IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

Review Paper on Dynamic Mechanisms of Data Leakage Detection and Prevention

Shivakumara T[1*], Rajshekhar M Patil[2], Muneshwara M S[3]

[1]Department of MCA, BMS Institute of Technology and Management, Bengaluru, Karnataka, India
[2]Department of CSE, Guru Nanak Institute of Technology, Hyderabad, India
[3]Department of CSE, BMS Institute of Technology and Management, Bengaluru, Karnataka, India

**Corresponding Author: shivakumarat@bmsit.in, Tel: 9060900986*

DOI: https://doi.org/10.26438/ijcse/v7i2.349358 | Available online at: www.ijcseonline.org

Accepted: 14/Feb/2019, Published: 28/Feb/2019

Abstract- Today's world needs most of the things would be automated, the machine would be able to do itself everything by its own, from initiation of job to till the decision-making capabilities. Similarly the Security of the data needs the automated, self-decision and self-protective proactive mechanisms. Since, the existing tools and mechanisms are semi-automated still it needs the human intervention to update and configure the security layers and features. This paper gives the summarized view of the existing work carried out so far in the area of data security – data leakage detection and prevention.

Keywords- *Data, Detection, Dynamic, Leakage, Prevention, Protection, Sensitive, Self-protective, Self-configurable*

I. INTRODUCTION

This world completely depends on the data will be generated by the humans for their own use. The companies are totally depends on the people who are generating the data. So, the data is a hot currency to the IT world. Due to this the hackers or intruders or the inside employees are trying to gain the monitory benefits, they are trying to get access to the data with or without authentication. In this case, Security plays a major role to protect the data or sensitive information which organizes needs to protect. This is the challenging issues to the organizations must prepare the security strategies and implementing them. In this regard organizations spending huge money rather than the usual business operations. Security became the headache to the companies. Security is not static, but it is in dynamic nature, it keeps updating the mechanism and strategies, because hackers are not static, they keeps trying to find the loop holes, weaknesses and avenues to attack the data centers of the victims. Sensitive data are more important than the just data itself. So, Sensitive data such as a person's credentials like bio-metric information – thumb impression, retina, voice, face. Card information's such as credit card numbers, pins, passwords, date of birth, contact numbers. Health related data such as diseases, treatment manuscripts. Insurance records, financial data etc.

The data protection needs at several levels, when the data is at rest, data is in motion and data is in use. This paper, mainly concentrates on the data is in rest and in use. When the data is in rest, the encryption mechanisms safeguard the data. It encrypts data and stores in the drives with key concepts. Storage technologies improved a lot, but the intentional leakers have the privileges to get access to this data, they can leak it to the outside world. Here is the gap that the dynamic mechanisms needed to fulfill, if someone with proper authentication tries to leak the data, the mechanisms must find and prevent the leakage of data. Here it is the challenging task, since it analysis the lot of the data of that particular leaker, his/her behavior, activities of past data, user policies and authorization. Today lots of tools available to watch the activities of the employee or the third-party, which they uses the organization data, still it required a lot of time to find and analyze the leakage happened, then lots of process involved it to recover the leaked data. But it requires instant leakage detection solutions. It means, completely automated tools required in place of semi-automated or manually controlled security solutions. Either sensitive data resided in the cloud or legacy devices, the security is must. This paper gives the glimpse of summarized work carried of automated solution of the cyber security.
Data is the hot currency to every corner of the world from small-scale to medium scale to large scale enterprise applications. The need of data protection lies in up-front of business world.

The data loss due to device crash, device theft and device destruction may happen with catastrophic failures. Every organization faces this problem, it may occur physically or logically remove or move data from the organization either intentionally or unintentionally. Whereas, the data leakage, is

an incident when the confidentiality of information compromised. It refers to unauthorized transmission of data from within an organization to an external destination.

The data loss / leakage prevention solutions detect and prevent unauthorized attempts to copy or send sensitive data both intentionally and / or unintentionally. The data leakage prevention solutions were designed to detect potential data breach incidents in timely manner and this happens by monitoring data. Data leakage prevention helps the organization to prevent the leakage of sensitive data or information namely, personal identifiable information, financial information, trade secrets, health information / records, aviation and consulting.

Data loss / leakage prevention is a computer security term which will be used to find, watch, and protect data in use, data in motion, and data at rest. The Data Leakage Prevention provides sensitive asset classification, sensitive asset audits, identity and access management audits, applying encryption to sensitive assets, applying enterprise digital rights management privileges to sensitive assets. When the data is in motion or transit the existing tools such as IDS, honey pot, firewalls, encryption, digital certificates, anti-virus tools, compression, authentication and monitoring tools helpful to watch the data. When the data is at rest, the authentication tools, drive-lockers, role-based access controls, a tokenization and access policy plays a role. But, the threat of data breach due to the guilt mind-set of employees in the organizations. Since, these are guilt mind-set, they have authentication access to the data centers, they can get access to and miss-utilize the data, which is more sensitive to specific to organization without any alert of security systems.

