ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

A Security System That Addresses the Problem of Dynamically Mining and Extracting Information

G R Rajitha
M.Tech Student,
Department of CSE,
School of Technology,
GITAM University, Hyderabad.

M Akkalakshmi, M.Tech, Ph.D
Professor,
Department of IT,
School of Technology,
GITAM University, Hyderabad.

Abstract:

For describing a current and/or future structure and behavior for an organization's security processes, information security systems, personnel and organizational sub-units, so that they align with the organization's core goals and strategic direction. Although often associated strictly with information security technology, it relates more broadly to the security practice of business optimiza- tion in that it addresses business security architecture, performance management and security process architecture as well.Large Enterprises routinely collect terabytes of security-relevant data on a regular basis. This volume and variety of data is going to increase in future. Existing data analytical techniques don't work well at large scales and typically produce so many false positives that their efficacy is undermined. The problem becomes worse as enterprises move to cloud architectures and collect much more data.Big data is a broad term for data sets so large or complex that traditional data processing ap- plications are inadequate. Challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set.Analysis of data sets can find new correlations, to "spot business trends, prevent diseases, combat crime and so on." Scientists, practitioners of media and advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research.In this pa- per we propose a system that addresses the problem of dynamically mining and extracting information from the large log data produced by an extensive variety of security products in a large Network.

Keywords:

Big Data, Analytics, Large enterprises, Security products, Dynamic mining, log data.

Introduction:

wireless sensor networks in large corporate enterprise. Relational database management systems and desktop statistics and visualization packages often have difficulty handling big data. The work instead requires "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make Big Data a moving target.Thus, what is considered to be "Big" in one year will become ordinary in later years. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant con- sideration." Big data can be described by the following characteristics:

Volume :

The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered Big Data or not. The name 'Big Data' itself contains a term which is related to size and hence the characteristic.

Variety :

The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

analysts. This helps the people, who are closely analyzing the data and are associated with it, to effectively use the data to their advantage and thus upholding the importance of the Big Data.

Velocity:

The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.

Variability:

This is a factor which can be a problem for those who analyse the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

Veracity:

The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.

Complexity :

Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data.Big data analytics consists of 6 Cs in the integrated industry 4.0 and Cyber Physical Systems environment. 6C system, that is, consist of connection (sensor and networks), Cloud (computing and data on demand), Cyber (model and memory), content/context (meaning and correlation), community (sharing and collaboration), and customization (personalization and value). In this scenario and in order to provide useful insight to the factory management and gain correct content, data has to be processed with advanced tools (analytics and algorithms) to generate meaningful information. Considering the presence of visible and invisible issues in an industrial factory, the information generation algorithm has to capable of detecting and addressing invisible issues such as machine degradation, component wear, etc. in the factory floor. Security information management (SIM) is an industry term related to information security referring to the collection of data (typically log files) into a central repository for trend analysis. SIM products generally are software agents running on the computer systems that are to be monitored, which then send the log information to a centralized server acting as a "security console". The console typically displays reports, charts, and graphs of that information, often in real time. Some software agents can incorporate local filters, to reduce and manipulate the data that they send to the server, although typically from a forensic point of view you would collect all audit and accounting logs to ensure you can recreate a security incident.

SIEM :

Security information and event management (SIEM) is a term for software products and services combining security information management (SIM) and security event management (SEM). SIEM technology provides real-time analysis of security alerts generated by network hardware and applications. SIEM is sold as software, appliances or managed services, and are also used to log security data and generate reports for compliance purposes.

SIEM Capabilities: Data aggregation :

Log management aggregates data from many sources, including network, security, servers, databases, applications, providing the ability to consolidate monitored data to help avoid missing crucial events.

Correlation :

looks for common attributes, and links events together into meaningful bundles. This technology provides the ability to perform a variety of correlation techniques to integrate different sources, in order to turn data into useful information. Correlation is typically a function of the Security Event Management portion of a full SIEM solution

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Alerting:

the automated analysis of correlated events and production of alerts, to notify recipients of immediate issues. Alerting can be to a dashboard, or sent via third party channels such as email.

Dashboards:

Tools can take event data and turn it into informational charts to assist in seeing patterns, or identifying activity that is not forming a standard pattern.

Compliance:

Applications can be employed to automate the gathering of compliance data, producing reports that adapt to existing security, governance and auditing processes.

Retention:

employing long-term storage of historical data to facilitate correlation of data over time, and to provide the retention necessary for compliance requirements. Long term log data retention is critical in forensic investigations as it is unlikely that discovery of a network breach will be at the time of the breach occurring. Forensic analysis: The ability to search across logs on different nodes and time periods based on specific criteria. This mitigates having to aggregate log information in your head or having to search through thousands and thousands of logs.

