International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Data Governance and Security - A Comprehensive Review of Snowflake and Databricks

Rameshbabu Lakshmanasamy

Senior Data Engineer, Jewelers Mutual Group

Abstract: ***Data governance and security are game changers in today's digital world to manage data assets effectively and to comply with regulations and protect sensitive information. Data assets are very crucial for organizations that leverage cloud based platforms for data storage, analysis and management and need stringent governance framework and strong security measures to protect this valuable resource. Two of the most known platforms are Snowflake and Databricks (Zhang, 2024). This paper reviews the Snowflake and Databricks data governance and security features comprehensively, their functional capability, encryption approaches, compliance, and certifications.***

Keywords: Data Governance, Security, Snowflake, Databricks, Compliance, Encryption, Network Security

1. Introduction

Data governance is the set of processes, roles, standards and metrics to enable an organization to use information effectively and efficiently (Janssen et al., 2020). A properly structured governance help organizations maintain data quality, manage data lineage and lets the firm exercise proper control on data access (Shabani et al., 2021). Without it, people cannot have trust in the data and cannot comply with legal and regulatory obligations.

Data governance features of Snowflake

Snowflake has a comprehensive data governance framework that facilitates effective control over data assets, and ensures legal compliance and the integrity of the data at all times. Figure 1 below shows all the snowflakes features.

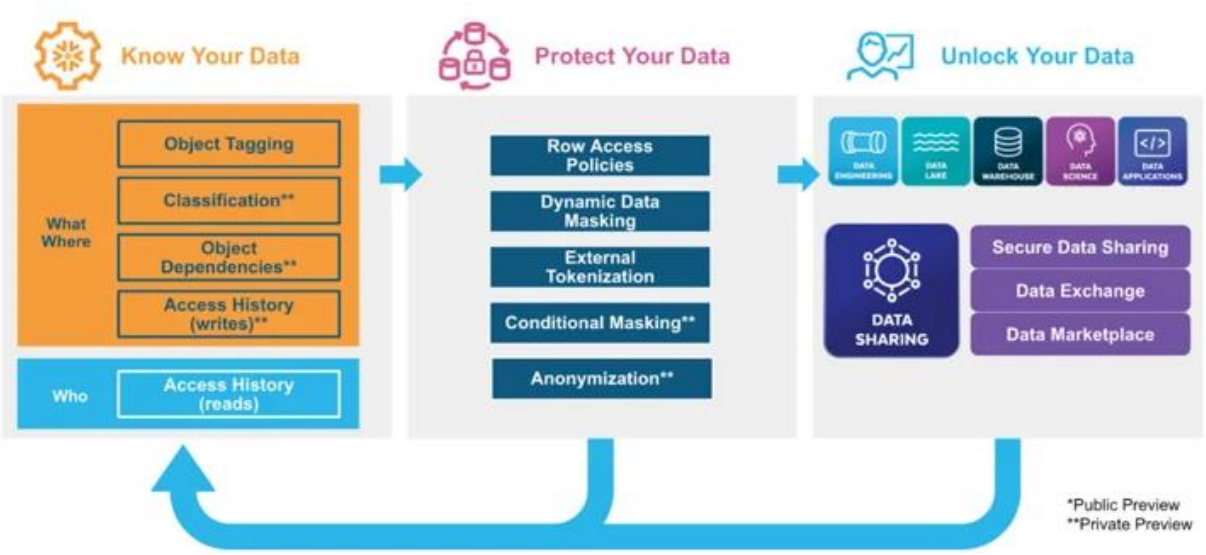


Figure 1: Snowflake's Governance and Security Features Adapted from Dataops

- Data Lineage: Data lineage is tracked across the Snowflake platform, allowing organizations to understand source, movement and transformation of data. Transparency and auditability are critical points for compliance and data quality and this feature makes certain that it happens in a right way.
- Access Controls: Snowflake's role based access control (RBAC) system is flexible and granular. Per Role, data administrators can assign permissions to users, allowing the users to have access to only the information they need. Snowflake also offers MFA and SSO for access security assurance.
- Data Sharing Governance: Snowflake's Secure Data Sharing feature lets organizations share live data with external partners whilst retaining control over the data access. Sharing can be limited by limiting it through encryption, masking the sensitive data, or by providing expiration dates for when access can dwindle.

