Malware and Log file Analysis Using Hadoop and Map Reduce

Mr. Sachin M Nehe
ME Computer IInd Year
Pravara Rural Engineering College, Loni, MH, INDIA,
ssachinnehe@gmail.com

Abstract - Today each and every day a lot of data is generated in increasing order. This is because of today's ecommerce and easy to use technologies. Also, there is increasing number of vulnerabilities in this large data. There are counter measures for these vulnerabilities like antiviruses 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. There are various applications which have a huge database. All databases maintain log files that keep records of database changes. This can include tracking various user events. Apache Hadoop can be used for log processing at scale. Log files have become a standard part of large applications and are essential in operating systems, computer networks and distributed systems. Log files are often the only way to identify and locate an error in software, because log file analysis is not affected by any timebased issues known as probe effect. This is opposite to analysis of a running program, when the analytical process can interfere with time-critical or resource critical conditions within the analyzed program. Log files are often very large and can have complex structure. Although the process of generating log files is quite simple and straightforward, log file analysis could be a tremendous task that requires enormous computational resources, long time and sophisticated procedures. This often leads to a common situation, when log files are continuously generated and occupy valuable space on storage devices, but nobody uses them and utilizes enclosed information. The overall goal of this project is to design a generic log analyzer using hadoop map-reduce framework. This generic log analyzer can analyze different kinds of log files such as- Email logs, Web logs, Firewall logs Server logs, Call data logs.
Index Terms - Malware, Hadoop, MapReduce, Log files, log analyzer, Heterogeneous database.

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 parallel. 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 actual MapReduce ideas possess the a couple distinct techniques in which Hadoop executes. The initial activity may be the place employment, which usually converts the data in the MapReduce variety. The results are independently break down from the tuples. The actual tuples are classified as the elements from the key/value twos. At this point, your mappers works based on the process inclined to this by simply your MapReduce. The actual reducers get the results from the mapper seeing that feedback and mix this specific element in your comparable information tuples with the referrals from the key/value set of two. For the reason that title is MapReduce your lower function is actually carried out following the map[14]. And so, the main intent behind your task would be to scan your spyware and adware from the large information with the aid of your Hadoop and MapReduce technology. The actual spyware and adware deciphering value needs to be prepared from the MapReduce. The actual cardstock is prepared the following: Segment II reveals your materials examine in your community connected with malware-detection algorithms. In Segment III, some sort of explanation about proposed system connected with spyware and adware detection utilizing Hadoop and MapReduce. Segment 4 reveals your debate and bottom line. Existing software applications often create (or can be configured to produce) a number of auxiliary textual content data files known as log data files. These kinds of data files are used during a variety of development of computer software improvement, primarily intended for debugging and also profiling purposes. Use of log data files assists examining by creating debugging easier. This lets you follow this reason from the system, on advanced, without having to work that in debug function. These days, log data files are normally utilized on customers installation when it comes to lasting computer software supervising and/or finetuning. Wood data files evolved into a

normal component of substantial application and so are essential in operating systems, personal computer sites and also spread methods.

II. RELATED WORK

Generally within significant facts arranged there is the give attention to the intrusion diagnosis process after that deciphering the malwares in the coordinator devices. It is because widely used Planet Broad Web. Because of the huge utilization of the online world every one of the vulnerability and episodes tend to be through with help with the World Wide Web. So, the awareness is performed inside the intrusion diagnosis process. Many functions do in this field simply by making use of different technological knowhow that happen to be as follows. Ibrahim Aljarah describes an intrusion diagnosis process (IDS) according to the parallel particle swarm search engine optimization clustering protocol using the MapReduce methodology [3].

