Fourth Generation Data Leak Prevention – A Paradigm Shift

By: Tarique Mustafa, Founder & CEO/CTO, GhangorCloud, Inc.

1. Executive Overview

Ever since the emergence of Data Leak Prevention (DLP) solutions, circa 1998, the industry has already been through three generations of DLP solutions, and is now witnessing the emergence of the fourth-generation solutions. It is instructive to compare and contrast the characteristics, capabilities and limitations of DLP solutions during these stages of evolution. Fig-1 enumerates some of the key characteristics of DLP solutions over these evolutionary stages.

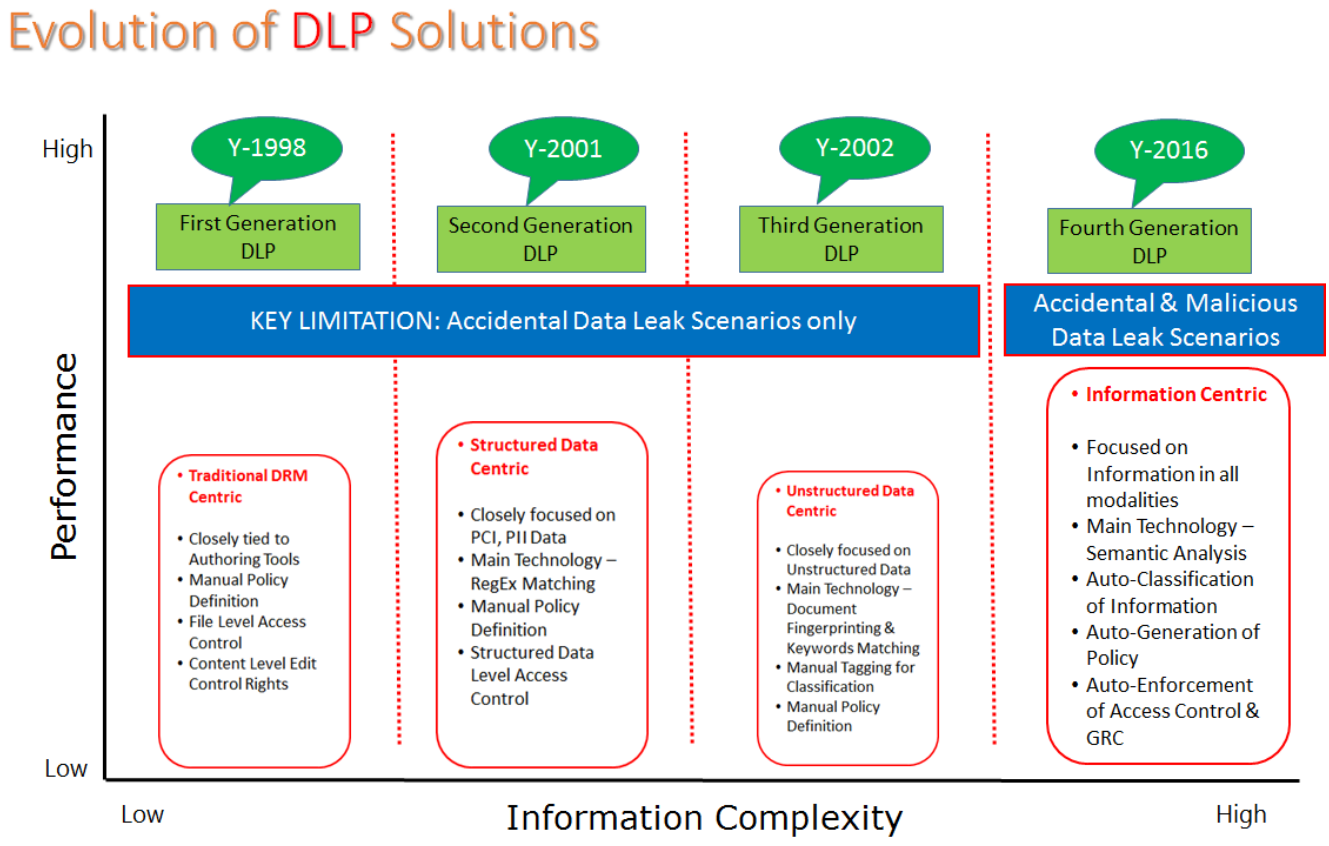

Fig-1: Evolution of Data Leak Prevention Solutions

First Generation DLP – Traditional DRM Centric:

First Generation DLP solutions were primarily focused on Digital Rights Management (DRM) functionalities. These solutions were tightly coupled with document creation and authoring tools. The main objective was to create controls for each piece of content in a document so that only authorized individuals could perform read/write/copy/delete and other operations on the constituent elements of the document. Credentials based controls (e.g. "Password Protection") were the main form of access control so that only authorized individuals with the knowledge of the assigned password could open the document and gain access to the content.

Second Generation DLP – Structured Data Centric:

Second Generation DLP solutions focused on identifying and protecting "Structured Data" such as Credit Card Information or Social Security Number. These solutions were built using basic "Signature Matching" technology such as Regular Expressions and Pattern Matching Algorithms for standard data items (such as PCI, PII, ePHI, etc.). These solutions were constrained to limited applicability, as they were only capable of identifying (and hence protecting) the Structured Data that can be mapped using the Signature Matching technology. Policies could then be defined to enforce appropriate security measures against theft and/or unauthorized disclosure of confidential data/information.

Third Generation DLP – Unstructured Data Centric:

Third Generation DLP solutions were primarily focused on identifying and protecting sensitive data in "Unstructured" form. This can be any confidential information found in loosely formatted documents (such as text files, word files, pdf, spreadsheets, emails, power point presentations, memos, etc.). Unfortunately, the Third Generation DLP solutions could not scale in their effectiveness and usability beyond the simple and trivial use case scenarios in real-life enterprise deployment.

These solutions heavily rely on 'Document Fingerprinting' technology which requires (i) a-priori manual identification of sensitive documents, (ii) extensive pre-processing to create the 'Document Fingerprints', and (iii) large storage to store the 'Document Fingerprints'. This technology essentially reduces to a form of pattern matching which requires manual identification and tagging of document types containing sensitive information. The Third Generation DLP solutions also have another grave vulnerability – their dependence on Manual Classification and Tagging of sensitive content. This dependency on manual intervention makes the Third Generation DLP solutions completely ineffective against 'Malicious Data Leak' scenarios wherein an unscrupulous employee (or a 'Mole') in the enterprise has no incentive to correctly identify and classify sensitive information, and tag it with the correct security tags. Due to their inability to correctly Identify, Classify and Protect the confidential data in automated fashion the Third Generation DLP solutions are inherently limited to the ***"Inadvertent or Innocent DLP"*** [1] scenarios and are NOT able to scale up to meet the challenges of the real-life ***"Malicious DLP"*** [2] scenarios. The Third

[1] Accidental or Innocent Data Leak – These are basic data leak scenarios where the incident happens inadvertently due to a mistake on part of the user. For example, including somebody in the recipient list of an email message that contains information or attachment with classified data while the recipient is not authorized to have access to the classified data.
[2] Malicious Data Leak – These are advanced real-life data leak scenarios that involve purposeful and intentional exposure or disclosure of confidential data/information to unauthorized individuals and parties.

generation DLP solutions can be very easily by-passed or "gamed" by simple Evasion Techniques hence defeating the whole purpose of these marginalized solutions.

Furthermore, the total cost of ownership (TCO) over the life cycle of these products is very high due to the cost associated with Manual Classification and Tagging.