Insiders are the main part of the organizations and some of them have access to sensitive or critical information. Insider attack is the most dangerous type of attack damaging the reputation of the institution. According to the statistics reported by InfoWatch Analytical Center, the rate of the data leakage caused by insiders is 72.8% and 43.7% of them is intentional; 49.7% of them is accidental. Besides 43.2% of data leaks occur on network channels. Due to advent of technologies and attack mechanisms, the automated tools exists to leak the organizational information, here manual involvement is less, everything the machine and algorithms to take care of data leakages. Now, the automated, self-defense mechanism needed to protect the data. Further, if the data is able to protect itself from inside or outside attacks, it is far better than any other external tools protection.

The security gap and strategies are entirely independent from the organization to organization. Security itself is not a just one time shot; it is a process of analyzing, identifying, selecting, and applying the suitable tool/s to prevent breaches.

DLP is a new approach for data protection beside the conventional security mechanisms such as antivirus, Firewalls, Network Intrusion Detection Systems (NIDS), etc. DLP systems have ability to find, watch and protect sensitive data according to predefined rules and or policies. In the rest of this review paper, we have discussed the related work and result discussion in the chapter-3, followed by chapter-4 conclusion and future scope of work.

II. RELATED WORK

The author [1] discussed "Self-Protective information capabilities must address at least three fundamental challenges

i. Referencing chain of successive custody, sources & operation.
ii. Incorporating notions of pedigrees & dependencies
iii. Tracking (including distribution and potentially source attribution)

Self-healing & self-protecting capabilities must enable the data itself to keep up key properties & provenance information overtime & at scale.
Self-protective data with the ability to support provenance & chain of evidence is also essential as a basis for information assurance & accountability.
Research challenges & goals: Self-Protecting Data Systems
These will have at least three critical capabilities
i. Attributes
ii. Access &
iii. Protection

Key Attributes include:
i. Origin & history (chain of evidence, transformations etc.,.)
ii. Access such as is necessary to protect private & classified data and must be actively governed by the data in contrast to externally governed data access in today's systems.

A scientific approach to cyber security requires development & application of creative approaches to quantify, process, display, & communicate existing, future, & potential threats. The Complex cyber world poses numerous analysis, challenges that must be addressed to collect manage, store, process, integrate, & understand massive, heterogeneous, distributed cyber data. A signification transformation will be required that makes data self-protecting rather than dependent on external operations, such an approach would render todays cyber security threats irrelevant. Today, mechanisms for tracking the provenance of data throughout the workflow exact only in rudimentary form & in a few large projects – no general purpose system is available.
The critical challenge to information or data integrity, accessibility, confidentiality or trust worthiness is to move

from today's paradigm of "passive data" to "active data". Detect & prevent unauthorized access or use. Recover from damage or manipulation retaining information about the nature of the event & initiator. Present verifiable credentials about its origins & subsequent transformations. Execute "defensive protocols" to find attackers & attack methods. Develop immunities through learning & communicating with peers. The concept of "Self-protection" also involves active self-maintenance of provenance, integrity & chain of evidence or "Self-advocacy".

The following approaches researchers needs their attention Mathematics Predictive Awareness of Secure Systems – real time detection of anomalous activity & adaptive immune systems response. This requires deeper understanding of complex applications & systems, through data-driven modeling, analysis, & simulation of architectures, techniques & processes.

Information: - Self-protective Data & Software – Create "active" data systems & protocols to enable self-protective, self-advocating & self-healing digital objects. This needs tackle the critical problem of data provenance & related research to give information integrity, awareness of attributes such as source, modification trace back, & actors & mechanisms to enforce policy on data confidentiality & access.

Platforms – Trustworthy Systems from untrusted Components – Develop techniques for specifying & maintaining overall trust properties for operating environments & platforms, quantifying & bounding security & protection, integrity, confidentiality, & access in the context of a "system" comprising each component for which these are varying degrees of trust.

Author [2], the survey of self-protecting software systems are a class of automatic systems capable of detecting and mitigating security threats at runtime. Self-protecting software systems allows the system to adapt to the changing environment through autonomic means without much human intervention. This paper proposed a comprehensive taxonomy to classify and characterize research efforts in this arena.

The author, discussed the techniques, CIA model, defense in-depth, and different architecture based on different operations and organizations, intrusion, firewalls. ARM (Architecture Manager) monitors and protects the system by implementing the Monitor, Analyze, Plan, Execute (MAPE) loop for self- adaptation.

