RSA Silver Tail – A Novel Approach on Big Data Analytics with cascading Machine Learning Technique

Dr.K.Vaishali, Dr G Swamy, Kaumudi Keerthana, Dr.S.Pothalaiah

Professor Department of CSE, Jyothismathi Institute of Technology and Science, Karimnagar, India

Vaishali5599@gmail.com

Associate Professor Department of CSE, Vignana Bharathi Institute of Technolgy, Hyderabad, India

swamygachikanti2010@gmail.com

Assistant Professor Department of CSE, Vignana Bharathi Institute Of Technology, Hyderabad, India

Keerthana.kommera@vbithyd.ac.in

Professor Department of ECE Vignana Bharathi Institute Of Technology, Hyderabad, India

pothalaiahs@gmail.com

Abstract – The challenge issue of Big data in security point of view to protect data against malicious users. Big data contains huge amounts of personal identifiable information stored data. The breaches affecting big data can have devastating consequences than potential affect. Large number of people are with consequences not only from reputational point of view but with enormous legal repercussions. In order to meet end-user demands, organizations continue to extend applications. At the same time hackers continue to proliferate an evolve leveraging anomaly to gain attacks on corporate data. Encryption is one of the technique to protect from hackers but cannot perform operations on Big Data. A solution for this is to use "Machine Learning Techniques" which allows stored data in the cloud to perform operations, prevents the intrusion and continues with the normal activity. One of the technique is RSA Silver Tail which protects the data in transit and ensure its confidentiality. In real time it monitors threat intelligence to prevent unauthorized access to the data. The aim of this paper is to identify the intrusion on big data operations whether the analytic process is a normal activity or abnormal data, so that the extracted data is secure and more efficient.

Keywords – Big Data Analytics, RSA Silver Tail, Intrusion, Malicious, Cloud Operations.

Introduction I

Big data analytics is a technique to operate on data sets which is really about two things big data and data analytics. Analytics helps to discover what has changed and how one should react. Most of the organizations are confused about selection of right form of analytics for their Big Data analysis. Though it have related experience in data warehousing, reporting and online analytic processing(OLAP), find that business and technical requirements are different for advanced form of analytics. Organizations are implementing specific forms of analytics called advanced analytics, collection of techniques and tools which includes predictive analytics, data mining, statistical analysis and complex SQL, and extends the list to cover data visualization, natural language processing and database capabilities that support analytics such as Map Reduce in-database analytics, in-memory databases, columnar data stores. Big data analytics is enabled by different types of analytic tools that are based on SQL queries, data mining, statistical analysis, fact clustering, data visualization, natural language processing, text analytics, artificial intelligence etc [8]. Big data focus on the size of the data in storage that is data volume along with its other attributes that are data variety, data velocity and data veracity, Thus Big Data View is driven by these four variables.

* Data volume is the primary attribute of big data that define big data in terabytes or sometimes petabytes to show how much data is involved.

* Data Variety defines different types of data such as structured, unstructured and semi structured data that is used by various applications such as credit files, utility records, or even face book posts[9].

* Data Velocity implies fast and frequent processing of data by applications as one can't waste time for massaging or cleansing data that is needed for time sensitive, real time processes.

* Data Veracity focus on accuracy of data that is how far the data is trust worthy for usage by business applications.

The big data scope even affects its quantification ,for example in real world the data collected for general data warehousing differs from data collected specifically for analytics. Different data sets are used for different forms of analytics. Big data analytics can be leveraged to improve information security, employed to analyse financial transactions, log files and network traffic to identify anomalies and suspicious activities and to correlate multiple sources of information into a coherent view. Data driven information security dates back fraud detection. Anomaly based intrusion detection systems[10] is one of the most visible uses for big data analytics in Fraud detection. Custom built infrastructure to mine big data for fraud detection was not economical to adapt for wide-scale fraud detection uses. Network packets and system an event for forensics and intrusion has traditionally been a significant problem.