Fourth Generation DLP – Information Centric:

Fourth Generation DLP solutions obviate the critical issues of the previous generations as well as address the challenge of 'Malicious Data Leak' scenarios. The Fourth Generation DLP solutions do not depend on manual processes and hence, are non-susceptible to human errors and invulnerable to purposeful acts of 'Malicious Data Leak'.

The key differentiating characteristics of Fourth Generation DLP solutions are the ability to automate the functions of data/information identification, classification, analysis and enforcement. The Fourth Generation DLP solutions must deliver;

1. **Automated Classification of data/content:** The ability to perform Automated Data Identification and Data Classification (regardless of its modality, i.e. Structured or Unstructured Data) results in the elimination of 'Manual Classification and Tagging' process hence enabling a more sophisticated and less error-prone DLP paradigm. This also eliminates the possibility of ***"Purposeful Misclassification"*** of data/content thus reducing the risk of ***"Malicious Data Leaks"***.

2. **Automated Generation of Policies:** The ability to perform Automatic Policy Generation without human intervention, eliminates the chances of human errors. This also obviates the possibility of ***"Purposeful Misconfiguration"*** of policies thus enabling better control and avoidance of the risk of ***"Malicious Data Leaks"***.

3. **Automated Enforcement of Advanced Access Control:** The ability to perform Automated Access Control eliminates the chances of ***"Purposeful Misauthorization"*** of Access Rights to critical pieces of data/content. This enables better Access Control and greatly reduces the possibility of ***"Malicious Data Leaks"***.

Collectively, these advanced capabilities enable the Fourth Generation DLP products as an ideal platform to provide protection against ***"Malicious Data Leaks"***. This also results in a much lower TCO over the life cycle of the Fourth Generation DLP products.

- ***The most differentiating characteristic of Fourth Generation DLP is its ability to protect data/information against 'Malicious Data Leak' via automation of key processes.***

2. Foundational Components of Fourth Generation DLP System

As depicted in Fig-2, in order to achieve the Fourth Generation DLP capabilities it is imperative that the solution must incorporate the following foundational technologies;

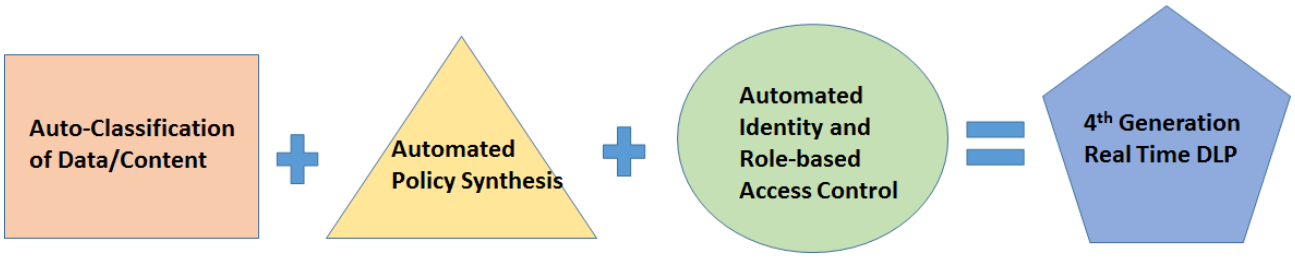

Fig-2: Key Components of Fourth Generation Data Leak Prevention Solutions

1) **Auto-Classification of Content –** the DLP system must be able to recognize and Auto-Classify confidential and/or mission critical data in any format without human intervention. Traditional Data Classification systems that require "Manual Input" are greatly ***ineffective*** as they are not only error-prone and unscalable but actually highly susceptible to Malicious Data Leak scenarios.

   To this effect, an Auto-Classification technology must exhibit following characteristics;
   - **No Manual Intervention:** The Auto-Classification process should NOT require any manual processing of data/content such as *Pre-Tagging* or *Pre-Marking* of data/content.
   - **No Pre-Processing:** The Auto-Classification process should NOT require any pre-processing of data/content such as *Document Fingerprinting* or *Hashing* of Data/Content.
   - **No Manual Heuristic Training:** The Auto-Classification process should NOT require manual heuristic training.
     NOTE: Heuristic Training based technologies are NOT *"Admissible"* as they are known to be ineffective for complex DLP use case scenarios.

