Big Data Analysis for Implementation of Enterprise Data Security

G Geethakumari
Dept.of Computer Science
BITS Pilani, Hyderabad Campus
Hyderabad, Andhrapradesh, India
geethamaruvada@gmail.com

Agrima Srivatsava
Dept.of Computer Science
BITS Pilani, Hyderabad Campus
Hyderabad, Andhrapradesh, India
Agrimasrivastava1@gmail.com

***Abstract* - In an Enterprise which can be retail, finance, health or manufacturing, there is a high volume of data to be analyzed. The data arrives for analysis at high velocity and there is a lot of data variety. Such a huge data is termed as the "Big Data". This Big Data when analyzed properly can play a significant role in securing the Enterprise data. Analyzing such huge quantum of data and then reducing it, identifying some patterns out of it can help the information security manager monitor even the slightest details of the activities going on in the enterprise. Making use of data mining and applying efficient algorithms of machine learning the security of an enterprise can be very well taken care of. In this paper we propose to investigate the impact of Big Data techniques when applied in the field of Enterprise data Security. We also propose to develop the analysis and design techniques to mitigate the security threats so as to secure the Enterprise data more efficiently.**

Keywords-Information Security, Enterprise Data, Big Data Analysis, Machine Learning.

I. INTRODUCTION

With the advancement of technology enterprises expanded. With this advancement the enterprise started producing more and more data. Managing such huge data became an important and challenging job. The focus of data processing moved from computer center to the terminals in individual offices and homes thereby making the job of the security manager increasingly difficult.

To keep track of all the ongoing events in an enterprise, the information security manager should analyze each and every data as every data set carries certain value [1]. This data can be processed and converted to some meaningful information. It involves a lot of expense in collecting information. It is always necessary to collect the right information. Data should be reliable and collecting reliable information is again a big and complex task in itself.

Hackers and people with malicious intentions are the biggest threats for the enterprise information [2][5]. Internal threats are more about people and processes. Partners in business, employees and consultants who have the access to an enterprise environment can pose potential threat to its security [2][5]. More importantly, detection and deep risk analysis suffer from its inability to collect each and every data and the unavailability of efficient processing platform for this Big Data. The aim here is to make use of the Big Data techniques to analyze the Enterprise data and apply the same to implement enhanced data security mechanisms.

II. RELEVANCE OF INFORMATION SECURITY

Information Security has a very important role to play for any organization [4][5]. Enterprises are spending huge amount of money just to improve upon their security. May it be a public or a private sector, a well-managed organization's security helps in developing a good structure for an organization. Nowadays security is one of the major challenges for organizations. Most of the organizations have already achieved success in implementing security at an appreciable level but still many are lagging behind. Making the information secure, preventing intrusions and stopping secret information disclosure is what every organization wants to go for. Considering all these facts the main challenge is to implement information security at a correct level so that it can address the appropriate threats with utmost efficiency.

III. THE MAJOR CONCERN – ENTERPRISE DATA SECURITY

Enterprise data security forms the basis of information security [2]. The way data is shared, read, modified and most importantly controlled is the key point of consideration. For each and every piece of information, the information security manager must establish and maintain a security program which must ensure three things: Confidentiality, Integrity and Availability [4].

Confidentiality is the protection of information in the system so that unauthorized persons cannot get an access to the enterprise data. User identification and authentication are the major aspects of confidentiality. Who gets to access what information also needs to be considered here. Hackers and unauthorized user activity, infected files when downloaded and virus attacks are some of the most commonly encountered threats to information confidentiality.

Integrity ensures that the data should not be modified unauthorized at any point of time. Whereas availability guarantees that the information is available to all the authorized users as and when required.

IV. BIG DATA – THE NEXT BIG THING

Voluminous amount of data produced by an enterprise that can be either unstructured or structured data describes Big Data. Loading Big Data into the relational database to perform analysis takes a lot of time and involves huge costs. When we talk about Big Data it doesn't describe any specific quantity but generally this term is used for petabytes and exabytes of data. In order to analyze Big Data we need to discover repeatable patterns. [6]

Big data has certain characteristics with which we can separate it from the normal data. Big data has Volume, Velocity, Variety and Value.