Database contains hidden information, to extract the information, interrelationship among the data has to be achieved or retrieved from complex data set. Extracting the data from large datasets would be difficult and time consuming. It needs to follow certain protocols and algorithms to classify the data for exact pattern. Due to increase in the amount of data in the field of genomics, meteorology, biology, environment it becomes difficult to find analyze patterns associations within large data. and customer analyze demands predict future possibilities in every aspect to increase innovation retain customers.

SECTION II

2. Related Work: In the context of information visualization shneideman describes a dataset as big when it is too big to fit on a screen at one item per pixel most desktops would stop at a few million data points. More often big data means data that cannot be handled and processed in straight forward manner. A spreadsheet fits in memory it is reasonable quick to determine if the data is clean the values are reasonable and the results can be computed rapidly, big dataset won't fir in memory so it will be hard to check whether it is clean. To understand the challenges of conducting big data analytics in more detail one of our analysts was a social psychologist who works intimately with Twitter, [5] data he receives the raw twitter fire hose dump of data often years' worth of it. Then he analyzes the feed to study trends such as changing sentiment over time or how information spreads through the Twitter feed. Several of the other analysts we interviewed do machine learning over very large datasets for example over all of the search queries coming to search engine, although each analyst workflow varies in specific way analyst activities are generally clustered into five steps acquiring data, choosing an architecture, shaping the data to the architecture writing and editing code, reflecting and iterating on the results. Cloud computation is fundamental different from local computing whether Microsoft Azure Amazon EC2 or a Hadoop cluster parallel computation is differ from a building code, highly parallel systems identical code is run on multiple virtual machines. Users typically get their code running on a single instance that runs locally and then deploy the code to a series of virtual machine that run remotely on a network. The data is stored across multiple servers to ensure that is can be processed as rapidly as possible the code itself contains rules that help decide which machine execute the code and in what sequence. [3] The first challenge identified was determining where the data in their big data systems came from, how they discover sources of data increasingly data is available in a wide variety of sources and formats, online databases of public statistics are provided by the U.S government and the united nations private companies sell data from data market places such as Microsoft Azure marketplace and Info chimps. Whatever the data big data analysis is performed on the platform organizes the computation around a set of programming abstractions substantially different from those of the normal desktop environment, analyst trained on the desktop environment have to learn these new abstractions and plan their computation around them often facing a new set of engineering trade-offs and failure modes. A completely new part of designing data analysis for the cloud is planning for the economic impact of the design choices, with cloud computing nearly every choice about computation uploading downloading data and storage has a direct dollar cost. Planning and monitoring these costs is unfamiliar and poorly supported for end users and making mistakes can be quite expensive, many of these decisions need to be made before the first byte is uploaded to the cloud and before the first line of code is written.

The potential benefits of big data are real significant and some initial success have already been achieved remain many technical challenges that must be addressed to fully realize this potential. Industry analysis companies like to point out that there are challenges not just in volume but also in variety usually mean heterogeneity of data types representation and semantic interpretation. Many people unfortunately focus just on the analysis modelling phase while that phase is crucial it is little use without the other phases of the data analysis pipeline, even in the analysis phase which has received much attention there are poorly understood complexities in the context of multi-tenanted clusters where several users programs run concurrently. [4] Many significant challenges extend beyond the analysis for example big data has to be managed in context which may be noisy heterogeneous and not include an upfront model, doing so raises the need to track provenance and to handle uncertainty and error topics that are crucial to success and yet rarely mentioned in te same as big data.

SECTION III

3. Problem Definition: The ability to extract analyzes and correlate potentially sensitive data sets, data used for analytics may include regulated information or intellectual property. System architects must ensure that the data is protected and used only according to regulations, however benefits of big data analytics tool extract and utilize the data violations of privacy easier and prevent abuse. Our research is to identify intrusion using Machine Learning Technique on Big Data that improves security and efforts of resource, specific work groups on big data such as confidentiality integrity.