Data governance features of Databricks'

Databricks' is built on Apache Spark, it is a powerful unified analytics platform with integrated governance features. Figure 2 below shows all the Databricks features.

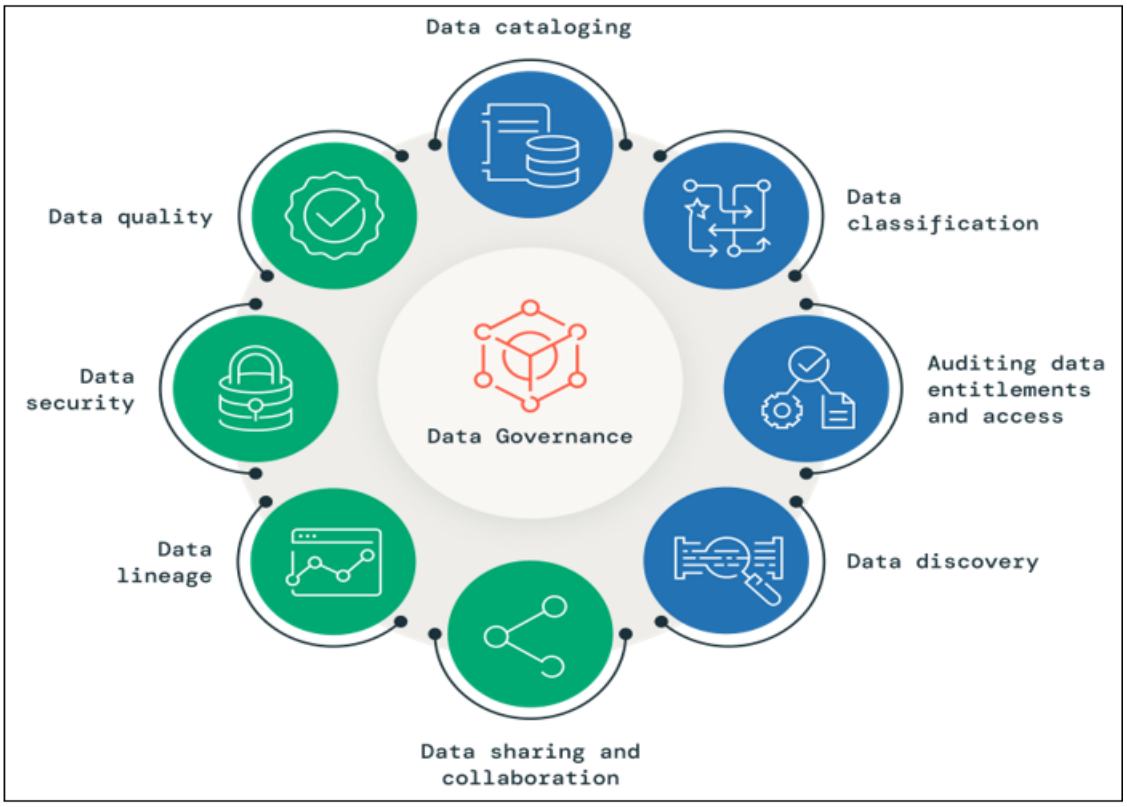
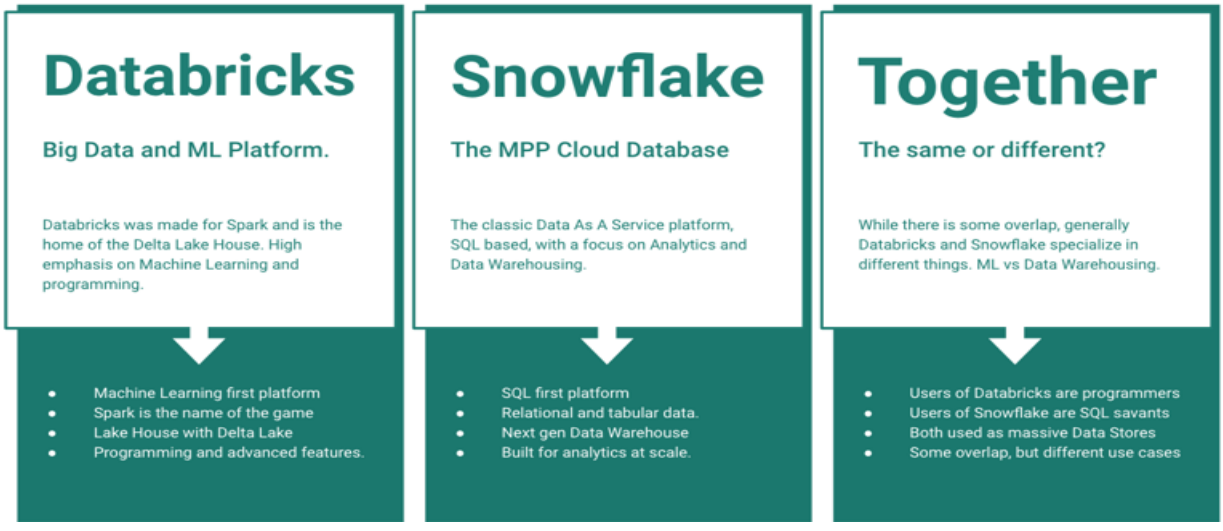


