Advanced Engineering Technology and Application

ISSN: 2090-9543 VOLUME 5, ISSUES 2, 2016 www.egyptfuture.org/ojs/

Using Big Data Technology for Vulnerability Scanners

Azamat Kulmanov[1,*] *and Lyazzat Atymtayeva*[2]

Kazakh British Technical University, Almaty, Kazakhstan

Received: 10 Mar. 2016, Revised: 20 Apr. 2016, Accepted: 26 Apr. 2016
Published online: 1 May 2016

Abstract: This paper presents the general characteristics of expert systems for processing of data analysis and secure information systems using vulnerability scanners and Big Data technologies. The work of vulnerability scanners is usually based on OWASP security standard recommendations that insist on the processing of various vulnerabilities and attacks. The number of queries to web-sites used by vulnerability scanners during the checking process may increase very rapidly and reach the size of Big Data. Thus, the consideration of Big data analysis becomes actual in this case.

Keywords: Big Data, Hadoop, MapReduce, vulnerability scanner, Cloudera, Hive, HDFS

1 Introduction

Big Data Processing is a challenge not only for the units working directly with clients, but also for the information security departments. Over the past ten years the demand for more reliable protection system led to the need to collect and analyze all the big context data about events and security threats. Below are statistics from the report Information Security Is Becoming a Big Data Analytics Problem, published by Gartner March 23, 2012:

The amount of data analyzed in enterprise information security units annually to double until 2016.

By 2016, 40 percent of companies in order to gather information about security threats will actively analyze at least 10 terabytes of data. In 2011, these companies were less than 3 percent.

In the area of information security management there are some special issues regarding the using Big Data in the processing of Internet queries. For providing the good level of information security any system may be checked by using vulnerability scanners that generate a lot of queries.

2 Vulnerability scanner

Vulnerability scanners is hardware or software serving for the diagnosis and monitoring of networked computers that allows you to scan network computers and applications to detect potential problems in the security system, to assess and address vulnerabilities.

Vulnerability scanners allow you to check a variety of applications in the system for the presence of ”holes” that can be exploited. Also, low-level tools can be used, such as a port scanner, to identify and analyze possible applications and protocols running on the system.

The number of threats is growing in proportion to the growth of the business, however, as demonstrated by long-term practice, 99% of attacks occur over a dozen standard validation error incoming data, or discovered vulnerabilities in the installed components of third-party software, or corny, for negligence of system administrators, using the settings and passwords set by default.

Community OWASP(Open Web Application Security Project) [1] is engaged by classification of attack vectors and vulnerabilities. It is an international non-profit organization focused on analyzing and improving software security.

OWASP has created a list of 10 most dangerous attack vectors to Web-based applications, this list is called OWASP TOP-10 [2], and it focused the most dangerous vulnerabilities, which can cost some people a lot of money, or of undermining the goodwill, up to loss of business.

In consideration of checking the security level of any system vulnerability scanners gives log data about vulnerability and mistakes found. Manual processing of

log data requires a huge amount of time, and hence it can be a tedious task. Since Volume, Velocity and Variety are being dealt in our case.

3 Using of Hadoop platform, HDFS and MapReduce Technology

Big Data technology often implies the using of Hadoop [3] platform. Hadoop is a complex system consisting of a large number of components. Install and configure a system on their own - a very difficult task. Therefore, many companies now offer ready the Hadoop distributions, including the deployment tools, administration and monitoring. Hadoop platform is usually distributed as a commercial (products from companies such as Intel, IBM, EMC, Oracle), and under free (Cloudera company products, Hortonworks and MapR) licenses [4].