Volume- Machine generated data in quantity is much larger than the data generated by people. For eg: a single jet engine can generate 10TB of data in 30 minutes. With more than 25,000 airline flights per day, the daily volume of just this single data source can go up to Peta bytes.

Smart meters and heavy industrial equipment like oil refineries and drilling rigs generate similar data volumes.

Velocity– Social media streams are a good example of explaining velocity of data. Large amount of Data is generated at a very high velocity. Even at 140 characters per tweet, the high velocity (or frequency) of Twitter data ensures large volumes (over 8 TB per day).This type of data comes to the Data Management System rapidly and requires quick analysis or decision making.

Variety-The data is not organized into simple, regular patterns as in a table rather the text, images etc are organized in highly varied structures that have to be analyzed for a deep insight of the information contained in them.

Value- Every data is important and carries Value. Good information may be hidden in unstructured non-traditional data. The challenge is identifying what is valuable and then transforming and extracting that data for analysis [6].

V. BIG DATA ANALYSIS

Basically Data can be broadly categorized into two sets - Data at Rest and Data in Motion - Data at Rest includes web logs, emails, social media etc i.e., collection of that type of data that has been streamed. Data at Motion includes stock market data, data collected by Sensors etc. Machine logs, RFID readers, sensor networks, vehicle GPS traces and retail transactions all contribute to the growing Big Data.

Also if the Data is in bulk it usually beats algorithms. For many years, enterprises have been making business decisions based on transactional data stored in relational databases. A lot of useful information can also be achieved using less structured data like the weblogs, social media, email, sensors and photographs that can be mined for useful information. The cost of data storage and the low compute power have made the task of data collection very easy which would otherwise have been of no use. As a result, a lot of companies are looking forward to include the non-traditional yet potentially very valuable data with their traditional enterprise data in their business intelligence analysis [6].

VI. BIG DATA ANALYSIS TECHNIQUES

Big Data can be analyzed using map reduce in implementation of basic methods like chaining, clustering, machine learning algorithms, time series analytics, text analytics, semantic analytics, latent sentiment analysis.

The data does not always move during the organization phase, so analysis of the Big Data may also be done in a distributed environment, where some data will stay where it was originally stored and be transparently accessed from a data warehouse. The infrastructure required for analyzing big data must be able to support deeper analytics such as statistical analysis and data mining, on a wider variety of data types stored in diverse systems and should be able to integrate analysis on the combination of big data and traditional enterprise data. New insight comes not just from analyzing new data, but from analyzing it within the context of the old to provide new perspectives on old problems [6].

VII. BIG DATA APPLICATION AREAS

There are various places where one can effectively make use of the Big Data technology, some of them are:

E-Commerce and Consumer Marketing - Storing, managing and acting on the sentiment expressed by consumers can be an example of Big Data problem. In a day or a week, consumers can have millions of electronic interactions concerning a brand. Some interactions can be private whereas, much of the interaction may be directly with the company via its websites, call centers, and stores; via email to the company; or, via open social media, such as Twitter or Facebook. When something important is happening – for example, consumers reacting to an economic, safety or environmental incident or to changes in price, product, supply or competition the company needs to quickly understand how best to react to the situation. A slow reaction can greatly increase the cost or difficulty of solving a problem or can cause an important opportunity to be missed [6].

Manufacturing –

Sensors and data capture devices in consumer and industrial products in engineering and testing processes and, in manufacturing processes now collectively generate truly enormous volumes of data. There are hundreds of sensors in every car manufactured today, each generating data many times each second about some component or system of the vehicle when it is in use (e.g, brakes, accelerator, steering, transmission, lubrication, cooling, engine). This data is stored in the car; captured when the car is in for maintenance; and, then analyzed to diagnose immediate problems, but also for insight into engineering, manufacturing and maintenance issues.