The working style of firewall is that it detects predefined policy threshold and so disables the HTTP connection. These ARM & Firewall show self-adaptive and self-protecting behavior at the system level. The meta-level subsystem is part of the software that is responsible for protecting (i.e., securing) the base-level subsystem. The meta-level subsystem would be organized in the form of feedback control loop, such as the MAPE-K architecture. Self-protection of software systems is becoming increasingly important as these systems face increasing external threats from the outside and adopt more dynamic architecture behavior from within. Self-protection, like other self-* properties, allows the system to adapt to the changing environment through autonomic means without much human intervention, and can thereby be responsive, agile, and cost-effective. Existing research has made significant progress towards software self-protection, such as in intrusion tolerance systems and adaptive security mechanisms at the application level. This paper proposes a comprehensive taxonomy to classify and characterize research efforts in this arena. The research has revealed patterns, trends and gaps in the existing literature and underlined key challenges and opportunities that will shape the focus of future research efforts. In particular, the survey shows self-protection research should advance from focusing primarily on the network and host layers to layer-independent and architecture-based approaches; from single mechanism and single-objective to multi-strategy and cost-sensitive decision-making, and from perimeter security to overall system protection. We believe the results of our review will help to advance the much needed research in this area and hope the taxonomy itself will become useful in the development and assessment of new research.

The author [3], proposed Bell-LaPadula security model. This model also called data confidentiality model. Biba-Integrity model which describe rule for the protection of data integrity. author using the concept of Bell-LaPadula Model for providing secured infrastructure; it is a state-machine model and used to apply access control in different environment such as-Military security - Army, Air-force, Navy, NATO, NASA etc. Commercial security Marketing Sales, Research and development, Human Resource department etc. In this model, the author chose, AES-Symmetric of 128-bit encryption and decryption. In addition to detect data leakage, it also protects the different types of active and passive attacks. The proposed work computationally cost-effective in terms of time and pace uses. Also, this can be used in distributed computing environment to protect data from data leakage. As author said that this model would not work for web environment where multiple numbers of users frequently accessing the data object.

The author [4], mainly focused on detecting the leakage of sensitive data over the mobile devices with the help of Labyrinth a run-time privacy enforcement system. Labyrinth supports both Android and iOS. The Labyrinth for commercial usage and enterprise production environments. Also enlighten about Packet Analyzer proxy will

dynamically detect matching of the values communicated between the client and the server with any of the security-sensitive values configured on the application. However, the Packet Analyzer will also correct any vulnerability occurring at run time by obfuscating confidential data, partially or in its entirety, before it is transmitted, in order to prevent leakage of security-sensitive values. The future work is specialized capabilities to handle encrypted communication with unauthorized third parties such as analytics and advertising servers. This work will require usage of statistical learning methods and / or databases of signatures.

Labyrinth features several novel contributions:

(i) it allows for visually configuring, directly atop the application's User Interface (UI), the fields that combine custom sources of private data;

(ii) it does not need operating-system instrumentation, but relies only an application-level instrumentation and on a proxy that intercepts the communication between the mobile device and the back-end servers; and

(iii) It performs an enhanced form of value similarity analysis to detect data leakage even when sensitive data (such as a password) is encoded or hashed. Labyrinth supports both Android and iOS.

The Labyrinth for commercial usage and enterprise production environments. Also enlighten about Packet Analyzer proxy will dynamically detect matching of the values communicated between the client and the server with any of the security-sensitive values configured on the application. However, the Packet Analyzer will also correct any vulnerability occurring at run time by obfuscating confidential data, partially or in its entirety, before it is transmitted, in order to prevent leakage of security-sensitive values.

The author [5], risk analysis of business flow, especially the sensitive data flow. The key elements of an IP data flow map includes: business flow, data types, data processing stakeholder, systems/applications, data transfer mode, data life cycle coverage, etc.

- IP data flow map in foundry can easily be drafted based on the life time of IC design data and its derivatives.
- Data type of IC design data is in the form of Graphic Data System II (GDSII- usually its size is >100 GB) or Open Artwork System Interchange Standard (Oasis).
- Key element 1, business flow: it is the basis of data flow. How to produce IC chips from IC design data is the main business flow in foundry. Then we can draw the data flows and pick up all sensitive data from the flow.

Key element 2, data types: of cause, IC design data is the first identified data type, whose format is usually GDSII or Oasis. The second category is the IC design's derivate data, including OPC data (Optical Proximity Correction), mask writer data (in E-Beam/MEBES format), and entities (reticle and wafer). The third category is IC process data used to process design data, including process recipes, analysis methodologies, production machine configurations, etc. And the rest data types include all kinds of documents (policy, operation instruction, report, etc.) and other business metadata.

Key element 3, data processing owner: In a foundry, the owners and authorized rights of each data type defined below:

Key element 4, systems/applications: each data processing step of IC design data will be identified its hardware and software, including server (model/OS), manufacturing machine, application system, etc.

Key element 5, data transfer mode: it may focus on the transfer method of IC design and its derivative, including network protocol (FTP, NFS/CIFS), manually or automatically by system.

Key element 6, data life cycle coverage: the whole life cycle of IC design data starts when being transferred into foundry from IC design house, and ends when wafer will be delivered, as well the clearing or destroying of IC design data and its derivative.

The limitation of network based DLP solution is that, once the sensitive data was being encrypted or be sending through encrypted protocol, including SSL, DLP will unable to inspect the packets' content and fail to detect the leakage of data-in-movement.