One of the distribution of Hadoop [3] is Cloudera [8]. Key of the product - CDH (Cloudera Distribution including Apache Hadoop) - a bunch of the most popular tools of Hadoop infrastructure for Cloudera Manager control. The manager takes over responsibility for the deployment of a cluster, the installation of all components and their further monitoring. Among related to Apache Hadoop software projects included in the distribution: Flume, HBase, Hive [7], Mahout, Oozie, Pig, Sqoop, Whirr, Zookeeper. We install last CDH4 version based on Hadoop 2.0 (including module YARN), in the CDH4 also included three own company product - Hue (browser Hadoop-cluster management interface), Impala and Search (full-text and faceted search in media HDFS and HBase).

Advantages of Hadoop platform is very nice and smooth scalability. Those, if we need to handle twice as much data, or to store twice the data, we simply add twice more machines in the cluster.

Zero cost software. There you have a lot of data, you can go to the Oracle company, buy a couple of million dollars a clustered Oracle, and as many advisers, it is not suitable for all companies, especially startup companies. Wdo not know well, some new social network, they simply can not afford to spend a few million on Oracle. They can afford to Hadoop, take the cluster and use the open-source Hadoop, as data analysis and storage.

When we talk about Hadoop, the first thing we have in mind is filesystem - HDFS (Hadoop Distributed File System) [6]. The easiest way to think about HDFS (Figure 1) is to present a normal file system, only bigger. The usual file system, by and large, consists of a table of file descriptors, and the data area. The HDFS table instead uses a special server - a name server (NameNode), and the data is scattered across data servers (DataNode).

The rest of the difference is not so much: the data is divided into blocks (usually 64MB or 128MB), for each file name server stores its way, a list of the blocks and

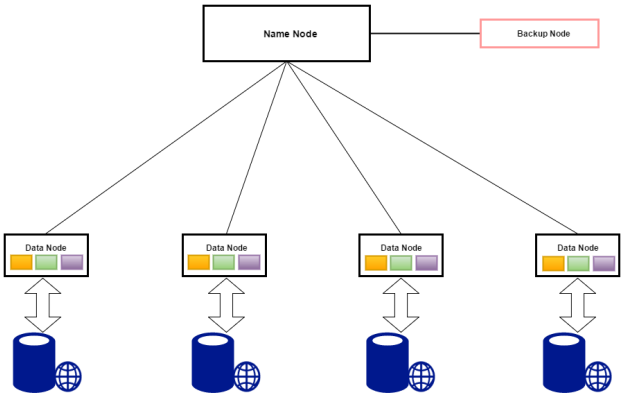

Fig. 1: Architecture of HDFS.

their replicas. And to ensure reliability, each block is stored in multiple instances on multiple machines. This ensures reliability, even if we fail, say, 10% of the machines in the cluster, most likely, we will not lose anything. Those, yes, we will lose some blocks, but since these blocks are stored in several copies, we can again, and to read and write.

We consider the using of Hadoop platform together with MapReduce technology by the reason that data processing is performed by using Hadoop MapReduce technology [5]. According to this technology a huge amount of information is divided into parts, and the processing of each of the these parts are entrusted to a separate server.

Typically, the data is processed on the same servers where they are stored, which allows for faster processing and avoid unnecessary data movement between servers. The results are then combined into a single unit.

MapReduce architecture (Figure 2) is built on the principle of master - workers. As the main acts JobTracker server, distributing tasks subordinate nodes in the cluster and controls their implementation.

Processing data is divided into the following stages:

1. Run the application: the transfer of the application code to the main (master) and slave units (workers);

2. Master assigns specific tasks (Map or Reduce) and distributes the input of the data on the compute nodes (workers);

3. Map-designated nodes read their input and start their processing;

4. Map-nodes locally store the intermediate results: each node stores the result in the local drives;

5. Reduce-nodes intermediate data read from the Map-Reduce nodes and perform data processing;

6. Reduce-nodes store the final results in the output files, usually in HDFS [6].

The advantage of MapReduce is that it allows distribution and operation pre-processing and convolution. Preprocessing operation operate independently of each other and can be performed in