This particular cardstock provides the parallel intrusion diagnosis process (IDSMRCPSO) good MapReduce framework due to the fact it is often verified to be a excellent parallelization methodology for a lot of software [3]. Furthermore, the recommended process incorporates clustering evaluation to make the diagnosis model simply by forming the intrusion diagnosis dilemma for search engine optimization problem [11]. With, creators give attention to the precise dilemma involving Big Facts going through within multilevel intrusion visitors. This conveys to the system challenges displayed through the todays Big Facts problems regarding multilevel intrusion problems [4]. In this paper describes the managing within major facts, multilevel topology gives a particular which often used HDFS and open fog up inside it [12]. In addition, it specifies the transmission challenges within scenario involving bandwidth. In[9] article author found Aesop, the scalable protocol in which determines detrimental executable data files by making use of Aesops ethical in which a man is known through the organization they continues. These people start using a significant dataset voluntarily offered through the associates involving Norton Group View, composed involving partial provides with the data files that exist on the devices, to name near romantic relationships between data files in which often glimpse with each other in devices. Aesop harnesses locality-sensitive hashing for you to evaluate the effectiveness of these kind of inter-file romantic relationships to build the graph, on what the idea executes significant degree inference simply by propagating facts through the described data files (as cancerous or perhaps malicious) on the preponderance involving unlabeled data files.

Author [10] suggests the fresh behavior spyware and adware diagnosis strategy according to the general system-wide quantitative facts move model. These people bottom their particular facts move evaluation for the incremental structure involving aggregated quantitative facts move charts. Most of these charts characterize transmission between different process entities for instance operations, electrical sockets, data files or perhaps process registries. Creators of these studies demonstrate the feasibility of our own strategy through a prototypical instantiation and enactment for that Windows operating system. This trials yield stimulating final results: inside our facts number of trials through common spyware and adware people and common non-malicious software.

III. PROPOSED SYSTEM

The particular scalability in addition to parallel running should be feasible together with average computer systems in addition to and this can be manufactured feasible with all the Hadoop platform. And also by simply the assistance of Linux OS that gets to be better in addition to reliable. The primary matter is actually writing the particular MapReduce rule regarding scanning the particular malwares within the significant information. Following your rule is actually composed the particular MapReduce may divided accomplishing this throughout Mappers in addition to Reducers.

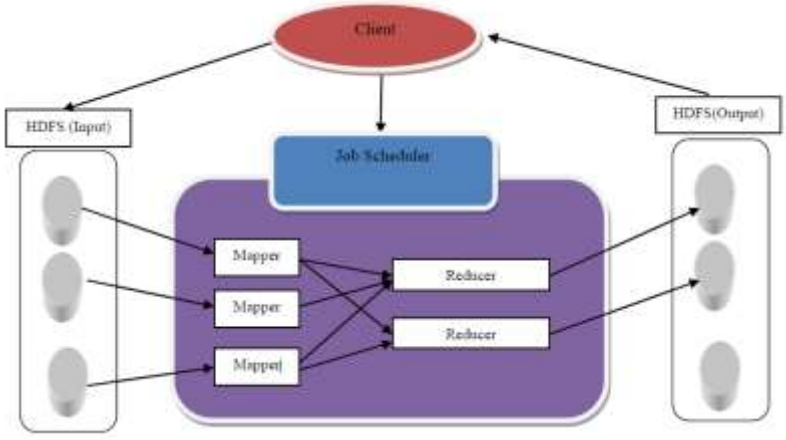

Fig. 1. Malware Detection Flow

MapReduce throughout Hadoop includes a choice of schedulers. The particular default is the first FIFO queue based scheduler, in addition to there are also multiuser schedulers called the particular Sensible Scheduler as well as the Ability Scheduler. All of us will probably run each of our code to the job driver. The project driver will probably backup the job configuration to the name node seeing that it has info on just about all facts nodes. Now the job driver will probably post this code to the job tracker. The project tracker will probably send out process configuration to the process tracker. The process configuration can have this adware and spyware signatures which in turn I've got to fit while using the facts which might be merchants from the HDFS. Over the process tracker this configuration can be presented for you to unique Mappers where this digesting can be sent out along with the malwares is going to be scanned parallel this facts which in turn exists from the data nodes. Immediately after checking the many files possibly although mapping is completed a similar malwares can be found. Next the reducer will kind the actual repetitive diagnosed malwares