Healthcare -The ultimate application for intelligent sensors and a source for both a great volume of data and great value is the human body. There are various medical devices that can increase and sustain the human health, safety, and mobility. Intelligent electronic devices some used by people at home and some that travel with them as they go about their day also capture and transmit data for analysis in managing chronic

diseases and conditions; for dealing with sleep disorders; monitoring exercise; and, for a rapidly growing array of health and medical uses.

In general, more frequent data about what is actually happening with the heart, the breathing process, the blood sugar or the blood pressure as the patient goes about daily life greatly increases the ability to make good clinical decisions [6].

VIII. BIG DATA APPROACH TO ENTERPRISE SECURITY

Big data is like traditional data in many ways. It must be captured, stored, organized, and analyzed, and the results of the analysis need to be integrated into established processes and influence how the business operates. But because big data comes from relatively new types of data sources that previously were not mined for insight, companies are not accustomed to collecting information from these sources, nor are they used to dealing with such large volumes of unstructured data. Therefore, much of the information available to enterprises is not captured or stored for long-term analysis and opportunities for gaining insight are missed.

Because of the huge data volumes, many companies do not keep their Big data, and thus do not realize any value from their Big data. Despite this fact, an astounding number of organizations do not make the most effective use of the information they already collect because security and operations management tools can generate enormous amounts of data. This suggests the growing need for techniques such as data mining to improve the usability of this information.

Organizations would benefit greatly from making this information more actionable in determining priorities for risk mitigation – particularly when managing IT risk at scale.

Emerging security trends can be identified through unsupervised learning and Bayesian inference to textual data. These algorithms can be parallely executed using hadoop/mapreduce, hence reducing the time complexity involved. The trend is centered on a large and growing volume of information – available from a myriad of sources – that can help assure more precise identification, more finely-grained authorization, and more accurate defense against identity and access fraud and abuse [6].

IX. THE ENTERPRISE DATA SECURITY PROBLEM

Data security is indispensable to any enterprise. Enterprise data is vulnerable to leaks by internal stake holders. We are trying to propose a risk assessment mechanism that will help the information security manager to get aware of the security threats and assessing the vulnerability in the end client device.

Many of the legacy security approaches certainly do not work the way we expect them to. Especially, as the data size increases, data variety and number of sources of data increases. The current approaches struggle with large volumes of data, variety of data sources and analysis tools which are unable to keep up with data velocity.

Big data techniques can take traditional security analytics to the next level. Each and every data is important and has some value attached to it. We need ways to better identify meaningful, actionable insight to process all this information. It is important to identify trends, detect new threats based on suspicious activity, identify geographic regions seeing similar disproportionate threat activities, global threat characteristics etc.

X. PROPOSED SOLUTION

Given a problem of securing the vital data of any enterprise, our main focus is to understand the various operations carried out in the system, status of the system, monitoring the typical behavior of the system and last but not the least reviewing its security [3].

The managers and information security officers in business or organizations have to understand the assets owned by the organizations including web servers, mail servers, file servers, personal computers, and all software and application service software. Here the data sets will be large, complex and dynamic with a need to capture, manage and process the data. Hence involving Big Data techniques can lead to entirely new set of security capabilities. This discovering process can be regarded as an auditing of enterprise data security. Figure 1 shows the proposed Big Data Analytics Model for implementation of enterprise data security.