   NOTE: Auto-Classification Engine must be able to perform ***Automatic Data Identification***. To this effect, sophisticated data identification techniques are imperative. Simple Keyword and Lexical Matches are NOT enough as they have been proven to be error prone resulting in high False Positive rates.

2) **Automatic Generation of Policy –** the DLP system must be able to automatically generate comprehensive set of policies for enforcement of data leak prevention. Traditional DLP systems are typically limited to manual policy creation process which is not only tedious but also error prone.

   DLP Policy definition is the single most critical process for successful deployment of a DLP regime. DLP Policies are inherently more complex and require deeper understanding of sophisticated use case scenarios[3] in order to ascertain acceptable level of "Completeness" and "Use Case Coverage".

   To this effect, an Auto-Policy Generation technology must exhibit following characteristics;

   - **No Manual Intervention:** The Auto-Policy Generation process should require little to no manual intervention in the Policy Generation process. All relevant *Policy Primitives* should be automatically generated thus eliminating human error and inconsistency.
   - **No Post-Processing:** The Auto-Policy Generation process should NOT require any manual post-processing such as *Manual Policy Embossing*[4] of Data/Content.
   - **No Manual Heuristic Training:** The Auto-Policy Generation process should NOT require manual heuristic training of the Policy Parameters to perform Incident Correlation.
     NOTE: Heuristic Training based technologies are NOT *"Admissible"* as they are known to be INEFFECTIVE for DLP Scenarios.

   Furthermore, the DLP Policies must have a simple syntax and be Tractable – the Policies must be **(a) Human Readable**, and **(b) Machine Executable**.

3) **Automatic Enforcement of Advanced Access Control –** the DLP system must be able to automatically derive comprehensive set of sophisticated Access Control Primitives. Traditional DLP systems are either dependent on extensive manual enumeration of Access Control Primitives or rely heavily on the Policy definition process – both of the two approaches is extremely cumbersome and constrained in its ability to provide the requisite coverage of "Use Case Scenarios" in sophisticated real-life DLP deployments.

   To this effect, an Advanced Access Control paradigm must exhibit the following characteristics;

   - **Identity and Role driven Access Control:** The Access Control mechanism must be based on "Identity and Role" based approach wherein access rights and corresponding violations can be automatically deduced from the identity and functional role of an actor (i.e. employee, application or device).

[3] Real-life DLP Use Case Scenarios are quite complex and hence far more difficult to enumerate a-priori.
[4] Policy Embossing pertains to the process of marking a data or content with a given policy. This is usually accomplished via insertion of policy description in the Metadata of the data item or content.

- **Contextual-Conceptual Correlation:** The Access Control mechanism must incorporate algorithms to perform sophisticated Contextual-Conceptual Correlation (i.e. who should have access to what type of information?) for detection and pre-emption of complex data leak scenarios.
- **Automatic Real-time Enforcement of GRC:** The Access Control mechanism must incorporate the GRC (Governance and Regulatory Compliance) principles from the ground up so that sophisticated SoD (Segmentation of Duty) principles can be enforced.

In summary, the Fourth Generation DLP solution must incorporate advanced features, as depicted in Fig-3, to deliver reliable efficacy for the real-life data leak scenarios.

Automation of these features is Imperative inasmuch as manual dependency is prone to inadvertent and/or malicious circumvention of these processes. Furthermore, automation also greatly reduces the Total Cost of Ownership (TCO) by eliminating the cost and time overhead associated with manual performance of these tasks on a continuous basis.