and will lessen the actual result. That way the actual planned method will work for seeking the malwares in the substantial files collection. Today while using guide in the determine 1 the client will deliver the actual feedback in the spyware checking on the HDFS through the employment drivers. Next the technique of mapper and also reducer will break up the procedure and also work inside parallel then this productivity will probably be stored in the HDFS productivity and will be presented time for the client. The process could possibly be rapid dependent upon the actual MapReduce signal plus the words which is acceptable to be able to use. Generic Log Analyzer can be used to analyze various kinds of logs such as:

1. Email logs
2. Web logs
3. Firewall logs 4. Server Log

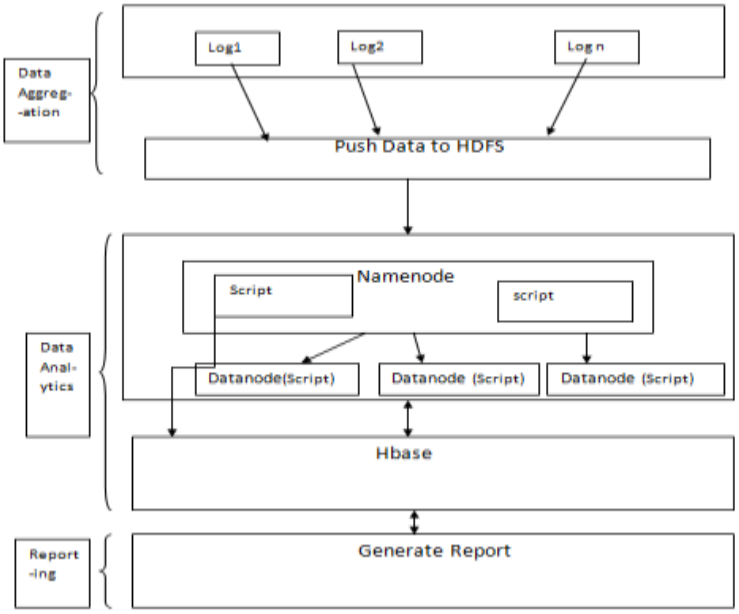

Fig. 2. Architecture of Log file Analyzer

This System will build Generic Log Analyzer for different types of large-scale log files by Taking advantage of Hadoop map-reduce framework and polymorphism for log analysis and will Increase efficiency and reliability of log analysis.

Algorithm1:Malware Analyzer

Input :Raw unstructured Data Output
Malwares Present in the Database begin
Collect file to database;
for each file in the database begin extract file signatures;
append signature in signature file;
end for
load signature file into Hadoop;
if signature file is in Apache format then Display or Store Result;
end if for each signature in signature
file begin
if signature has a match database then filter out the corresponding file ;
End if
End For

IV. RESULTS AND DISCUSSION

In the proposed approach for Malware detection through MapReduce we have performed different experiments; from those experiments we have analyzed different aspect with proper flow of steps in above chapter. First of all we are working on the Hadoop framework which executes each of its processes in MapReduce which is parallel processing. For proposed operation, we are using Apache Pig for MapReduce coding. This programming tool works on data flow; so we execute each step with proper flow such that optimized output should come.