Proposed system:

In the proposed system, data will be normalized by removing any noise or irrelevant data. Then a dynamic method is discussed to observe any incidents in anomalous host behavior.

Step 1: Data Normalization:

Data normalization is the process of reducing data to its canonical form. For instance, Database normalization is the process of organizing the fields and tables of a relational database to minimize redundancy and dependency. In the field of software security, a common vulnerability is unchecked malicious input.

The mitigation for this problem is proper input validation. Before input validation may be performed, the input must be normalized, i.e., eliminating encoding (for in- stance HTML encoding) and reducing the input data to a single common character set.

Timestamp Normalization:

An enterprise that runs on a global network, produce critical logs by many different devices which are located at various geographical locations. The timestamps made by these devices may differ widely based on their local time zones, UTC, or other time representations. The proposed approach addresses this issue by making use of a central SIEM system for log management that records the time at which the log was received by the SIEM and hence tags each log entry with its own timestamp tsiem. First, for each log entry i, time difference is computed for each device that sends logs to the SIEM system $\Delta i = tsiem,i - tdevice,i$. Then for the device, timestamp correction value is determined by setting i to the value Δi. The normalized timestamp value is achieved by applying this correction value to each device timestamp $tnormalized,i = tdevice,i + \Delta correction$.

IP address-to-Host Mapping:

The problem of varying IP addresses of hosts through DHCP, Beehive analyzes the DHCP server logs collected in the SIEM system and constructs a database of IP-to-host mappings over time. Each binding is represented as a tuple {IP address, hostname, MAC address, start-time, end-time} which maps an IP address to a host in a specific time interval. To update existing bindings, this algorithm is run daily as new DHCP logs become available.

Static IP Address Detection:

Beehive also adopts a method for automatically identifying hosts with static IP addresses. We first retrieve all IP addresses and represent it by set A. Next, IP addresses are retrieved from logs that must only contain hosts with dynamic IP addresses, such as DHCP and VPN logs, and is denoted by D. We then compute the set difference $S = A - D$ which will now contains static IP addresses. Periodically, this procedure is repeated, as new logs become available to harvest new IP addresses and update the sets A, D and S accordingly. But, we

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

resolve IP addresses in S to their host names and compare the names to the previously stored values with each iteration. We conclude that the given IP address is not statically assigned if the host names changed between two iterations, and we remove it from the set S. In this way, with each iteration we refine the pool of potentially static IP addresses. If Beehive fails to find a corresponding binding for an IP-to-host lookup, but instead finds the given address in S, we treat that IP address as a static host and use the host name found in S.

Step 2: Feature Extraction:

In machine learning, pattern recognition and in image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative, non redundant, facilitating the subsequent learning and generalization steps, in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction.When the input data to an algorithm is too large to be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meters, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named features vector).

This process is called feature extraction. The extracted features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced Feature extraction involves reducing the amount of resources required to describe a large set of data. When performing analysis of complex data one of the major problems stems from the number of variables involved. representation instead of the complete initial data.Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which overfits the training sample and generalizes poorly to new samples.

Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy.For each dedicated host in the enterprise, daily a feature vector that includes the 15 features is generated . The features can be grouped into four categories: Destination-Based Features:

New Destinations - The number of new external destinations contacted by each host per day is termed as destination-based feature. We build a history of external destinations contacted by internal hosts over time. If on a particular day if a destination has never been contacted by hosts in the enterprise within the observation period, the destination is considered as new destination. A number of data reduction techniques are employed, to make Beehive scalable, which includes filtering, custom whitelisting and domain "folding." First the "popular" destinations are filtered by creating a custom whitelist, where "popularity" is defined over hosts in the enterprise. The whitelist includes external destinations (both domains and IP subnets) whose number of interacting internal hosts over time (i.e., a training period of one week) exceeds a threshold.

New destinations without whitelisted referrer - Second destination-based feature counts the number of "new" destinations contacted by a host without a whitelisted HTTP referrer.

Unpopular IP destinations - counts the number of destinations contacted by a host that are both unpopular (not on the custom whitelist described above) and are IP addresses.

Host-Based Features - We build a history of UA strings per host over a month- long period, during which every UA string observed from the host is stored. Afterwards, a UA string is considered "new" if it is sufficiently distinct from all UA strings in that host's history.

Policy-Based Features - Blocked domains (and connections) are a coarse indicator of host misbehavior. Upon visiting an unknown destination, the user must explicitly agree to adhere to the company's policies before being allowed to proceed. For a host, we count the number of domains (and connections) contacted by the host that are blocked, challenged, or consented.