Figure 2: Key Elements of Data Governance Adapted From Databricks'

- **Data Lineage**: With Unity Catalog, all data lineage is tracked by Databricks and data repository is managed by the central repository, which is the Unity Catalog. Data lineage can trace data transformation and dependencies among different sets of datasets and models.
- **Access Controls**: Databricks also enables the role based access control (RBAC). Fine grained access control, view and column level is also provided by the Unity Catalog for the Unity Catalog, which gives data access strong security. Also, Databricks works with identity providers like Azure AD (AAD) for SSO and MFA.
- **Data Quality Management**: Delta Lake provides schema enforcement and auditing and Delta Lake is integrated with Databricks. This means that the data is always high quality, logging out all schema changes and data mutations automatically. And time traveling features of Delta Lake make governance even better with time travel features: We can query historical data versions for increased transparency and auditability.

Comparison of Governance Features:

Snowflake and Databricks both have a full suite of data governance features, however the implementations are slightly different. An advantage to organizations that need to partner with other teams or outside stakeholders via secure data sharing are the focus of Snowflake (Shah, 2024). However, integrating with Delta Lake gives Databricks more granular control over data quality and lets user track data at a more granular level. This both creates two strong data lineage tracking and access control mechanisms and serves for organizations that give precedence to governance and compliance (Koppula, 2022). Figure 3 below shows a comparison between the two platforms.



Security Measures:

Importance of Data Security

The ability of an organization to store and process data is only becoming more reliant on cloud services, which makes security of the utmost importance. Data breaches, unauthorized access and compliance violations can result in financial loss, reputation damage, and legal repercussions. While both Snowflake and Databricks stress security via encryption by these encryption methods, access control mechanisms and compliance certifications (Yulchiev, 2024; Bandari, 2023).

Snowflake's Security Controls

- Encryption: Data captured by Snowflake is encrypted in transit, and at rest employing AES-256 encryption, a well adopted standard in securing data. Continual rotation of encryption keys provides the maximum protection. It also comes bundled with service provided encryption keys (CPK) allowing organizations to control the encryption keys they use which benefits them in terms of security (Kashyap, 2023).
- Network Security: Across the clients, Snowflake retains TLS encryption of the network connections to the Snowflake platform. Network policies are now another feature that does this very same thing, meaning we can restrict access to IP addresses or regions if access isn't desired and use untrusted networks (Kashyap, 2023).
- Compliance Certifications: Snowflake is certified to multiple industry standards and regulations – SOC 1, SOC 2, GDPR, HIPAA, and PCI DSS among others. Snowflake said the certifications are designed to confirm that Snowflake consistently meets the same exacting standards for security and privacy required for so many industries, including healthcare, finance and retail (Kashyap, 2023).