When Data-in-use In IC design data processing area, endpoint based DLP will be enforced on all relevant servers to monitor and control all user activities about IC design data, including accession, modification, deletion, copy, etc

Endpoint based DLP usually identify GDSII/OASIS by its file header, and monitor all user activities on IC design data in relevant data processing servers. Some DLP products can attach a tag on specified sensitive data, such as, ftp from appointed IP address, and then trace these data, even they will be renamed or transfer to different endpoints.

Endpoint based DLP may watch the following activities:
Access, alter, copy, move and remove IC design data.
Transferring IC design data out of one host to movable storage device such as U-disk.
Printing or screen-capture operations involving IC design graphics.

Data-at-rest Storage based DLP solutions scans and protects data that will be stored in computer storage. In foundry, host based DLP can be used to scan public and private shares and

databases in internal office zone and manufacture zone to detect if IC design data exists in unauthorized places, they can be encrypted and removed, and highlighted to data owner.

Central management: - DLP consists of components designed to work together to watch and protect sensitive data wherever the data will be stored and when it is sent outside organization. SIEM (Security Information Event Management) collect and aggregate log data from various devices and applications through software called agents or connectors; filter uninteresting data and normalize to a proprietary format, analyses through correlation using contextual information and alert administrators in case of attack

In order to build up a central management platform, we use a SIEM system to gather all data access logs from different DLP detection servers and other information: firewall logs, server logs, and product information. SIEM system can analysis these information and present the information from website. It provides a management console to manage all data leakage prevention policies, workflow, reporting, users, roles, system management, and security.

Host-based firewalls provide a layer of software on one host that controls network traffic in and out of that single machine.

Anti-malware programs provide real-time protection against the malware, viruses, spyware and other harmful program on a computer.

Deploy Network Behavior Anomaly Detection (NBAD) solution to detect external and internal threat in IP data flow on East-West and North-South direction. Deploy Network Anti-Malware solution in the Internet network to detect latest malware and protect DMZ (De-militarized zone)/inside area.

Deploy Anti-DDoS, network Anti-Malware solutions at the internet edge to detect and mitigate DDoS, malware threat to protect company inside area.

Data set used in this paper is IC Design data with its derivatives, IC Process Technologies etc.

Building a modern 12 inch wafer fabrication facility (fab) with 40 nanometer process technology requires more than 4 Billion USD.

To keep up competitive advantages, semiconductor industry has strived for continuous technology migrations, high yield, cycle time, and customer service. However, information security is foundation of all competitiveness, and as well as an important index for IC design houses investigating on foundries.

To meet the information security requirements of IC design houses' various product, foundries requested to pass various information security certification authorities' audits, for example, ISO27001, Bankcard (CC, EAL, and PBOC), automotive, etc.

An Information Security Management System (ISMS) has all instrument and methods by which the semiconductor industry can use to satisfy information security in all IC application fields. One of the most important assets is "data" when build up ISMS, therefore, the data protection must take the first priority. The confidential data in semiconductor industry includes: IC design data with its derivatives, IC process technologies, etc.

IC design data is the source of the semiconductor industry chain and the core IP asset of IC design house. Such data leakage can severely impact a foundry's competitive advantage and reputation, design house's confidence, and sometimes even results in the closure of the design house. Design house, the asset owner of IC design data, have much concern of design data's protection, and as its custodian, foundry should define it as the most critical data asset in ISMS.

Another type of confidential data assets in foundry is the IC process technology. The IC process technology is the core competitive advantage of one foundry and includes massive data all over the manufacturing steps. Protect all these confidential data is an important item in ISMS of semiconductor industry.

Data-In-Use are defined as any data being used on end-user systems. Usually an endpoint based DLP product will be used to monitor data as the user interacts with them or transports from an endpoint device or client through different output channels to peripheral devices.

Data-At-Rest are any sensitive data stored in data repositories throughout the information system. Data encryption and access control are traditional technical solutions to protect data-at-rest from being accessed, stolen, or altered by unauthorized people. To detect sensitive data being revealed to unauthorized terminals, storage based DLP products is the solution mostly used to detect sensitive data being revealed to unauthorized terminals (by simply inspecting data content). Once detected, DLP will isolate the data, and a warning would be sent to data owner.

Data-In-Movement are any data that are moving across communication channels over a network. Network based DLP products is a proper solution to monitor and control data which will be sent between objects in an information system. DLP is able to detect the packets' content, no matter the movement through known protocols, such as FTP, email, http, or unknown protocols.

Study about SIEM system as a case study
In this paper, mainly considered the IC design business sensitive data flow. DLP was designed from classifying the zone level security from low to high. If any transaction or data flow from high to low, the Network level DLP filters, applies the policies and authorization and monitors, It may block the flow or stops and sends alert message to data owner.

After analysis of sensitive data categories, work flows and evaluation of different DLP products, we make use of network DLP, endpoint DLP, anti-DDoS, IPS, anti-malware, network behavior anomaly detection(NBAD), firewall, Email gateway, web gateway and SIEM to build a multi-layer data leakage detection and protection system, to prevent data loss risk from inside and outside.