Big data security analytics is flexible query capabilities should blend data processing power and custom rules for strong accuracy includes machine learning algorithms like 21CT, LogRhythm, silver Tail and behaviour anomaly detection security Lancope Netskope Solera networks etc. organizations also use Splunk as a foundation for custom algorithms visualization for security remains extremely emerging area today but there is an increasing amount of research and development happening primarily in places like U.S national labs institute. When malware targets an unpatched system it's an emergency, when malware is headed for a patched system, the situation isn't very critical time, big data security analytics will blend threat detection/forensics with continuous monitoring to calculate risk scores associated with cyber-attacks. McAfee will push this agenda by integrating McAfee Security Manager (i.e. Nitro) with ePO. RSA will do the same by bridging its big data security analytics and Archer. HP will also follow this path. Nevertheless, security automation is a growing requirement as the security staff can no longer keep up, Cisco will use its network infrastructure, SDN, and cloud-based big data security intelligence for network security automation. Other network security specialists like Check Point and Palo Alto Networks will also pursue this course, IBM will also be aggressive, integrating its network security portfolio (i.e. ISS) and Trusteer (i.e. endpoint security) with QRadar, IBM Security Intelligence with Big Data, and X-force security intelligence. The analytics stage where statistics algorithms simulations and fuzzy matching come into play which analysis technique to apply to know that the activity is normal or abnormal. Trend analysis to build a baseline of normal network behaviour and alert network managers in real-time managers when there is an anomaly. At the end of the analysis process there is usually analyst who needs to make sense of insights surfaced by the analytic engine that has run against the entirely of available data. Network analysts visualizing trends and patterns of the underlying flow of data is extremely powerful for identifying outliers. Network managers aren't the only folks in developers with millions of lines of code and complex multi-tiered applications, apps teams also struggle with managing huge volumes of end-user experience and transaction data and extracting meaningful intelligence.

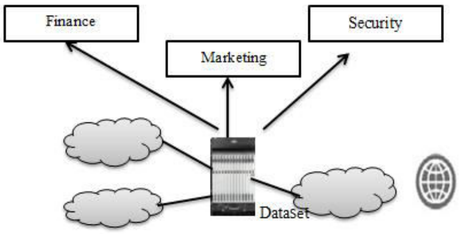


Figure 2 Proposed systems for big data analytics

Machine learning research provides data-driven models for predicting one-dimensional targets such as binary outputs in classification and real valued outputs in regression. at the edge if statistics computer science and emerging application focuses on the development of fast and efficient alogirthm for real-time processing of data with as a main goal to deliver accurate predictions of various kinds of segmentation of customers fraud detection can solve such applications using a set of generic methods that differ from more traditional statistical techniques.

*Machine Learning Algorithms Using R :*Machine learning is to improve the learning that it becomes automatic task of learning or predictive modelling for finding predictive patterns. Typical applications of machine learning can be classified in to scientific knowledge discovery ranging from the anti-spam filtering systems.

The KNN is one of the machine learning algorithm instance based learning where new data is identifies based on stored labeled instances, the distance between stored data and the ne instance is calculated by means of similarity measure expressed by a distance measure such as the Euclidean distance cosine similarity or the Manhattan distance. The data is calculated for any point of input into the system similarity value to perform predictive modelling is either classification or assigning a label new instance or regression. The k-nearest neighbour algorithm adds to the basic algorithm that after distance of the new point to all stored data points have been calculated the distance values are sorted and the k-nearest neighbour are determined, the neighbours are gathered and a majority used for classification or regression purposes.

SECTION IV

4.1. Linear Regression: Linear regression estimate value is number of calls total sales based on [6] continuous variables establish relationship between independent and dependent variables by fitting regression line and represented by linear equation Y = aXb. to understand linear regression is to relive the childhood by increasing order of weight without asking them their weight what do you think the child will do at the height and build of people and arrange them using a combination of these visible parameters. For above equation Y is Dependent variable a Slope X Independent variable b Intercept.

