Malware Analysis Using Hadoop and MapReduce.

Nilesh Kisanrao Dengle
Department of Information Technology
Pune Institute of Computer Technology
Pune, India
nileshdengle20@gmail.com

Prof. S. C. Dharmadhikari
Department of Information Technology
Pune Institute of Computer Technology
Pune, India
scdharmadhikar@pict.edu

***Abstract*—Today each and every day a lot of data is generated in increasing order. This is because of today's e-commerce and easy to use technologies. Also, there is increasing number of vulnerabilities in this large data. There are counter measures for these vulnerabilities like anti-viruses or anti-malwares. But, for scanning a large data in less time its difficult. So using Hadoop and MapReduce technology we can scan it parallely in less time. In this project we are scanning malware using Hadoop and MapReduce.**

Keywords-Malware,Hadoop,MapReduce

I. INTRODUCTION

A large amount of data is created by each and every individual in today's era and will continue in exponential manner. Now there are many technologies to store this large amount of data. Also, there are cloud in which the security can be provided. There are lot servers which are provided by the companies like Amazon, EMC etc.
Now, as there is ease of storing and loading data easily which is an advantage. There is also disadvantage that there are a lot vulnerability which can infect the data. There are lot different anti-malware software's which detect the malware and avoid affecting the valuable data. But the main concern is about time and optimization to scan the malwares.

To scan malwares in large data we can do it with parallel functionality. This can be done with the help of Hadoop [2] framework. The MapReduce [10] developed by the Google works for assigning the job parallely. Let's talk about Hadoop, the main architecture of Apache Hadoop consists of Hadoop Distributed File System which is used for storage and MapReduce for the parallel processing. Hadoop divided the file into the blocks and makes the replication of the blocks in different nodes. To work in parallel we have to submit the code to the Hadoop MapReduce. The nodes take the configuration and work accordingly. Due to this, there is the advantage of parallel working with data which are distributed in different locality. With high end architecture of today's generation and high speed net there is a reliable result with less fault tolerance[13].

The MapReduce concepts have the two separate method that Hadoop performs. The first task is the map job, which converts the data into the MapReduce form. The data is individually break down in the tuples. The tuples are the elements in the key/value pairs. Now, the mappers will work according to the process given to it by the MapReduce. The reducers take the outcome of the mapper as input and combine this element in the similar data tuples with the reference of the key/value pair. As the name is MapReduce the reduce function is always performed after the map[14].

So, the main purpose of the project is to scan the malware in the large data with the help of the Hadoop and MapReduce technologies. The malware scanning code should be written in the MapReduce. The paper is organized as follows: Section II presents the literature review in the area of malware-detection algorithms. In Section III, a description about proposed system of malware detection using Hadoop and MapReduce. Section IV presents the discussion and conclusion.

.

II. RELATED WORK

Mostly in large data set there has been the focus on the intrusion detection system then scanning the malwares on the host machines. This is because widely used World Wide Web. Due to the vast use of the internet all the vulnerability and attacks are done with help of the internet. So, the concentration is done in the intrusion detection system. Many works have done in this area by using different technologies which are as follows.

Ibrahim Aljarah describes an intrusion detection system (IDS) based on a parallel particle swarm optimization clustering algorithm using the MapReduce methodology [3]. This paper presents a parallel intrusion detection system (IDS-MRCPSO) based on the MapReduce framework since it has been confirmed as a good parallelization methodology for many applications [3]. In addition, the proposed system incorporates clustering analysis to build the detection model by formulating the intrusion detection problem as an optimization problem[11].

In, authors focus on the specific problem of Big Data facing in network intrusion traffic. It tells the system challenges presented by the today's Big Data problems associated with network intrusion problems[4]. It describes the management in big data, network topology which gives a specific which used HDFS and public cloud in it [12]. It also defines the communication challenges in case of bandwidth.