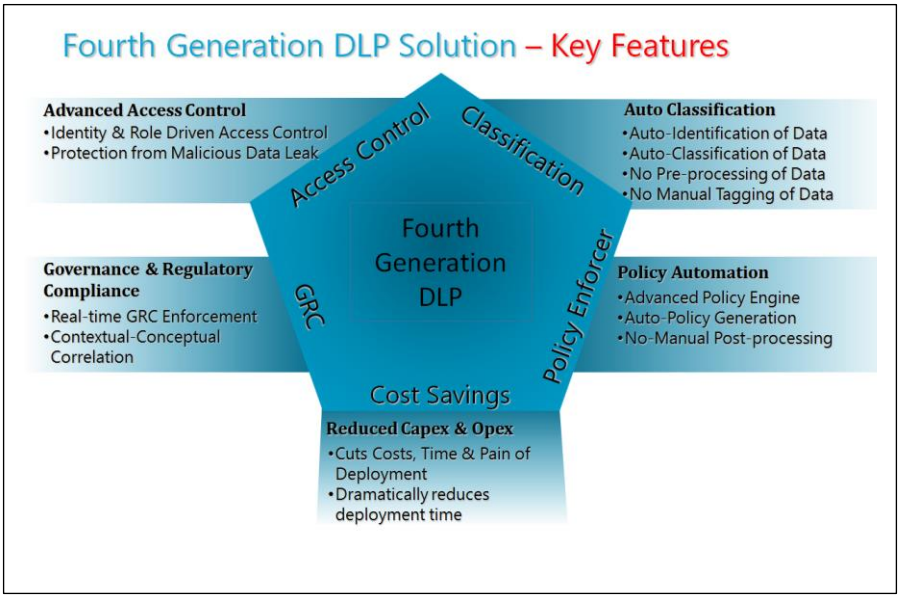

Fig-3: Key Features of Fourth Generation Data Leak Prevention Solutions

3. GhangorCloud's Information Security Compliance Enforcer (ISE) – a Fourth Generation DLP System

GhangorCloud's Information Security Compliance Enforcer (ISE) is the pioneering Fourth Generation Data Leak Prevention solution that has been built from the ground up to address the deficiencies and constraints of the previous generation DLP solutions.

As illustrated in Figure-4, the ISE platform is comprised of key technology components and sophisticated data security features such as Automated Security based Data Classification, Data Identification, Automated Policy Synthesis and Advanced Access Control. The ISE platform eliminates Manual Classification and Tagging, eliminates the pre-processing cost and enables protection against Evasion Attacks, thus enabling true 'Malicious DLP'.

GhangorCloud Solution Key Features

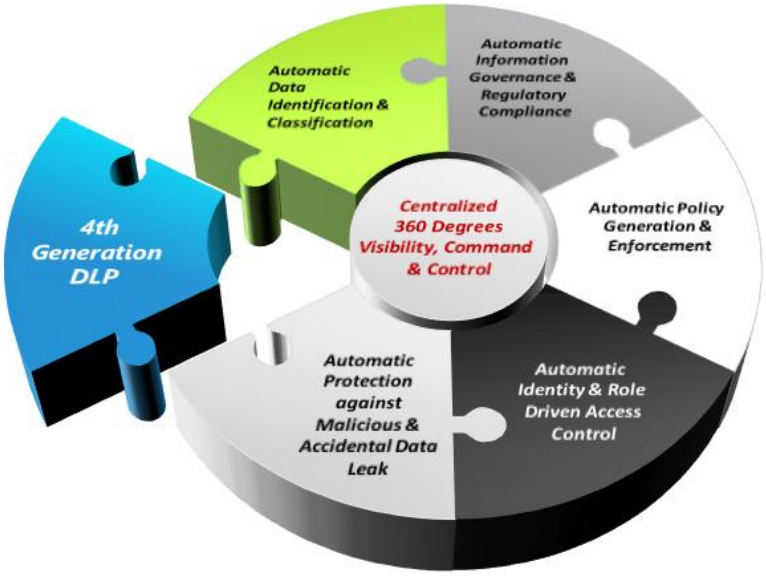

Fig-4: GhangorCloud – Information Security Compliance Enforcer (ISE) Platform, Fourth Generation DLP