The author [6], "data lineage" as used here refers to a data life cycle that includes the data's origins, where it moves over time, and what happens to the data as it is transformed. For data lineage, the tool used is InfoSphere DataStage from IBM. InfoSphere, can help to provide visibility into the analytics pipeline and simplify tracing errors back to their sources. Data lineage tools can provide a visual representation to discover the data flow/movement from its source to destination via various changes and hops on its way in the enterprise environment.

Data lineage reports with clear information about where the sensitive data leak(s) happens. Data lineage displays the data flows in ETL jobs. The data flow analysis engine can be hardware, firmware, software or any combination thereof. Real-time sensitive data alerts can be generated by database activity monitoring (DAM) and file activity monitoring (FAM) systems. These alerts will be generated for example by a security system, such as Guardium from IBM. DAM is a database security technology for monitoring and analyzing database activity that operates independently of the database management system (DBMS), and is typically performed continuously and in real-time.

A FAM system is used to find sensitive data that will be stored in files and can include: discovery to inventory files and metadata; classification to crawl through the files to look for potentially sensitive data and monitoring access to files.
Alerts generated by DAM & FAM systems could be used to inform the analysis engine about data sources contain sensitive information. DAM/or FAM can send a sensitivity status change alert in real-time to the analysis engine which can cancel or halt, the execution of the ETL, job to prevent leakage of sensitive information.

Data lineage reports with static sensitivity information, since DAM / FAM alerts are generated in real-time and use static sensitivity information.

Predictive reporting will be used to correlate between the real-time sensitive data alerts and an expected data flow as provided by data lineage reports. It can utilize a graphing algorithm (e.g., breadth-first search or depth-first search) which finds the path between two nodes (nodes representing data locations) in a graph. The analysis allows the system to determine how any of the sensitive data flows in the graph from one data location to another.

Enriched data lineage report includes detailed information about why the submitted ETL job is or is not permitted be executed. This report includes a graphical display of the data flows between data sourced and processes and a flat list of all reached data sources, along with indications about sensitivity. Problematic flows can be emphasized by line color or thickness. Sensitive data sources can be emphasized by color or bold font. Information in the enriched data lineage report can be used by users to it's the problematic areas or to remove the sensitive data sources.
When the user submits an ETL job to be run, before the job is actually executed, an analysis is performed to determine if the anticipated behavior of the ETL job leads to any leakage of sensitive-data-down the line. If the analysis engine detects, it deny the ETL execution, otherwise it process the job. The analysis engine can use data lineage reports for predicting whether running an ETL job will result in a data leakage.

When combining two or more data files or database; results in non-sensitive to sensitive information leads to sensitive data leakage. The analysis engine will prevent an ETL, job with comments that when combine creates sensitive data from running to prevent the leakage of sensitive data.
The author claims that, it is not only works on personal environment, but also works in cloud environment deployment models of any kind such as private, public, community or hybrid.

The ETL job based on predicting that execution of the ETL, job will not result in sensitive information being made accessible to an unauthorized user.

Execution of the ETL job will result in sensitive information being made accessible to an unauthorized user.
The author [7], now the company determines the data leakage by monitoring the work activities of internal employees are being studied. An intrusion detection system exists as a typical method and it is divided into an

Abnormal behavior intrusion detection method– Is to detect the behavior that deviates from the scope of the specified action by specifying the range of normal action from normal business activities. However, if the range of normal behavior specified is ambiguous, normal business activities also have the risk of being regarded as data leakage in the system

Misuse intrusion detection method– Is the one that detects the behavior similar with or same as the data leakage pattern inputted, by inputting the data leakage behavior pattern into the system which appeared in data leakage accidents in the past. This method can correctly detect the data leakage behavior pattern inputted to the system; it cannot detect the data leakage behavior pattern which is not inputted into the system.

Proposed system that can cope with the data leakage behavior pattern not analyzed by the manager is needed to prevent the data leakage by the internal staff. Suggests a method to find the data leakage through convolution neural network after writing the security log collected from the work activities of internal employees as graphs according to the scenario, by creating a scenario to find data leakage with Apriori Algorithm.

Insider Threat Protection Tool (ITPT) –Its an intrusion detection method. It judged information misuse and intentional access by first recording the behavior patterns of existing information users, and comparing the behavior of each information user with the behavior patterns of existing information users, and comparing the behavior of each information user with the behavior patterns stored in the database through the monitoring criteria stored in the ITPT. In this case a lot of supporting data will be required to input the data leakage behavior pattern, and the security manager may have to analyze the respective ground data, which can take time cost.

Machine Learning – Method of predicting behaviors and commands be taken next by each user after studying a behavior pattern of a user in usual time by introducing machine learning into an abnormal behavior intrusion detection method. Analyzed the system calls generated by the daemon programmer the root privilege program through the Bayesian network, and then notifies the administrator of the abnormal behavior when is different from the existing system call pattern. If the wrong data is used to create a behavior pattern of a user in the usual time, there is a risk that the data leakage behavior may be erroneously determined as a normal behavior pattern.