In[9] author present Aesop, a scalable algorithm that identifies malicious executable files by applying Aesop's moral that "a man is known by the company he keeps." They use a large dataset voluntarily contributed by the members of Norton Community Watch, consisting of partial lists of the files that exist on their machines, to

identify close relationships between files that often appear together on machines. Aesop leverages locality-sensitive hashing to measure the strength of these inter-file relationships to construct a graph, on which it performs large scale inference by propagating information from the labeled files (as benign or malicious) to the preponderance of unlabeled files.

Author[10] proposes a novel behavioral malware detection approach based on a generic system-wide quantitative data flow model. They base their data flow analysis on the incremental construction of aggregated quantitative data flow graphs. These graphs represent communication between different system entities such as processes, sockets, files or system registries. Authors demonstrate the feasibility of our approach through a prototypical instantiation and implementation for the Windows operating system. The experiments yield encouraging results: in our data set of samples from common malware families and popular non-malicious applications.

III. PROPOED SYSTEM

The scalability and parallel processing should be possible with average computer hardware and which can be made possible with the Hadoop platform. And by the help of Linux OS it becomes more secure and reliable.

The main concern is writing the MapReduce code for scanning the malwares in the large data. After the code is written the MapReduce will split the process in Mappers and Reducers. MapReduce in Hadoop comes with a choice of schedulers. The default is the original FIFO queue-based scheduler, and there are also multiuser schedulers called the Fair Scheduler and the Capacity Scheduler.

We will run our code to the job driver. The job driver will copy the job configuration to the name node as it has information about all data nodes. Now the job driver will submit the code to the job tracker. The job tracker will distribute task configuration to the task tracker. The task configuration will have the malware signatures which I have to match with the data which are stores in the HDFS. Through the task tracker the configuration is given to different Mappers through which the processing is distributed and the malwares will be scanned parallely the data which resides in the datanodes.

After scanning all the data either while mapping is done the same malwares can be found. Then the reducer will sort the repeated detected malwares and will minimize the result. In this way the proposed system will work for finding the malwares in the large data set.

Now with the reference of the figure 1 the client will give the input of the malware scanning to the HDFS through the job driver. Then the process of mapper and reducer will split the process and work in parallel then the output will be saved in the HDFS output and will be given back to the client. The process may be fast depending on the MapReduce code and the language which is suitable to use.