The sum of squared difference of distance between data points and regression line, linear regression is simple linear regression is characterized by one independent variable and multiple linear regressions is characterized by multiple independent variables finding the best fit polynomial regression curvilinear regression.

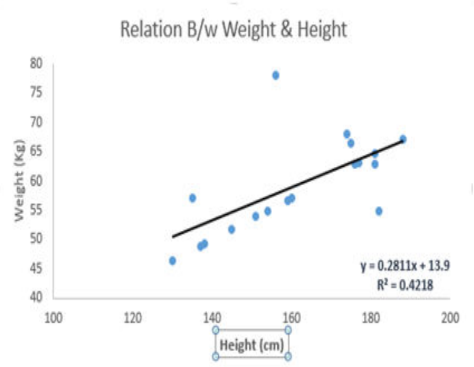


Figure 3 Linear Regression

4.2. Logistic Regression: Logistic is a classification not an regression algorithm used to estimate discrete values like binary values 0/1 yes/no true/false etc. it predcts the probability of occurrence of an event by fitting data to a logit function as logit regression output values lies between o and 1.

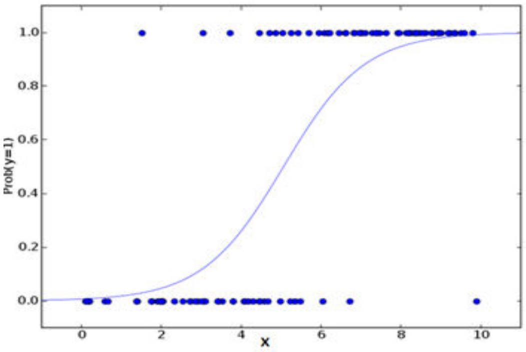
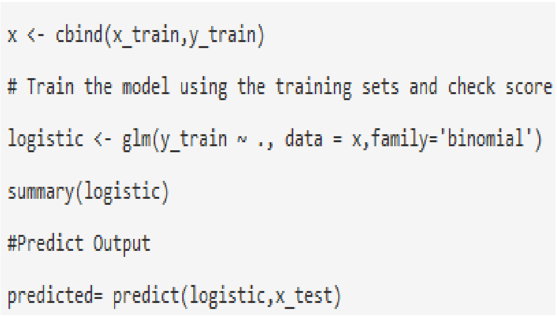


Figure 4 Graph of Logistic Regression

In the graph p is the probability of presence of the characteristic of interest parameter that maximize the likelihood of observing the sample values rather than that minimize the sum of squared errors. For the simple way to replicate a step function can go more details but that will beat the purpose.

R code for Logistics



```
x <- cbind(x_train,y_train)
# Train the model using the training sets and check score
logistic <- glm(y_train ~ ., data = x,family='binomial')
summary(logistic)
#Predict Output
predicted= predict(logistic,x_test)
```

4.3 Decision Tree: Supervised learning algorithm that is mostly used for classification problems, suprisingly categorical and continuous dependent variables split the population into two or more homogeneous sets. The most significant attributes independent variables to make as distinct groups as possible, the population is classified into four different groups based on multiple attributes to identify different hetrogeneous groups uses various techniques like information.