In the proposed system, after the data leakage staff will be identified, the file and folder that is accessed at the time of business activity is specified in the security log list to which the Apriori algorithm is applied so as to efficiently determine the file and folder accessed by the employee, through the association analysis, the administrator can efficiently identify the behavior scenarios and the files accessed by the employee.

In addition, it is possible to cope with data leakage more efficiently at data leakage with availability for identifying the type and contents of documents accessed by employees who leaked data to the outside, through a set of 'action' and 'access file and category' items that can be derived by analyzing the association between security logs.

Determining Data leakage scenario graphs – Convolutional neural network is a deep learning algorithm that extracts features of an image through convolution layer and pooling layer, and the learns the feature extracted by combining neural networks or judges the image. In this work, a 5 x 5 size filter was used and the filter was applied by skipping one space at a time. In the pooling layer, the space of the feature map created in the convolution layer is reduced through the Max-Pooling method, which represents the largest value in each image area. Then, the neural network layer and the softmax classifier learn the corresponding image though the image feature map created in the convolution layer and the pooling layer, or classify the input image by finding an image having a similar feature map, and let the user know the result of what kind of image it is. The author conclusion is, data leakage detection accuracy is 95% or more regardless of the number of internal employees.

In the existing scenarios, the behavior pattern related to the data leakage fed into system in advance and define the employee as the staff who leaked the data, whose behavior pattern will be detected when such fed behavior pattern is detected, but it fails, because, if the fed behavior patterns is insufficient in detecting the user behavior pattern, then the system fails to identify the data leakage.

The security log shows the content of the internal employee's business activities and could be collected from the security system/s such as virtual private networks, patch management system, data loss prevention, firewall, data rights management, database, host based intrusion detection system, Messenger, mail, etc.

The association analysis algorithm will be used to the security log history that occurred in the past data leakage accident, and the data leakage of each internal employee is judged through the data leakage judgment scenario with a set of security logs that can be displayed together when data leakage is leaked. This algorithm can represent a data leakage behavior pattern that the security administrator does not understand, so that the data leakage behavior pattern can be identified more flexibly than the existing data leakage judgment scenario.

The system implemented in Docker based Tensorflow framework environment. The security log will collects from a syslog that records events that occur on the operating system, a network-based intrusion detection system that records traffic occurring on the network, a data loss prevention and data rights management solution that records overall business activities, such as

access logs of the database holding important information of the company, and document download. This system judged whether of data leakage with higher accuracy than the data leak detection system which does not apply association analysis algorithm, also it showed lower percentage of false positive and false negative. Also, less likely to misjudge data leakage.

The author [8], pointed out that, detecting the exposure of sensitive data information is challenging due to data transformation in the content. Transformation result in highly unpredictable leak patterns. This work uses sequence alignment techniques used for detecting complex data-leak patterns. It achieves good detection accuracy in recognizing transformed leaks. The transformations such as insertion and deletion result in highly unpredictable leak patterns. Asymmetric cryptography, help the creation of a verifiable association between a public key and the identity of the holder of the corresponding private key for uses such as authenticating the identity of a specific entity, ensuring the integrity of information, providing support for non-repudiation, and establishing an encrypted communications section. Big data analysis system concept for detecting unknown attacks, big data analysis techniques that can extract information from a variety of sources to detect future attacks. To defend against these unknown attacks APT detection tool will be used. Big Data analytics with Hadoop used to analyze targeted attacks on enterprise data. Big data analytics is the process of analyzing big data to find hidden patterns, unknown correlations and other useful information that can be extracted to make better decisions. Zero day attack signatures detection using honeypot it uses LCS algorithm. Cloud model based outlier detection algorithm for categorical data. Cloud computing-based forensic analysis used for collaborative network security management system. The authors' objective is i. securely transforming the data from one place to other by using key attribute. ii. It contains the transmission of the data to the long distances. Iii. Sign to sign key parameter in according to the level of authority and iv. Hybrid data encryption.

The author [9], Modification attacks on sensitive words. Generate a list of sensitive words from the sensitive document set. Boyer Moore (BM) algorithm used to search exact sensitive strings exposed to whitespace attack. Smith Waterman (SW) sequential alignment algorithm was also employed to detect modified string attacks. TF-IDF(term frequency–inverse document frequency) method was used to extract the sensitive words of sensitive documents. Latent Semantic Indexing (LSI) was preferred to model document topics. Zemberek was used for extracting and analyzing Turkish.

Traditional Network security systems such as NIDS and firewalls use string searching algorithms to detect malicious codes, worms or Trojans. Knuth-Morris-Pratt algorithm is another string searching algorithm, but Boyer Moore is better than in average case, according to author. SW local alignment algorithm is preferred to cope with the possible modification attacks to sensitive data. It is the most optimal algorithm. In order to create the sensitive words list TF-IDF method used. LSI was used to determine the topic of the documents monitored in the network flow which are called as test documents in here.
This paper focused on the data-in-motion leakage

E-mail based data leakage prevention system – Emails with images in attachment gathered by SMTP proxy server to check if hidden information was embedded in images with online information hiding tools. If a payload is detected system, generates an alert to prevent sensitive data leakage.