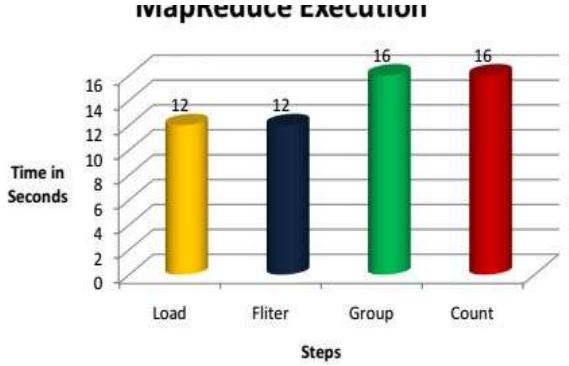

Fig. 3. MapReduce Execution

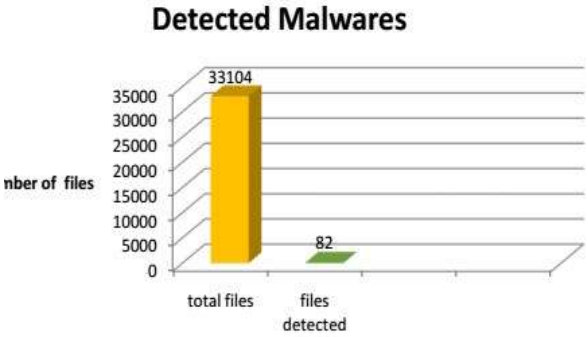

Fig. 4. Detected Malware

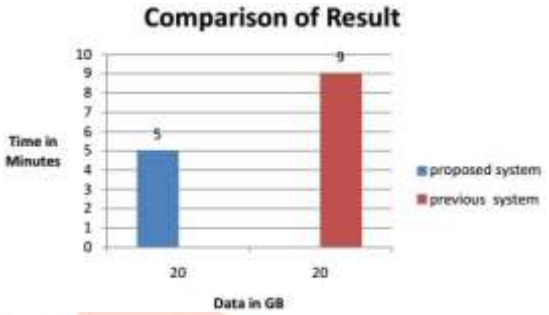

Fig. 5. Comparison of Result

The main advantage of the Apache Pig is that each of its functions in which we have to display or store is done in MapReduce. The above graph gives us the clear timing in seconds for the execution of different steps. The execution is done for output purpose of each step. The output is either displaying it on terminal or storing it in the HDFS.

V. CONCLUSION AND FUTURE WORK

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.

Many systems based on these models have been developed to deal with various real problems of data integration, such as collecting information from different web sites, integrating spatially-related environmental data, and aggregating information from multiple enterprise systems. But all that system is work on only single type of log files.so with the help of map reduce framework This System will be able to analyze many types of log files. Due to use of Hadoop framework, Efficiency of log analysis has improved If any new standard format log file is created then it will be easy to extend our project to analyze that log file. Our project can also be implemented on windows so that novice users find it easy to use.

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.

VI. REFERENCES

[1] J. Dean and S. Ghemawat, Mapreduce: Simplified data processing on large clusters, in Proceedings of the OSDI 04, 2004, pp. 137150.
[2] Apache Hadoop, available at:http://hadoop.apache.org.
[3] Ibrahim Aljarah and Simone A. Ludwig.MapReduce Intrusion Detection System based ona Particle Swarm Optimization ClusteringAlgorithm, Evolutionary Computation (CEC), 2015IEEE Congress, June 2015.
[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 Wchner, Martn Ochoa and Alexander Pretschner. Malware Detection with Quantitative Data Flow Graphs. ACM 978-14503-2800-5/14/06.
[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.OReilly Media, Jun. 2009.
[9] Acar Tamersoy, Kevin Roundy and Duen Horng Chau , Guilt by Association: Large Scale Malware Detection by Mining Filerelation Graphs, KDD14, August 2427, 2014.
[10] Ibrahim Aljarah and Simone A. Ludwig. Towards a Scalable Intrusion Detection System based on Parallel PSO clustering Using MapReduce. ACM 978-1-4503- 19645/13/07.
[11] 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. 110, May 2010.
[12] http://wiki.apache.org/hadoop.
[13] http://www-01.ibm.com/software/data/infosphere/Hadoop /MapReduce/.
[14] J.H. Andrews,Theory and Practice of Log File Analysis Technical Report,Pennsylvania Western Ontario.
[15] Jan Waldamn,Log File Analysis Technical Report. Hadoop