STORM. New Algorithm can be developed for efficient classification of attack.

REFERENCES

[1]Tom White, "Hadoop: The definitive Guide," Third Edition, ISBN: 978-1-449-31152-0-1327616795
[2]Michal Hubczyk, Adam Domanski, and Joanna Domanska, "Local and Remote file inclusion",Springer 2012
[3]T. Condie, N. Conway, P. Alvaro, J. M. Hellerstein, K. Elmeleegy,and R. Sears, "MapReduce online," in Proc. 7th USENIX Conf.Netw. Syst. Des. Implementation, 2010, p. 21.
[4]Sharma, N., & Mukherjee, S. 2012. Layered approach for intrusion detection using naïve Bayes classifier. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics* (pp. 639-644). ACM. White.
[5]Ian Mitchell, Mark Locke and Aundy Fuller, "The White
Book of Big Data"