Traffic-Based Features - Traffic-based features attempt to capture interesting activities by the amount of time when a host is generating abnormally high volumes of traffic.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Step 3: Clustering:

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them.A clustering algorithm is applied to the projected vectors after PCA. Our algorithm is an adaptation of the K-means clustering algorithm, but this does not require the number of clusters to be specified in advance.

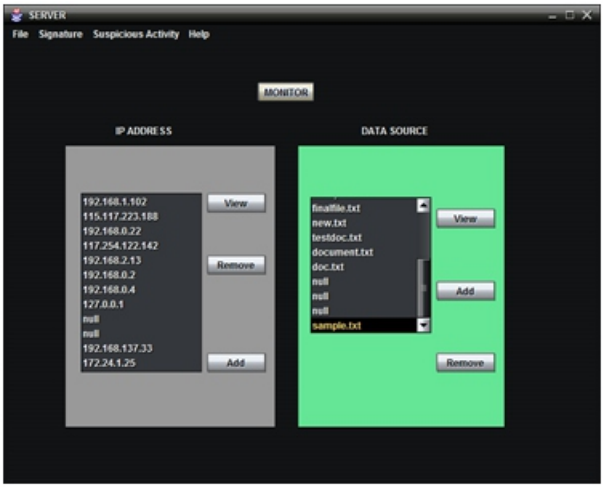
1. Randomly, a vector is selected as the first cluster hub. All the vectors are assigned to this cluster.
2. A vector which is farthest away from its hub is selected as a new hub. Every vector is reassigned to the cluster with the closest hub.
3. Step 2 is repeated until no vector is further away from its hub than half of the average hub-to-hub distance. Vectors are compared via L1 distance, i.e., for vectors v1 and v2, their distance is L1Dist(v1,v2) = summation|v1[i]-v2[i]|. The clustering algorithm is used and applied daily on feature vectors in the enterprise. Incidents are generated for the hosts and they are reported to the security analyst.

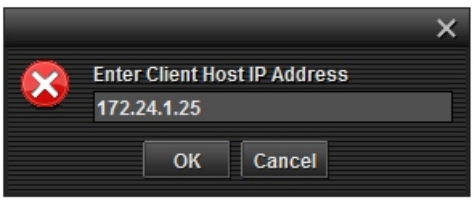
Step 4: Evaluation:

Evaluation is a systematic determination of a subject's merit, worth and significance, using criteria governed by a set of standards. It can assist an organization, program, project or any other intervention or initiative to assess any aim, realisable concept/proposal, or any alternative, to help in decision-making; or to ascertain the degree of achievement or value in regard to the aim and objectives and results of any such action that has been completed. The primary purpose of evaluation, in addition to gaining insight into prior or existing initiatives, is to enable reflection and assist in the identification of future change.

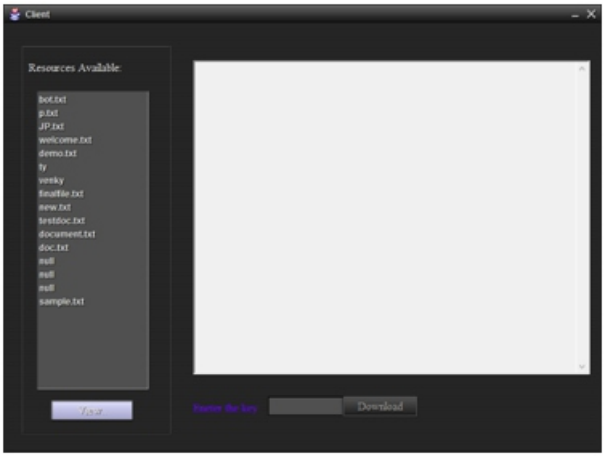
Conclusion:

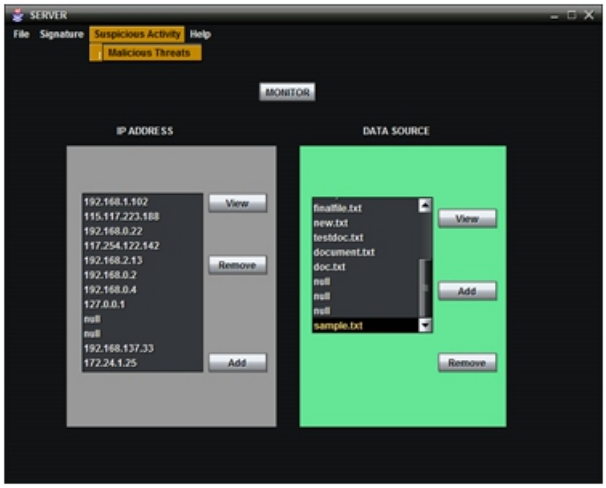
In this paper we proposed a system that addresses the problem of dynamically mining and extracting information from the large log data produced by an extensive variety of security products in a large Network.Big data is changing the landscape of security technologies for network monitoring, SIEM, and forensics. However, in the eternal arms race of attack and defense, big data is not a panacea, and security researchers must keep exploring novel ways to contain sophisticated attackers. Big data can also create a world where maintaining control over the revelation of our personal information is constantly challenged.Therefore, we need to increase our efforts to educate a new generation of computer scientists and engineers on the value of privacy and work with them to develop the tools for designing big data systems that follow commonly agreed privacy guidelines.