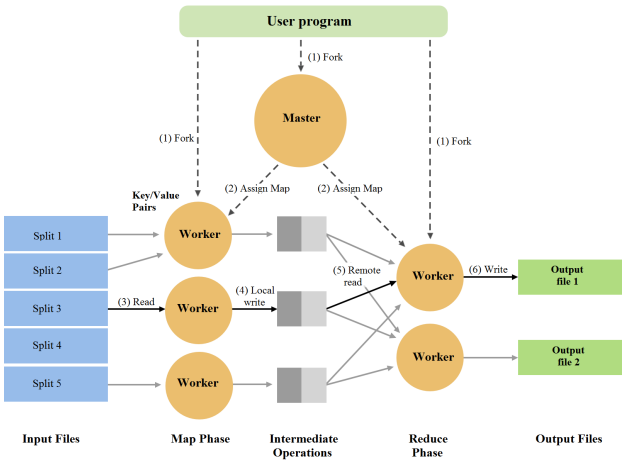

Fig. 2: Architecture of MapReduce.

parallel (although in practice it is limited to the input source and / or the number of processors used). Similarly, a plurality of operating units can perform the convolution - for it is only necessary that all of the pre-treatment with one particular key value is processed by one worker node at a time. Although this process may be less effective than a sequential algorithm, the MapReduce can be applied to large data volumes, which can handle a large number of servers. Thus, the MapReduce can be used to sort a petabyte of data, which will only take a few hours. Parallelism also gives some possibilities of recovery after partial server failures: if the working unit, the producing step pretreatment or convolutions fails, its operation can be transferred to another working unit (assuming that the input of an operation for available).

The framework is largely based on the functions map and reduce, commonly used in functional programming, although the actual semantics of the framework is different from the prototype.

4 Using HiveQL for queries and results

MapReduce is a very powerful data processing tool, but it can be quite difficult to establish and maintain, while many companies operate business analysts who are able to write well in SQL queries, but do not know how to write code in Java. Also, many organizations have programmers who can write code in scripting languages. Hive [7] and Pig - it's two projects, which were developed independently of each other and are designed to help these analysts and programmers to effectively use MapReduce analysis for large data sets.

Hive - an add-on Hadoop in order to facilitate tasks such as the accumulation of data, non-programmable queries and analysis of large data sets:

Hive can be used by those who know SQL;

Hive creates MapReduce jobs that run on Hadoop cluster;

Definitions of tables in Hive are built on the data in HDFS [6].

Hive can be used for interactive data exploration or create reusable tasks batch job processing. Hive allows you to create the structure for the mostly unstructured data. After determining the structure you can use to create a Hive query the data without knowledge of Java or MapReduce. HiveQL (Hive Query Language) allows you to create queries using operators such operators MySQL. Hive understands how to work with structured and semi-structured data, such as text files in which fields are separated by special characters.

In our case, we will process the data using Hive. We scanned http://baskino.club/ site and received the report about the vulnerabilities and prepared the report in CSV format. Then, we create a table, and a file recorded in the table on Hadoop data.

Creating table:

```
create table log_file_table
pluginid int, alert string,
riskcode int, confidence int,
riskdesc string, url_site string)
row format delimited fields
terminated by ',';
```

Record file to the database:

```
load data local
inpath '/tmp/log.csv' overwrite
into table log_file_table;
```

We now have the data base is ready, we scanned only 1 site, now imagine, Kazakhstan has about 7000 sites, if we scan all the data will be very large.

Now consider the queries, if they are many, they can be parallel, and expressed as the Map-Reduce tasks.

For example, the standard HiveQL query [7]:

```
select alert, count(url_site)
from log_file_table group by alert
```

Here Map:

```
line => ({alert}, url_site1)
```

Reduce:

```
({alert}, [url_site1, ..., url_site n])
=> ({alert }, count(url_site))
```

We brought only one example of data analysis and processing, but using Hive can extract any data in the right form for vulnerability analysis (Figure 3).