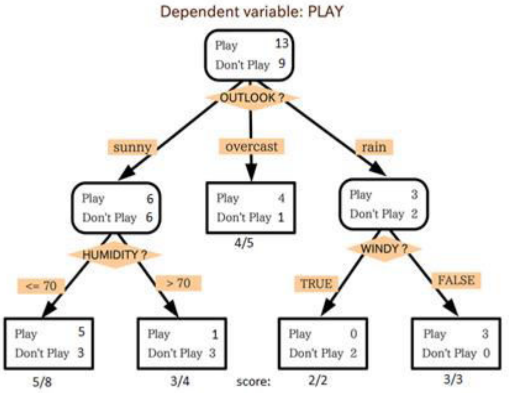
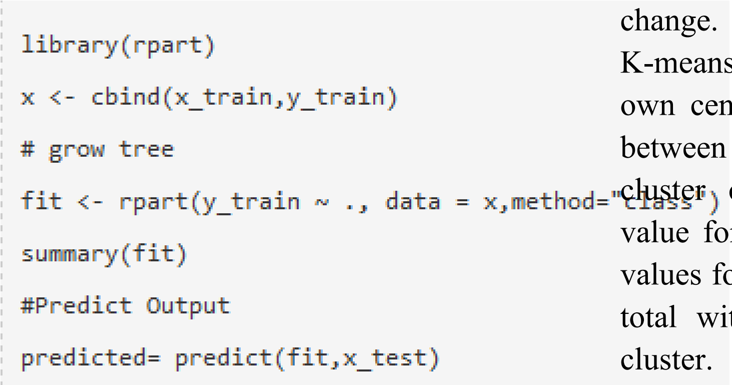


Figure 5 Decision Tree Algorithm

Decision tree is a type of supervised learning algorithm having pre-defined target variable that is mostly used in classification problems, categorical and continous input and output variables. Homogenous sets based on most significant splitter in input variables. [7] Decision tree is easy to understand even for people from non-analytical and statistical knowledge to read and interpret them represents intuitive and users can easily relate the hypothesis and fastest way to identify variables and relation between two or more variables with the help decision trees can create new variables/ features that has better power to predict target variable.

R Code for Decision Tree



```
library(rpart)
x <- cbind(x_train,y_train)
# grow tree
fit <- rpart(y_train ~ ., data = x,method="class")
summary(fit)
#Predict Output
predicted= predict(fit,x_test)
```

4.4. K-Means: Unsupervised algorithm which solves the clustering problem follows a simple and easy way to classify given data set through a certain number of clusters assumes k clusters. Data points inside a cluster are homogeneous and heterogeneous.

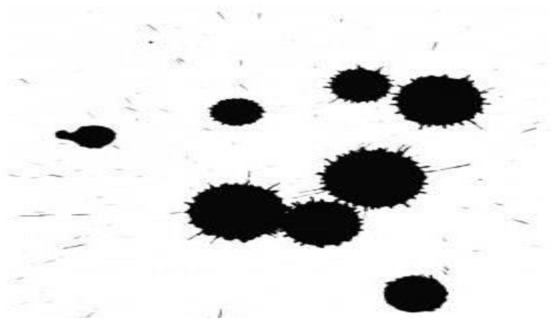


Figure 6 Clustering of K-Means Algorithm

K-means picks k number of points for each cluster known as centroids

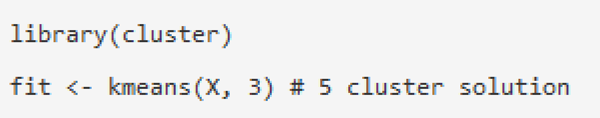
Each data point forms a cluster with the closest centroids that are k clusters.

Finds the centroid of each cluster based on existing cluster members here new centroids.

As we have new centroids repeat step 2 and 3 to fin closest distance for each data point from new centroids and get associated with new k-clusters. Repeat process until convergence occurs that centroids does not change.

K-means is a cluster and each cluster has its own centroids sum of square of difference between centroid and the data points within a cluster constitutes within sum of square value for that cluster. When sum of square values for all the clusters are added becomes total within sum of square value for the cluster.

R Code for K-Means



```
library(cluster)
fit <- kmeans(X, 3) # 5 cluster solution
```

SECTION V

4. The term big data refer large scale information management and analysis technologies that exceed the capability of traditional data processing in amount of data, the rate of data generation and transmission of structured or unstructured data. The process of analysing mining big data can produce operational and business knowledge at an unprecedented scale in storage processing and analysis of big data include that rapidly decreasing cost of storage and CPU power in recent years.

