happiest minds
The Mindful IT Company
Born **Digital** . Born **Agile**

SECURING THE CLOUD DATA LAKES

The present digital world is all about data, which probably matters the most for any organization today. Every day over 2.5 quintillion bytes of data is generated. Enterprises moving to the cloud have provided the flexibility to access the data from anywhere, any device. These data stored on the cloud data lake platform brings a unified analytical environment that includes cloud storage, multiple data processing engines, advanced analytical tools, and more, enabling scalability, agility, and cost-benefit for an enterprise.

This whitepaper provides a comprehensive guide on securing your cloud data lake platform with industry best practices endorsed by leading IT security experts.

Let us understand some of the infrastructural and security challenges that every organization faces today before delving into the concept of securing the cloud data lake platform. Though Data lakes provide scalability, agility, and cost-effective features, it possesses a unique infrastructure and security challenges.

READ ABOUT CLOUD DATA LAKES CHALLENGES →

CLOUD DATA LAKES **CHALLENGES**

Data Migration 1

The first and foremost challenge for any organization is migrating the data into the cloud. It's not only complex but also requires huge investments, especially when it is done repeatedly.

Data Analytics 4

It is very difficult for the security team to filter and detect malicious activity. Traditional SIEMs have limited capabilities that rely on restrictive languages to query and interact with the data but cannot handle advanced analytics. Any organization moves to the cloud mainly to its analytics feature that combines, transforms, and organizes disparate data sources. Though many cloud service providers offer analytics solutions, a robust solution is required to effectively utilize and hook into these analytics platforms.

Unstructured Data 6

The major challenge is handling unstructured data, making it difficult for the security team to search and analyze huge volumes of data. In addition, most security tools leave data normalization to the users, making it more challenging for security analysts to understand relationships between malicious indicators and events across time.

Data Management 2

Though data lakes support all data types, managing those data in multi and hybrid environments is the biggest challenge, and it is an intensive process. When things go wrong, data swamps may happen, and poor data management requires many fixations.

Scalability 5

The modern EDR and XDR solutions generate large amounts of data but are not built or fully capable enough to manage the data produced. Hence, when these data are pushed to the SIEM solutions, the time it takes to search, efforts to maintain, and scale are massive.

Data Storage Cost 3

Most organizations intentionally reduce the security data collection required for defending against attacks due to its high license cost. This is the primary reason organizations lack an effective investigation, which is a huge anti-pattern where breaches get unnoticed. Organizations depend on third-party cloud service providers. These cloud service providers charge based on the time more than the size of the data stored. The cost gradually increases over time. This may become a huge burden for the businesses where the existing engineering and IT costs might be invested to rent cloud services.

HOW TO SECURE YOUR CLOUD DATA LAKES

To overcome the above challenges related to scaling, detection, cost, and analytics, organizations must separate the storage and adopt serverless services that reduce the overheads and provide flexibility in processing data at a large scale. Having an effective security data lake helps you to centralize data and enhances the power of threat detection, analytics, and compliance initiatives. This eventually supports complex use cases for security analysis, including threat hunting at scale.

01

Implement Data Loss Prevention (DLP) Strategy

Cloud data lakes leverage persistent data in cloud objects to optimize and maintain data integrity and availability. The capabilities like object versioning and retention capabilities provide crucial redundancy in the accidental deletion or object replacement. Ensure every service that manages, or stores data is identified and classified based on their sensitive level deploy the appropriate level of security and control. The sensitive level is based on security and regulation standards. Ensuring a proper evaluation of all the services that manage and store data is crucial. In addition, limiting the access from deletion or updating functions will eventually reduce data loss, and having a backup plan will enhance the overall data retention capabilities.

02

Separate Security Functions

The foremost practice is to separate security from non-security functions, which is essential to mitigate risk. User's access must be restricted from critical business data and provide access to those required to perform the task. When it comes to cloud data lake platforms, access to both cloud and data lake platforms should be limited to only experienced security personnel and ensure only this security personnel have access to alter cloud security controls. A minor misconfiguration or lack of knowledge can become vulnerable to a security breach.

03

Hardening the Cloud Platform

Harden and isolate your cloud data lake deployment with a unique cloud account. Cloud services like AWS, Azure, Google, and more can easily leverage organizations' services to create and manage new accounts. The most compelling model for logical data separation on cloud platforms is to use a unique cloud account for your deployment. Implementing hardening protection in line with CIS, Benchmarks will ensure security by providing logical data separation from your other cloud services.

04

Secure the Network Perimeter

After isolation and hardening the cloud account, building a secure network perimeter for the environment is important. You adopt any method to secure the network perimeter, but the method you select must be in line with specific circumstances. Key compliance or bandwidth requirements may well indicate that a private connection or a cloud VPN (Virtual Private Network) is required. The firewall is crucial for maintaining traffic control and visibility of any sensitive data stored in the cloud, and non-private connections are allowed. Leveraging a third-party next-gen firewall will offer you the features of intrusion prevention, application awareness, and threat intelligence and generally complement native cloud security tools. By deploying a virtualized enterprise firewall in a hub and spoke design, you can ensure effective security in place with consistent compliance throughout your cloud environment.