Automated Real-time Data Classification: The ISE DLP System incorporates a unique Auto-Classification Engine that works with *Embedded Ontologies* to Auto-Identify and Auto-Classify sensitive information. The Auto-Classification engine looks at all the content in a data transaction, not just the file type or transmission protocol. It can identify sensitive information as granular as specific words and phrases. The Auto-Classification Engine completely replaces the requirement for manual tagging or fingerprinting of sensitive information. It can readily work out-of-the-box and does not require any pre-processing of data or a laborious training / learning process.

Automated Policy Synthesis: The ISE DLP System incorporates a powerful Auto-Policy Generation Engine. Sophisticated Policies are automatically created by correlating types of Actors with types of sensitive Information. Each class of Actor has policies associated with them which define what Operations they can conduct, what other Actors they can share Information with and what Information is permitted to sending and receiving Actors. For example, a finance person can send revenue information to the CEO but not to someone in sales.

Automated Access Control Enforcement: ISE DLP System incorporates a sophisticated Auto-Access Control Engine. Advanced DLP algorithms correlate Actors (people, devices or machines), Operations (data transactions) and Information (data being communicated) with Policies. The result of the correlation is policy enforcement based on *who is transferring/transmitting what data to whom using what communication medium*. Automated controls are enacted depending on the sensitivity of data being transmitted.

Based on the severity of an exfiltration attempt, the system can enforce actions to allow, quarantine or block transmissions. Attempts to send lower levels of sensitive information to inappropriate receivers cause the transmission to be halted and the sender queried for permission. More serious levels of information are quarantined and sent to management levels for approval. Highly sensitive data transmissions are blocked completely.

Automatic Enforcement of GRC: The ISE DLP System incorporates unique Identity and Role-based GRC Enforcement Engine. The system can enforce policy based on roles and job functions of users. Segmentation of Duty policies use business logic to compare content with users and make automated decisions on information control. Individuals or groups are Actors. Each Actor is limited to certain types of data for their own usage and for sharing with other Actors. Actors are limited in the Operations that they can apply. Different Operations such as emails and downloads are limited for each Actor. For example, a billing clerk can email information about billing records, but not upload them to the web or send them to outside persons.

4. Summary

Ever since the inception of data leak prevention solutions (circa Year-1998 to date), the data leak scenarios and data leak exfiltration attacks have tremendously evolved into more and more complex data security threats. The advent of Advanced Persistent Threats has further compounded the challenge from Accidental to Malicious Data Leak and Exfiltration Attacks. Traditional 3rd Generation DLP solutions (circa Year-2002 to date) seriously lack in their ability to address these advanced and ever-evolving data leak and exfiltration scenarios. A new set of technological innovations is essential to address the new and emerging Data Leak and Exfiltration Attack Vectors.

GhangorCloud's unique (patented and patent-pending) 4th Generation DLP technologies address key issues and limitations in the existing traditional Data Leak/Theft Prevention solutions for Enterprise and Service Provider networks, while providing superior accuracy, performance and maintainability. GhangorCloud's ISE product, which has a highly scalable architecture, can be quickly deployed across an enterprise. The ability to automatically identify content, automatically classify it, automatically generate security policies in real-time without requiring constant tedious manual intervention, and provide real-time protection against purposeful Malicious Data Leak makes it unique.

GhangorCloud is the leader in the emerging next generation enterprise information security & access management market. To schedule a demo please email to info@ghangorcloud.com or submit a request at: http://www.ghangorcloud.com/request-demo/

CONTACT:
GhangorCloud, Inc.
2001 Gateway Place, Ste: 710 West Tower
San Jose, CA 95110
www.GhangorCloud.com