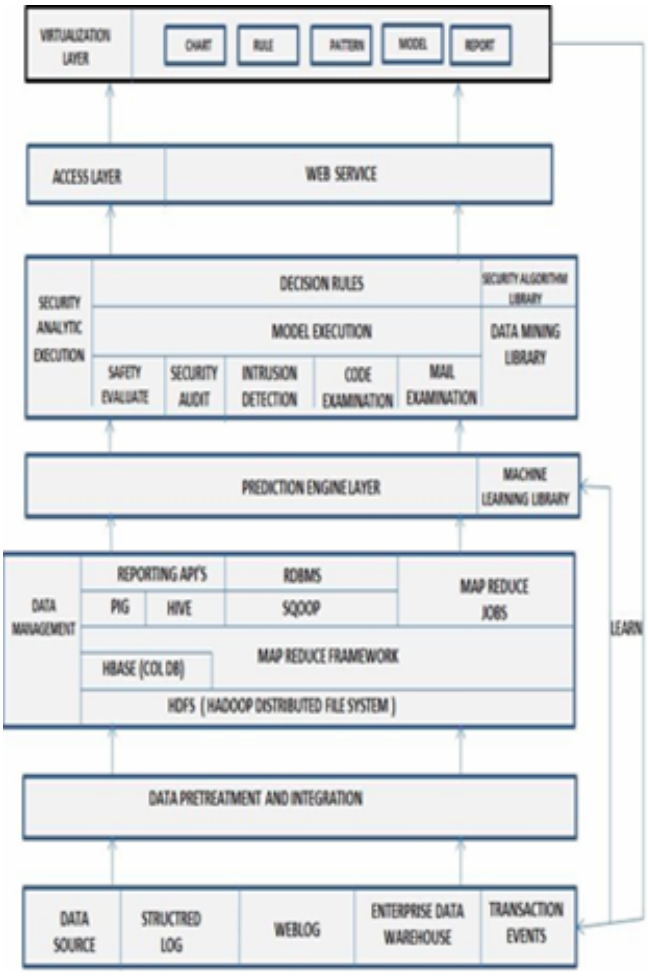

Figure 1. Big Data Analytics Model for Enterprise Security

In the first level of the proposed model, we collect the time series data of the enterprise. Time series data is a sequence of data points measured typically at successive time instants spaced at uniform time instants. Time series prediction or forecasting is based on the analysis of the available data. Initially data from the enterprise can come in any of the forms - it can be structured, unstructured, a weblog, an enterprise data warehouse or a record of transaction events.

The structured data is the one that is well organized and search-able in a structure and is easily identifiable. The unstructured data doesn't have an identifiable structure. Earlier the enterprises used to make use of the traditional databases that had structured data stored in it to analyze and make business related decisions. Unstructured data carries an equal importance as that of the structured data. Nontraditional databases such as a log file of a company, mails, photos etc. can be analyzed and mined as well to get lots and lots of useful information.

The Enterprise data warehouse integrates data from multiple source systems and enables a central view across the enterprise. In the next level of the model, this data undergoes data pretreatment and integration wherein we can classify the data as to whether the data of the enterprise is sensitive or security data.

Level 3 of the proposed model depicts the use of Big Data analysis technique that deals with the data management part of the solution. Using this technique we can manage huge amount of data effectively. When big data is distilled and analyzed in combination with traditional enterprise data, enterprises can develop a more thorough and insightful understanding of their business, which can lead to enhanced productivity, a stronger competitive position and greater innovation. The characteristics derived out of this analysis helps the security administrator of the enterprise to come up with appropriate security policies and implementations.

Big data technique can be applied for any distributed file system but here in specific we propose to use HDFS i.e. the Hadoop Distributed File System. Map Reduce, a part of the Hadoop Framework, is a distributed processing model that provides with an execution environment so that large clusters of commodity machines can run on it.

We propose to use Pig as the data flow language and execution environment for exploring very large datasets. We plan to adopt Hive as the distributed data warehouse. It manages the data that is stored in HDFS and provides a query language based on SQL for querying the data. Hbase, a distributed, column-oriented database of the Hadoop framework is proposed to be used as the required database for the Big Data analysis. HBase uses HDFS for its underlying storage. To efficiently moving data between relational databases and HDFS we plan to use Sqoop, a tool for the same.

Making use of efficient Machine Learning algorithms we can analyze the Big data and observe varying patterns thereby deriving results out of it. The recently emerged DLP (Data Loss Prevention) detection technology, enables organizations to use software that learns to detect the types of confidential data that require protection.

While machine learning as a concept has been around for decades and has been used in everything from anti-spam engines to Google TM algorithms for translating text, it will be applied to DLP content analysis. As a DLP detection technology, Vector Machine Learning learns to recognize sensitive information that must be protected using security algorithms.