Vulnerabilities of DLP Systems –Conventional security systems might be inadequate to detect transformed sensitive data; therefore string matching algorithms are not sufficient to detect leakage. In this SW algorithm is used to detect modified strings.

Content based evasion attacks – Solutions include, methods of fingerprinting, NLP, keyword matching, pattern matching, transform matching and LSI

A flow based model - A system for protecting the data in the state of data-in rest, data-in-motion and data-in-use proposed. It aims to prevent data leakage for all available hardware and software systems.

An application for preventing sensitive data leakage is improved. It was focused on threats by users who have social network accounts and cause private information leaks. The approach used was Named Entity Recognition and Twitter data was used.

Context based data leakage prevention system - When a user tries to reach a web application, a request is sent the network firewall and in the firewall, the DLP system compares the request context with pre-defined rules. If it conflicts with the rules, the system blocks the request.

Text – preprocessing used to obtain text mining for both sensitive and non-sensitive words of document, punctuation marks such as single quote, full stop, removed, and then converted all words to lowercase. With the help of zemberek library A Turkish NLP tool, the roots of all words are obtained. This step used to create a sensitive words list and inspect words in outgoing content on network.
Calculation of TF-IDF scores provide us to generate sensitive words list that one must pay attention to these list when inspecting network package whether has sensitive data or not.

Topic modeling is the key stage of the proposed system. Because three different topic related documents used as data set and determining the topic of the document is important to judge whether the topic of the inspected document is sensitive or not. In this study, Cyber Security, IoT and Economic were determined as the topics. The topics of the inspected test documents are firstly investigated to decide whether a deeper inspection is required or not. Train phase: Sensitive word should be taught to security systems, some of these systems make this process by defining policies, rules and critical word sequences, it follows a dynamic approach that update the critical words according to given predefined sensitive document set. It uses TF-IDF to extract sensitive words from document set. The train phase includes a generating sensitive words list from the predefined sensitive documents set

Detection phase: In the detection phase, firstly the topic of the preprocessed test document is extracted via LSI method. If the topic is sensitive, which we defined Cyber Security as a sensitive topic among others, a deeper inspection required. Deeper inspection includes, searching content for exact sensitive words with BM algorithm and then searching for modified sensitive word sequences via SW algorithm. This cascaded system is necessary for detecting both original and modified versions of the sensitive data. If the sensitive data is found in the content, the system generates an alarm to system administrator. The data set used is Turkish texts for the topic cyber security, IoT and Economics 5 different attack types were evaluated on the basis of criteria such as SW score, accuracy, sensitivity, specificity and different number of sensitive words. Different positions of words were considered to attacks. SW score has a determining factor to obtain a better performance. It was seen that SW score with 40 gives the highest accuracy. 10, 20 and 30 number of sensitive words was used in the experiments.

The author [10], given clear insight of usage of MirrodDroid tool to prevent sensitive data leakage from unauthorized traffic or users to the sink or agents. He used an efficient Runtime monitoring system to detect sensitive information leakage from android based mobile devices. Also, he disclosed his idea of future work to keep track the inter component communication between two separate processes and include more sensitive information sources and sinks as part of his current work.

The author [11], network security device, protecting an enterprise network, maintains a filter database containing multiple filtering rules. Each filtering rule specifies a watermark value, a set of network services for which the filtering rule is active and an action be taken by the network security device. The network services include a web-based electronic mail (email) service, Simple Mail Transfer Protocol (SMTP), Internet Message Access Protocol (IMAP), Post Office Protocol 3 (POP3), an instant messaging program, a file sharing service and / or a device synchronization service. Network traffic received by network security device that is, originated within the enterprise network. The network traffic is directed to the destination residing outside of the enterprise network, is associated with a particular network service and contains a file. A watermark value embedded within the file is identified by the network security device. A determination is made by the network security device regarding whether there exists a filtering rule specifying a watermark value matching the watermark value embedded within the file and for which the filtering rule is active for the particular network service. When the determination is affirmative, the action specified by the filtering rule is performed by the network security device. The DLP sensor is a module that is capable of detecting watermarks in files and / or extracting the information contained in the watermark and the file. The user may log into the gateway and configure the DLP sensor to detect a particular watermark and responsive to the detection perform a specified action. The action may include either blocking or passing the file at the gateway

The digital watermark is a technology for embedding information, such as the name of the copyright holder, reproduction history, and the like in data, such as an image, document, voice and the like. A watermark includes plain text (a visible watermark) or encoded information (an invisible watermark) or watermark is a value (e.g., hash value) that can be used to look up the associated company identifier and sensitivity level. The watermarking program may operate through a Common Internet File System (CIFS) share. CIFS is an application layer network protocol used for providing shared access to files, printers, serial ports, and other communications between nodes, such as computers. CIFS serves thus make their file systems and other resource available to clients on the network. A sensitivity level of the document or information contained therein (e.g., critical, high, medium, low). Data before leave the organization, the organization gateway monitors the files, and water marking files, it fetches the watermark and check against the database of the organization, if it matches, applies the filtering rules, if any violation, just it blocks the file or traffic at the gateway itself.