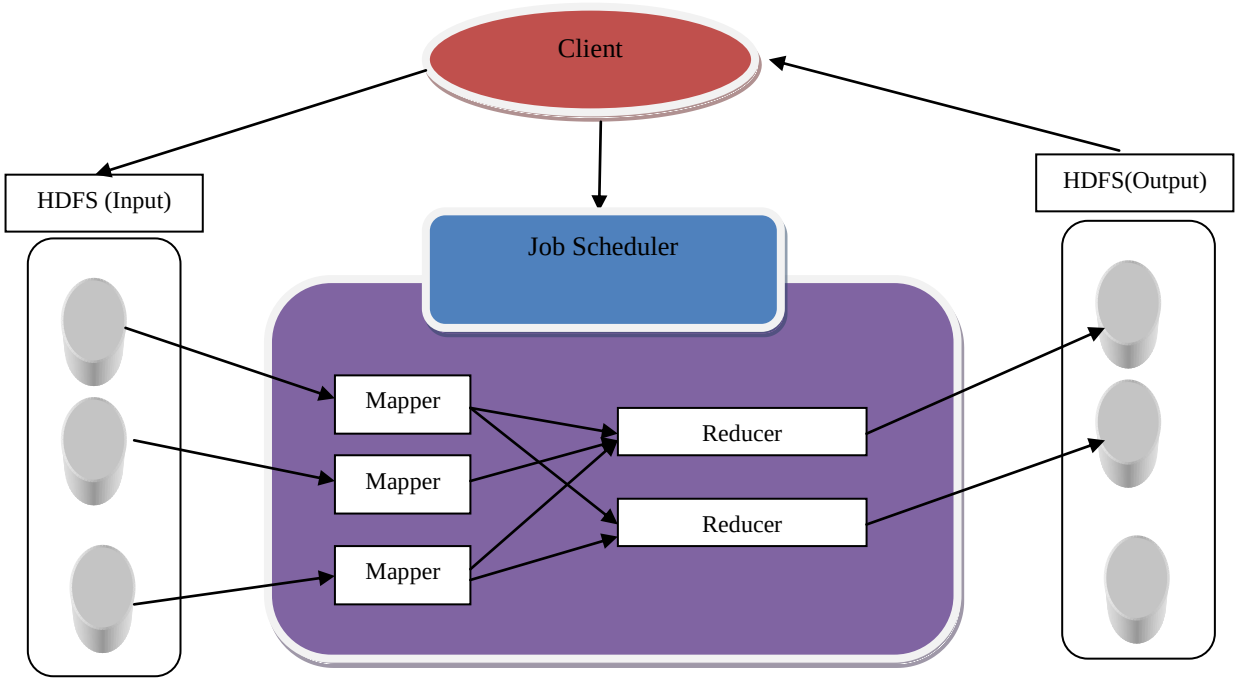

Fig no 1 Malware Detection Flow

IV. Discussion and Conclusion

In this paper we proposed an efficient and latest method to scan the malware in large data set in minimum time. It can be possible due to the recent technologies like Hadoop and MapReduce. The main concern of the project is that how the MapReduce configuration and code to be written. There are many languages in which MapReduce can be written like java, python, and ruby. There are also different platform which are built over the MapReduce like Apache pig and Hive. Apache pig is scripting language which is called the pig latin.

With the literature review we found that most of the detection system are based on the intrusion detection system (IDS) this because wide use of internet in today's era. Other system uses the clustering method for finding the malware also there are some limitations in that. So by this we came to conclusion that there should a parallel processing through which we can detect the malware. So the Hadoop and MapReduce technology are the recent one which can fault tolerance and reliable. Also we are trying to study the pig latin which is an scripting language. This scripting language can be reliable for writing the malware detection code.

Referrences

[1] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," in Proceedings of the OSDI '04, 2004, pp. 137–150.

[2] Apache Hadoop, available at: http://hadoop.apache.org (2013).

[3] Ibrahim Aljarah and Simone A. Ludwig. "MapReduce Intrusion Detection System based on a Particle Swarm Optimization Clustering Algorithm," Evolutionary Computation (CEC), 2013 *IEEE Congress*, June 2013.

[4] Shan Suthaharan. "Big data classification: problems and challenges in network intrusion prediction with machine learning." ACM, March 2014.

[5] http://wiki.apache.org/hadoop.

[6] Tobias Wüchner, Martín Ochoa and Alexander Pretschner. "Malware Detection with Quantitative Data Flow Graphs." ACM 978-1-4503-2800-5/14/06

[7] Zhiyong Shan and Xin Wang. "Growing Grapes in Your Computer to Defend Against Malware."IEEE, VOL. 9, NO. 2, FEBRUARY 2014

[8] T. White, Hadoop: The Definitive Guide, original ed.O'Reilly Media, Jun. 2009.

[9] Acar Tamersoy, Kevin Roundy and Duen Horng Chau , "Guilt by Association: Large Scale Malware Detection by Mining File-relation Graphs," KDD'14, August 24–27, 2014

[10] Tobias Wüchner, Martín Ochoa and Alexander Pretschner. "Malware Detection with Quantitative Data Flow Graphs." ACM 978-1-4503-2800-5/14/06

[11] Ibrahim Aljarah and Simone A. Ludwig. "Towards a Scalable Intrusion Detection System based on Parallel PSO clustering Using MapReduce." ACM 978-1-4503-1964-5/13/07

[12] K. Shvachko, H. Kuang, S. Radia, R. Chansler, "The Hadoop Distributed File System," 26th IEEE Symposium on Mass Storage Systems and technologies, Yahoo!, Sunnyvale, pp. 1-10, May 2010

[13] http://wiki.apache.org/hadoop.

[14] http://www-01.ibm.com/software/data/infosphere /Hadoop /MapReduce/