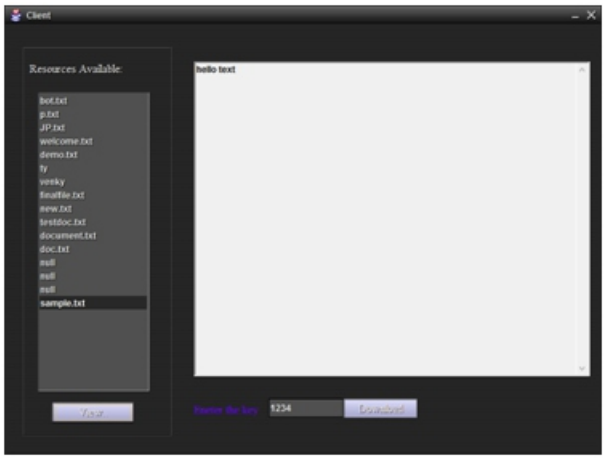


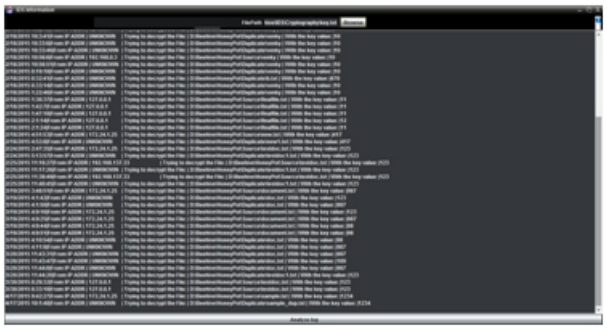


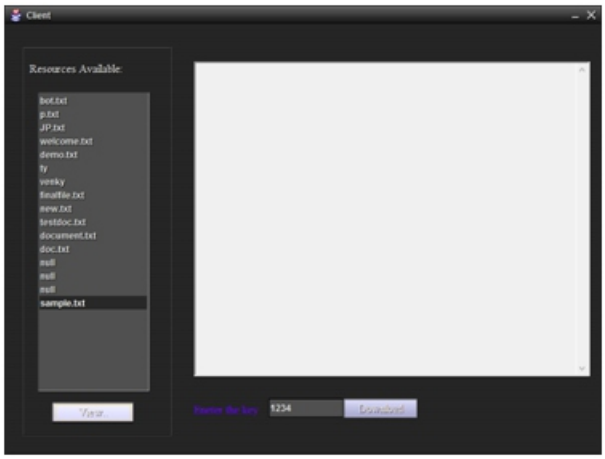
ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal











References:

1.T.-F. Yen et al., "Beehive: Large-Scale Log Analysis for Detecting Suspicious Activity in Enterprise Networks," to be published in Proc. Ann. Computer Security Applications Conference (ACSAC 13), ACM, Dec. 2013.

2.A. Rowstron et al. Nobody ever got fired for using hadoop on a cluster. In HotCDP, 2012.

3.E. Chickowski, "A Case Study in Security Big Data Analysis," Dark Reading, 9 Mar. 2012.

4.E. Ryvkina et al. Revision processing in a stream processing engine: A high-level design. In ICDE, 2006.

5.J. François et al., "BotCloud: Detecting Botnets Using MapReduce," Proc. Workshop Information Forensics and Security, IEEE, 2011,pp. 1–6.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

6.N. Pansare, V. R. Borkar, C. Jermaine, and T. Condie. Online aggregation for large mapreduce jobs. PVLDB, 2011.

7.P. Upadhyaya, Y. Kwon, and M. Balazinska. A latency andfault-tolerance optimizer for online parallel query plans. InSIGMOD, 2011.

8.P. Giura and W. Wang, "Using Large Scale Distributed Computing to Unveil Advanced Persistent Threats," Science J., vol. 1, no. 3, 2012, pp. 93–105.

9.T. Dumitras and D. Shou, "Toward a Standard Benchmark for Computer Security Research: The Worldwide Intelligence Network Environment (WINE)," Proc. EuroSys BADGERS Workshop, ACM, 2011, pp. 89–96.

10.T. White. Hadoop: The Definitive Guide. 2009.

11.V. Raman, B. Raman, and J. M. Hellerstein. Online dy-namic reordering for interactive data processing. VLDB'99.