4.1. Security in Big Data Network: In the case study of Zions Bancorporation using Hadoop cluster and business intelligence tools to parse more data more quickly than with traditional SIEM tools, the quantity of data and frequency analysis of events are too much for traditional SIEMs to handle searching among months load of data could take between 20minutes and an hour that get the same results. The security of data warehouse driving this implementation not only enables users to mine meaningful security information from sources of firewalls and security devices but also from website traffic business processes and day-to-day transactions.

4.2. Enterprise Analytics: Enterprise collect terabyte of security relevant data that network events, software application events and people action events for several reasons including the need for regulatory compliance and post forensic analysis. Enterprise can barely store the data for example it is estimated that an enterprise as large as HP generates 1 trillion events per day or roughly 12 million events per second which grows as enterprise enable event logging in more sources more employees that deploy more devices and run more software. Earlier analytics do not work well at the scale and typically produce so many false positives that efficacy is undermined. the effort of HP labs is to move toward scenario that data leads to better analytics and actionable information designed in order to identify actionable security information from large enterprise data sets and drive false positive rates down to manageable. Challenges must be overcome to realize the true potential of big data analysis among the challenges are the legal privacy and technical issues regarding scalable data collection transport storage analysis and visualization.

The challenges group at HP labs has successfully addressed several big data analytics for security challenges which are highlighted first large scale graph inference approach was introduced to identify malware infected hosts in an enterprise network and the malicious domains accessed by the enterprises hosts. A host domain access graph constructed from large enterprise event data sets by adding edges between every host in the enterprise and the domains visited by the host. Truth information from a black list and a white list and belief propagation was used to estimate the likelihood that a host or domain is malicious. HTTP request data set collected at a large enterprise a 1 billion DNS request data set collected at an ISP and a 35 billion network intrusion detection system alert data set collected from over 900 enterprise worldwide showed that high true positive rates and low false positive rates can be achieved with truth information. Terabyte of DNS events consisting of billions of DNS requests and responses collected at an ISP to identify botnets malicious domains and other malicious activities in a network. Malicious fast flux domains tend to last for a short time whereas good domains such as hp.com last much longer and resolve to many geographically distributed IPS. Set of features were computed including one derived from domain names, time stamps and DNS response time-live values then classification technique used to identify infected hosts and malicious domains.

SECTION V

5.1. Misuse of Big Data: Data in the wrong hands whether malicious manipulative or naïve can be downright dangerous goes bad unfortunately the learning game is no stranger to both abuse of data. Data gathering in education stands outs as truly evil in 1939 the CEO of IBM Thomas Watson flew across the Atlantic to meet Hitler results that mechanical equivalent of a learning management system, data was stored as holes in punch cards to record details of people includes misuse data used daily. It was a vital piece of apparatus used in the final solution to execute the very categories stored on the apparently cards, as books are documented with misuse data. Fortunately, the Internet, social media, and other modern technologies are not only a major source of the big data that is sometimes used for bad purposes, but these technologies often also help expose these abuses. Individuals who become privy to damaging information about governmental behavior can publicize that behavior on the

Internet to vast audiences. For example, since information about these events is quickly posted online, The Internet has forced [7] China to admit to events that in the past would have been kept secret, such as governmental land grabs, protests against local pollution levels, local health epidemics, and deaths from mining disasters, How to harness big data to socially valuable purposes while keeping down abuses is becoming an important priority in democratic societies.

5.2. Silver Tail Technique to detect Intrusion in Big Data: RSA silver Tail is a streaming analytics supports click– by-click in memory threat scoring for faster detection of suspicious activity on big data analysis. In real-time silver tail behavioural analysis on hourly basis tracking more than 330,000 clicks per second on some larger sites including mobile web traffic, functionality makes its online fraud detection more applicable to the needs of the market observed Baylor.

Achieving the right balance of security without compromising the user experience is a challenge for organization, RSA is adaptive authentication for organization that want to protect users accessing web sites and online portals browsers secure sockets layer virtual private network application web access management applications. Adaptive authentication is a comprehensive authentication and fraud detection by RSA risk based authentication technology adaptive authentication is designed to measure the risk associated with users login and post login activities by evaluating a variety risk indicators.