The author [12], Time-based file assured deletion, introducing in the existing system which means that files can securely deleted and can remain permanently inaccessible by anyone after a pre-defined duration of time. The main idea is that a file is encrypted with a data key only by the owner of the file, and again this data key is further encrypted with a control key by a separate key manager. The key manager is a server who is responsible for cryptographic key management. The control key is time-based, meaning that it will be completely removed by the key manager when an expiration time is reached, where this expiration time will be specified when the file is first declared. Without the

control key, the data key and hence the data file remain encrypted and are inaccessible. FADE leverages existing cryptographic techniques, including attribute based encryption (ABE), and a quorum of key managers based on threshold secret sharing. To protect the integrity of a file, the client computes HMAC signature on every encrypted file and the HMAC signature will be stored together with the file in the cloud. Experimental results provide the depth of knowledge about the performance-security trade-off when FADE is deployed in practice. Provide information about the cloud backup, storage files data safely and securely.

III. CONCLUSION AND FUTURE SCOPE

Data security is not an easy task even in top leading organizations suffering fear of data leakage, this paper discussed most of the latest work carried out in the area of detection of data leakage and prevention algorithms, tools and technologies supported. Future of the work needs the development of automation solutions in which the systems would be able to learn itself for its self-defensive capabilities and the human involvement in the defense mechanism will be kept minimal.

REFERENCES

[1] Carlie Carlett, et. Al, "A Scientific Research and Development Approach to Cyber Security", report submitted to Department of Energy, Dec 2008.

[2] Eric Yuan, et. Al, "A Systematic Survey of Self-Protecting Software Systems", ACM, 1539-9087/2010/03-ART39.

[3] Neeraj Kumar, et. Al, "Detection of Data Leakage in Cloud Computing Environment", Sixth International Conference on Computational Intelligence and Communication Networks, 978-1-4799-6929-6/2014,IEEE

[4] Marco Pistoia Omer Tripp, et. Al.," Labyrinth: Visually Configurable Data-leakage Detection in Mobile Applications", 16th IEEE International Conference on Mobile Data Management, 2015.

[5] Sherry Zhu, An Efficient Data Leakage Prevention Framework for Semiconductor Industry, Proceedings of the 2016 IEEE IEEM, 978-1-5090-3665-3/16/$31.00 ©2016 IEEE

[6] Becker et. Al., Real-Time Data Leakage Prevention and Reporting, 21st September 2017, US Patent publication No. US 2017/0270310 A1

[7] Min-JiSeo, et. Al., An advanced data leakage detection system analyzing relations between data leak activity, International journal of Applied Engineering Research ISSN 0973-4562 Volume 12, No.21,2017, pp 11546-11554

[8] Neha Ramteke, et. Al., Review paper on prevention of data leaks using 3DES Encryption, International journal of innovations & advancement in computer science, IJIACS, ISSN 2347-8616, Vol 6, Issue 11, Nov 2017.

[9] YavuzCanbay, A Turkish language based data leakage prevention system, 2017 5th International Symposium on Digital Forensic and Security (ISDFS), 978-1-5090-5835-8/17/$31.00 ©2017 IEEE

[10] Sarker T. Ahmed Rumee, et. Al., "MirrorDroid :A Framework to Detect Sensitive Information Leakage in Android by Duplicate Program Execution", 978-1-5090-4780-2/17, IEEE, 2017

[11] Michael D Nelson, Data Leak Protection, 10 May 2018, US Patent Publication, us 2018/0131674 A1

[12] Shivakumara T, et. Al, "To incorporate value-added security features into current data for outsourcing cloud data applications through Secured Access Control and Assured Deletion", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 7, Issue 3, May-June 2018 , ISSN 2278-6856

Authors Profile

SHIVAKUMARA T working as Assistant Professor for Department of MCA, BMS Institute of Technology and Management Bangalore. He has completed his masters' degree (Master of Computer Applications). Teaching the masters' degree computer applications courses prescribed by Visvesvaraya Technological University (VTU). Actively involved in teaching-learning process, as an outcome of it he was able to publish 3 text books, laboratory manuals, learning materials in coordination with co-authors in the same field. He has published few national conference papers and journals. His current research focuses on data and information security - data leakage prevention. He has been engaged to create linkage between industry and academia and organized talks, workshops, and alumni interactions.

Dr. Rajashekhar M Patil received B.E. in Computer Science and Engg from REC Bhalki(Gulbarga University) in 1996 and M.Tech from AIT, Bangalore 2004 and doctorate obtained from Dr. MGR University in 2013. He has actively involved in research area like IDS and Network Security. Several papers published in reputed Journals and conferences.

Muneshwara M S received B.E. and M.Tech from VTU in 2005 and 2012 respectively. During 2006 to present affiliated to BMS Institute of Technology and Management as Assistant professor at Department of CS&E, actively involved in research area like Distributed Network security and Cloud Computing. Several papers published in reputed Journals and conferences pursuing the PhD in Computer Science under VTU