Throughout your cloud infrastructure environments, only firewalls should have public IP addresses. Use robust entry and exit policies with breach prevention profiles to reduce the risk of unauthorized access and data exfiltration.

05

Implement Host-Based Security

Host security is a broad attempt and must adapt to specific service and function use cases.

Host Intrusion Detection

Host intrusion detection is a crucial component that runs on the host and uses various detection techniques to find suspicious activity, either known threat signatures or behavioral anomalies. It alerts the administrators if any unusual event is detected. Leveraging Machine learning algorithms combined with either threat or anomaly-based systems can even offer higher level detection and respond to potential threats and attacks.

File Integrated Monitoring

Considering most exploits, attackers require elevated rights to get into the system and corrupt files or services. FIM solutions provide you with the ability to detect and track the changes made by the attackers and alerts you with the detailed changes made within the system. Some File integrity monitoring (FIM) also provides an advanced feature to restore files to their previous state.

Log Management

Log management is very crucial and needs proper attention while implementing this in your security practice. The analysis of logged events provides a vital role in investigating security incidents. Log storage, retentions, and deletion policies should be carefully designed with proper procedure and control to meet regulatory compliance requirements. The most common method to enforce secure log management policies will copy logs into storage in real-time. Many open-source log management tools and licensed versions of log management tools are designed to integrate with cloud-based solutions, which offer additional data visualization capabilities and usage alerts.

06

Implement Strong Identity Management and Authentication Measures

Identity management is the main pillar for having robust access control for cloud data lakes. The first step in building a secured data lake is integrating your identity provider with the cloud provider. Managing third-party applications or deploying data lakes with multiple services requires a patchwork of authentication services such as SAML clients and providers to use Auth0, OpenLDAP, Kerberos, Apache Knox, or others.

For example, AWS provides help with SSO integrations for federated EMR Notebook access.

07

Leverage Authorization Controls

Cloud providers provide data and resource access controls and column level filtering to secure sensitive data as part of their platform-as-a-service solutions. This Identity and Access Management (IAM) policies and role-based access controls (RBAC) allow you to limit access control using the principle of least privilege. Cloud providers offer fine-grained access control through their Lake Formation service, which automates the process to secure your data lake.

Depending on the number of services running in the cloud data lake, you may need to extend this approach with other open-source or 3rd party projects such as Apache Ranger to ensure fine-grained authorization across all services.

08

Enforce Encryption

Following the encryption guidance provided by the cloud service providers is crucial for cluster and data security. It requires a strong understanding of Identity and Access Management (IAM) key rotation policies and application configurations to effectively leverage these fundamental security functions. Encryption must protect both data at rest and data in motion. You may provide a self-certificate in case if you are using integrated third-party cloud services. Amazon S3 supports multiple encryption options where AWS Key Management System (KMS) provides centralized control over the encryption keys to protect data assets. This KMS offers the flexibility to rotate, disable, delete, define usage policies, and audit the use of encryption keys.

09

Vulnerability and Patch Management

Leverage a comprehensive vulnerability and security patching strategy that combines automated detection, risk assessment and complexity, testing, and patch deployment. Using alternative mitigation techniques, turning off unnecessary services, and employing firewall controls will reduce the vulnerability time. Having clear visibility on your vulnerability management program is crucial and understanding the risk factor within your environment will reduce exploitation and data loss.

10

Compliance Monitoring and Incident Response

Compliance monitoring and incident response are the heart of any cloud security functions, including early threat detection, investigation, and response. Integrate them to perform cloud monitoring if you already have existing security information and event management (SIEM). Cloud deployments have unique threats that require experience and practice to identify and resolve the issues correctly. Bring in the best incident response strategy in place to react quickly to security incidents.

Conclusion

Organizations moving to the cloud aim for a robust and flexible environment with analytical-driven AI/ML capabilities that benefit enterprises from agility, scalability, and cost-effective solution. Securing this complex environment requires skills and adequate security measures to protect data and the surrounding environment, such as cloud platforms, storage, data processing engines, and analytical tools that carry the risk of exploitation. Following the above security best practices, adhering to compliance, and utilizing the maximum analytical benefits of cloud data lakes can help you manage and protect data effectively.

About the Author

Vinayak S
Senior Practice Manager,
Cloud & Infra Security

Vinayak S has over 13+ years of experience across various IT security service domains like Cloud Security, Infrastructure Security, and Cyber Security Practice. He is currently working as a Senior Practice Manager for Cloud and Infrastructure Security Service at Happiest Minds Technologies Ltd. Vinayak carries a niche experience across the BFSI, Health, and Global sectors. He is responsible for designing and implementing the security infrastructure for multiple customers, multiple technologies covering the Cloud and On-Prem in Compliance with all security specifications that are being followed as the best information security practice.

About Happiest Minds Technologies

For more details, write to us at
Business@happiestminds.com

happiest minds
The Mindful IT Company
Born **Digital** . Born **Agile**

www.happiestminds.com