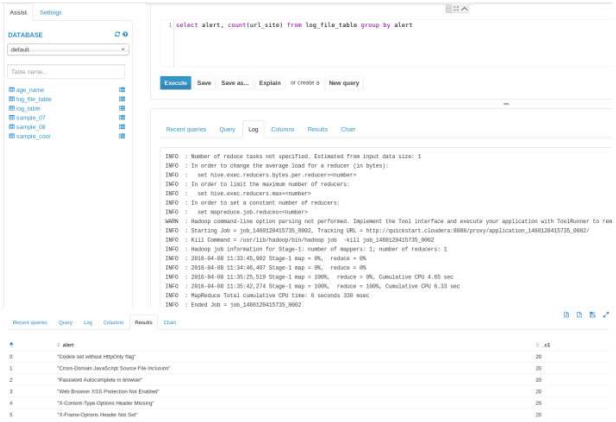

Fig. 3: HiveQL query and result.

5 Conclusion

By implementing the current technologies (IT security, Big Data, expert systems) within organization, there results a proper environment for analysis and development. Combining conceptual models of Big Data and IT expert systems [9]-[11] and worthless-considered data analysis. Looking ahead, Big Data have become one of the most discussed topics in recent years, and information systems analysis is an important factor in shaping business decision making systems.

References

[1] Troy Hunt, OWASP Top 10 for .NET developers, Release 1.0.8, pp. 1415, 19 Dec 2011.

[2] Troy Hunt, OWASP Top 10 for .NET developers, Release 1.0.8, pp. 1213, 19 Dec 2011.

[3] Tom White, Hadoop: The Definitive Guide, Third Edition, pp. 115, 2012.

[4] Tom White, Hadoop: The Definitive Guide, Third Edition, pp. 613619, 2012.

[5] J. Dean and S. Ghemawat, Mapreduce: simplified data processing on large clusters, Comm. ACM 51:1, pp. 107113, 2008.

[6] Tom White, Hadoop: The Definitive Guide, Third Edition, pp. 4580, 2012.

[7] Venner, Jason (2009). Pro Hadoop. Apress. ISBN 978-1-4302-1942-2.

[8] Tom White, Hadoop: The Definitive Guide, Third Edition, pp. 619621, 2012.

[9] Atymtayeva L., Akzhalova A., Kozhakhmet K. Main Issues of the Software Development for Knowledge Base Processing in the Intelligent Applications for Information Security Audit. // Journal ”Recent Advances in Computer Engineering, Communications and Information Technology”, ISBN: 978-960-474-361-2, pp. 271-280.

[10] Atymtayeva L., Kozhakhmet K., Bortsova G. Building a Knowledge Base for Expert System in Information Security. // Springer Journal Advances in Intelligent Systems and Computing, Volume 270 ”Soft Computing in Artificial Intelligence ”, pp. 57-77

[11] Sheriyev M., Atymtayeva L. Automation of HCI Engineering processes: System Architecture and Knowledge Representation // Int.Journal ”Advanced Engineering Technology and Application (AETA)”, Natural Science Publishing, Vol.4, N2 (May 2015), ISSN 2090-9535, pp. 41-46.

Azamat Kulmanov received the master degree in Information Systems at Kazakh British Technical University in Almaty, Kazakhstan. The Author has been pursuing his research work under the guidance of Lyazzat Atymtayeva, Doctor of Science degree in Mechanics, Mathematics and Computer Science. His research Interest includes Big Data Technology, Hadoop, Expert systems and Vulnerability Scanners.

Lyazzat Atymtayeva received the PhD and Doctor of Science degree in Mechanics, Mathematics and Computer Science at al-Farabi Kazakh National University, Kazakhstan. Her research interests are in the areas of mechanics, applied mathematics and computer science including the numerical and rigorous mathematical methods and models for mechanical engineering and computer science, intelligent and expert systems in Information Security, Project Management and HCI. She has published research papers in reputed international journals of mathematical and computer sciences. She is reviewer and editor of international journals in mathematics and information sciences.