Adaptive authentication leverages a series of technologies and components to provide cross-channel protection including the RSA risk engine policy management, device & Behavior profiling, RSA efraud detection.

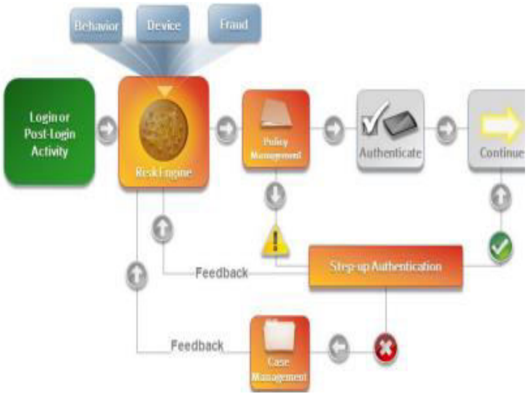


Figure 7 Machine Learning RSA Silver Tail Technique

RSA siler tail is a self learning statistical machine learning technology that utilizes over 100 indicators to evaluate the risk of an acitivity in real-time. Authentication leverages the risk engine to generate a unnnique scooore for each acitivity that ranges from 0 to 1000 where 1000 indicates the greatest level of risk and score is reflective of device profiling behaioral profiling and efraud data. RSA policy management application translates risk policies into decision and actions through the use of comprehensive rules framework the policy management application can be used to set the risk score that will require later review in the case of management application prompt assurance or stp-up authentication deny transactions of fraud is very high.

RSA efraud activity is cross function repository of fraud patterns from RSA extensive application of customer's researcher's and third party contributors across the global. When intrusion elements such as IP address device fingerprint and payee account are identified and shared with the efraud then it provides direct feeds to the risk engine when an activity is attempted from a device or IP that appears in the repository in high risk.

A highly effective fraud management that enables the tracking of activities that rigger policy engine rules and determines if flagged activities are normal or abnormal then organizations use this case management to analyze fraud activities. Case management API is an interface of adaptive authentication capabilities that allows an organization to share information with an external case management system, consolidating the cases into one system provides an organization with the ability to more efficiently confirm and prevent the abnormal activity. RSA silver tail is procedure that validates users identity usually prompted by high risk transactions according to organization policy rules that prevents the malicious attack on big data analytics.

CONCLUSION

Current research of Big data gained lots of interest due to its perceived unprecedene opportunities and advantages, big data analytics can be applied to leverage business enhance decisions. Big data analysis is valuable information can be stored and extracted to support informed decisions, consequently for different areas, where big data analytics can support and it found vast amount of opportunities in various applications such as customer vendor and employee fraud prevention and detection. According to the our research security to the big data business applications to identify malicious attacks by using RSA silver tail machine learning, this gives the user more secure data for big data operation. Believe that big data analytics is great significance in the data storage and management on financial, marketing and sales provide unforeseen insights and benefits of various areas.

REFERENCE

[1] Asur, S., Huberman, B.A.: Predicting the Future with Social Media. In: ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 492–499 (2010)

[2] Bakshi, K.: Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7 (2012)

[3] Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills: New Analysis Practices for Big Data. Proceedings of the ACM VLDB Endowment 2(2), 1481–1492 (2009)

[4] Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104 (2011)

[5] Economist Intelligence Unit: The Deciding Factor: Big Data & Decision Making. In: Cap Gemini Reports, pp. 1–24 (2012)

[6] Bilge, L. & T. Dumitras. (2012, October) Before We Knew It: An empirical study of zero-day attacks in the real world. Paper presented at the ACM Conference on Computer and Communications Security (CCS), Raleigh, NC.

[7] Dumitras, T. & D. Shou. (2011, April). Toward a Standard Benchmark for Computer Security Research: The Worldwide Intelligence Network Enviornment (WINE). Paper presented at the EuroSys BADGERS Workshop, Salzburg, Austria.