ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 8, Issue 3, pp: (235-237), Month: July - September 2020, Available at: www.researchpublish.com

Data Lakes Security Challenges

Ahmad Alhadab

Hassa, Saudi Arabia

Abstract: **organizations and enterprises of all sizes implement data lakes as it is a main requirement for digital transformation initiatives. Securing data lakes has many challenges due to the special nature of the data lake technology, it's usage within the organization, and the technical implementation details. Securing any data lake should take into considerations such challenges in order to maintain the value of implementing data lake and maximize ROI. Failure to secure data lakes could lead to potential losses both financially and intellectually. This article discusses three challenges and advise three best practices that could be used to secure data lakes while addressing these challenges.**

Keywords: **Data lakes, Security, Advanced Analytics, Cloud.**

I. INTRODUCTION

Date lakes are considered the next step in logical evolution for business data hosting history after data warehouses and data marts. There are many different definitions for this concept and that's depending on the way they are implemented to serve a purpose or solve a business problem. Nevertheless, data lakes can be generically defined as "A repository of enterprise-wide raw data, but combined with big data and search engines, a data lake can deliver impactful benefits. Data lakes bring together data from separate sources and make it easily searchable, maximizing discovery, analytics, and reporting capabilities for end-users" [1]

The term "enterprise" could be added in to indicate that it is a private data lake. Big enterprises, mid-size and small-size organizations all alike looking for a competitive edge by enhancing their products, optimizing their operations, or improving their services. Big data analytics could be the starting point of all that. Data lakes is usually the starting point for all AI, ML advanced analytics projects

II. CONCENRS

In their rush to implement data lakes, organizations sometimes overlook the importance of having clear and solid security measures in place before going life and start ingesting business data in. This could lead to grave risks and massive financial losses due to hacker ransomwares and paying paneities due regulatory bodies because of incompliance to industry regulations.

There is no one security standard or one security measure that could fit all data lakes. Each data lake implementation has different purpose, serving different company and has different user types. All that pose several challenges in implementing security measure that should secure, not harm. Some main concerns explained as follows.

ELT: Late Processing

Early data ingestion and late data processing is one key innovation of data lakes [2]. Data from different sources (IoT devices, sensors, log files, or any structured data source) is to be replicated (E: Extract) into the data lake as soon as they are ready to. Data processing (T: Transform) however, is to happen when there is a need to produce data insights or generate business reports. In addition, the way data is to be processed could be defined on the fly as seen possible based on the current specific requirements for each reading request. That means that data shouldn't be altered from their original state ingested in. alternatively, different copies of the data should be produced for each reading or processing request. This will mean that we will have data in the lake with different stages of cleaning and enrichment.

Cloud Hosting

Data lakes is a fairly new concept that both big enterprise and organizations of all sizes want to embrace in order to have a competitive edge quickly. It's more related to the amount of data generated rather than the organization size. For that reason, small and med-sized organizations tend to have their data lakes hosted on public cloud service powered by a third party and accessible over the internet. That could be the only feasible way to have a business data lake but it imposes risks on having all business data gathered on one "basket" reachable by all outsiders.

Driven Value

One of the drivers for enterprise data lakes is the value added for business analysts and planners by eliminating all data silos within the same enterprise by creating one source that contains all kinds of business data coming from all organizations and departments. On the other hand, there is a risk on having restricted or highly classified business data accessed by people who's not supposed to.

III. BEST PRACTICIES

The following security and design measures could be implemented to address the discussed concerns.

Data Lake Zoning

Data lakes should have multiple zones that could allow for defining logical and/or physical separations of data based on the amount of enchainment and pre-processing. This will help in better secure and organize the data in the data lake by applying different security measures for each zone. As a best practice, it's recommended to have four zones in the data lake [2]:

1- **Temporal Zone:** Serves as a transitional zone for temporary copies of data. Data in this zone should be deleted as soon as they are not needed anymore and it should be only accessible by data engineers and data lake administration team.

2- **Raw Zone:** This zone will host the raw data ingested from different sources. This is the are suitable to run data mining and ML and advanced data analytics application based on vast datasets. Data scientists and data analysts should be granted read-only access to this zone but with as any data manipulation or modelling should be conducted on copies of the original data. The raw data should be left un-altered as further different kind of data modelling or processing will be needed for other business needs in the future [2]. Sensitive data must be encrypted.

3- **Trusted Zone:** This zone should have data copies after running QA and validation checks and conducting any required data processing to model the data and make it ready for consumption by downstream application for self-service data analysis tools [2] (i.e. reporting and dashboarding tools) by business end users. This zone should be accessible by business normal end users.

4- **Refined Zone:** This zone should have enriched and manipulated copies of the data. This could be the best places to store the output from data analysis tools that need to write data back.

Encryption

As data could be anywhere (on premises, in cloud, or in both places), and it could be accessed and manipulated from many different points, it is imperative to ensure data security by applying data encryption where data is [3]. Data encryption should be applied for both data at rest and data movement wherever the data is: on premise or on cloud.

Authorization

Applying authorization means – in a nutshell, granting different access permissions to different users to access specific and/or perform specific tasks. in the context of data lakes, however, this need to be carefully planned and executed in order to avoid ending up with data silos all over again and consequently diminish the value of having and enterprise data lake.

Alongside implementing four zones within the data lake, document-level access control could be applied to different datasets in one zone [1]. This access controls could be inherited from the source system from where data was initially ingested from. Furthermore, access controls could be relaxed for some parts of the data lake for some users. For example, we could allow data scientists from one business line to have access to all raw data in the "Raw Zone" related to that business line. this will allow them to run ML models and data mining application on the data they are more familiar with. As a result, this will help in both maintain the value of the data lake and security standards.

IV. CONCLUSION

The concept of data lakes is still on the evolving side. Securing data lakes is not as straight forward as it sound due to special nature of data lake technology, the purpose it is to serve, and the technical details of the data lake implementation. That's possibly the reason there is still no one agreed on industry standard for securing data lakes. In the meantime, administrators should start with the basics and build their security policy with the aim to secure but hinder or harm the value of having the data lake at the first place.

REFERENCES

[1] Maroto, 2020. Data Lake Security.[https://www.searchtechnologies.com/blog/data-lake-security]. Accessed on September 2, 2020

[2] K. Coenye, 2019. Data Lake Security and Governance Best Practices. [https://kohera.be/blog/azure-cloud/data-lake-security-and-governance-best-practices/]. Accessed on September 2, 2020

[3] M. Gualtieri, J. Kindervag, K. Mak et al. 2016. Big Data Security Strategies For Hadoop Enterprise Data Lakes. [https://www.forrester.com/report/Big+Data+Security+Strategies+For+Hadoop+Enterprise+Data+Lakes/-/E-RES122747#]. Accessed on September 2, 2020