This approach continuously improves the accuracy and reliability of finding sensitive information. Applying machine learning to DLP and using Vector Machine Learning (VML) technique, we can quickly and efficiently protect IP and confidential information among increasing amounts of unstructured data.

At the prediction layer engine Machine Learning concepts like Vector Machine Learning and DLP will be applied to the refined data obtained [7]. By predicting any threats to enterprise data security we can improve on its security up to a greater extent.

Using this technique and implementing highly efficient security algorithms we can build upon an enterprise whose data will be well managed and highly secure. It is non preventive in approach but we can also have predefined solutions for the encountered threats thereby converting it as a preventive model to handle enterprise data security.

In the security analytic execution the safety evaluation, security audit, intrusion detection, code examination and mail examination is performed on the data that comes out of the Prediction Engine Layer using the data mining library.

The Web Service in the Access Layer would support the machine to machine interaction over a network. At the Virtualization Layer using the charts, model, pattern etc we can graphically present the result of our analysis on the data effectively to the user.

XI. CONCLUSION AND FUTURE WORK

Enterprise data security is a challenging task to implement and calls for strong support in terms of security policy formulation and mechanisms. In this paper we propose the use of Big data techniques for analyzing the enterprise data and applying the analysis results for better implementation for data security in enterprises. We plan to leverage on the Big data characteristics of large volumes of enterprise data and apply machine learning algorithms to prevent unauthorized access and modification of enterprise data. We have proposed a broad model for the same.

In future we plan to take up the data collection, pretreatment, integration, map reduce and prediction using machine learning techniques. We then would use the results for securing as well as preventing future threats to the enterprise data.

ACKNOWLEDGMENT

Our sincere thanks to BITS-Pilani, Hyderabad Campus for providing us the research environment.

REFERENCES

[1] Wang Cheng, Zeng Min, Liu qiong-mei. Practices of Agile Manufacturing Enterprise Data Security and Software protection. 2nd International Conference on Industrial Mechatronics and Automation, 2010.

[2] Wenguang Chai . Analyzes and Solves the Top Enterprise Network Data Security Issues with the Web Data Mining Technology. 2009 First International Workshop on Database Technology and Applications, 2009.

[3] Goutam Chakraborty, Hiromitsu Watanabe and Basabi Chakraborty. Prediction in Dynamic System - A Divide and Conquer Approach. IEEE Mid-Summer Workshop on Soft Computing Methods in Industrial Applications, 2005.

[4] M.M. Anwar, M.F. Zafar , Z. Ahmed. A Proposed Preventive Information Security System. IEEE International Conference on Electrical Engineering, April, 2007.

[5] Li Xuemei, Li Yan2, Ding Lixing . Study on Information Security of Industry Management. Asia-Pacific Conference on Information Processing, 2009.

[6] Oracle: Oracle: Big data For enterprise. An Oracle White Paper, January 2012

[7] White Paper: Data Loss Prevention Machine Learning Sets New Standard for Data Loss Prevention, 2012.

AUTHORS PROFILE

Dr. G. Geethakumari (corresponding author) is presently Asst. Professor, Dept. of Computer Science and Information Systems at BITS-Pilani, Hyderabad Campus. Prior to joining

BITS, she was a faculty member at CSE Dept, NIT-Warangal. Her research interests include: grid computing, information security, access control modelling, mobile application security, parallel computing, cloud computing, and cloud security. She has many international publications to her credit. She has been a Program Committee member for many international conferences and is also in the book review panel of publishers such as TMH and Springer. Dr. Geetha has been in the forefront of technical activities at BITS-Pilani, Hyderabad Campus. She has held various positions such as Faculty Advisor, Computer Science Association in BITS-Pilani, Hyderabad Campus.

Ms. Agrima Srivatsava is currently a research scholar with the Department of Computer Sciences at BITS Pilani, Hyderabad Campus. Her research interests include Information Security, Big Data and Machine Learning.