Databricks' Security Controls

- Encryption: Databricks uses AES 256 encryption. Databricks also encrypts each data set using encryption keys via integration with Delta Lake at the file level (L'Esteve, 2022).
- Network Security: Organizations can create secure, private connections to their data environments on Databricks through Virtual Private Cloud (VPC) peering. It helps data to not travel on the public internet, while reducing the chance of interception. Network policies on the Databricks platform also limit access to the platform according to predefined rule (L'Esteve, 2022).
- Compliance Certifications: Databricks itself is also SOC2, GDPR, HIPAA and ISO 27001 compliant. What they are saying with these certifications is that Databricks values high standards of security, privacy and data management (L'Esteve, 2022).

Security Comparison:

Data in transit is encrypted both in transit (AES 256) and at rest (AES 256) in both Snowflake and Databricks. But for those who want full control over how their encryption keys are stored, Snowflake has a provision for an extra customer-managed key (CMK). On the network security side, VPC peering from Databricks furthers more secure networks and IP restriction policies by Snowflake offer the additional layer of defense. Both have similar security certifications which makes them viable companies (L'Esteve, 2022).

Compliance and Certifications

In industries that involve handling sensitive data it is important to comply with data protection regulations. The level of certification of a platform helps you understand whether track with standard legislation and how it manages data safely (Vinnikainen, 2023).

Snowflakes

Snowflake holds several key certifications, including:

- SOC 1 and SOC 2: Snowflake's certifications meant that not only are these checks and balances being done, but that these controls are effective in the case of financial reporting as well as with data security.
- HIPAA: This allows Snowflake to house and digest protected health information (PHI), and is therefore ideal for healthcare organizations.
- GDPR: Snowflake is 100 percent compliant to the European Union's General Data Protection Regulation, which means that it follows all the heavy privacy and data protection requirements that EU citizens demand when subjecting their personal data.
- PCI DSS: As with any other data that is not protected, Snowflake is compliant with the Payment Card Industry Data Security Standard, meaning it can handle and store payment data.

Databricks

Databricks also holds important certifications, including:

- SOC 2: Like Snowflake, the Databricks runs in a SOC 2 certified environment implying strong internal controls for data security.
- HIPAA: Databricks compliance to the HIPAA rules makes healthcare data a possibility on the platform.
- ISO 27001: This is a standard for information security management, which ensured that Databricks is a secure place for handling sensitive data.
- GDPR: Databricks is GDPR compliant as it can manage the data of EU citizens in a secure and privacy conscious way.

Compliance Certifications Comparison

Snowflake's additional PCI DSS affirmation, however, might verify to be a more compelling decision for organizations in finance that process installments data, both of which have a conformity confirmations. If a company has a stringent security requirement, the ISO 27001 certification of Databricks may look attractive due to its assurance of information security management systems.

Conclusion

Snowflake and Databricks have extensive data governance and security features that are in the forefront within cloud data management space. It's snowflake's forte in secure data sharing and marketplace capabilities, whereas data quality and advanced analytics are provided by Databricks through Delta Lake. Strong encryption, access controls and compliance certification is implemented across all platforms and they are both fit for purpose for organizations that either have very stringent data security and regulatory requirements

or just want to ensure sensitive customer data is secure. Ultimately, whether an organization needs to train across teams for modeling or charting purposes, the need to integrate third-party ingested data like a DQ source or the need to interact with stream data, the primary function an organization requires from a BI tool will determine the advantage between Snowflake vs Databricks.

References

[1] Bandari, V. (2023). Enterprise data security measures: a comparative review of effectiveness and risks across different industries and organization types. International Journal of Business Intelligence and Big Data Analytics, 6(1), 1-11.
[2] Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. Government information quarterly, 37(3), 101493.
[3] Kashyap, R. (2023). Data Sharing, Disaster Management, and Security Capabilities of Snowflake a Cloud Datawarehouse. International Journal of Computer Trends and Technology, 71(2), 78-86.
[4] Koppula, R. S. (2022). Implementing Data Lakes with Databricks for Advanced Analytics. North American Journal of Engineering Research, 3(2).
[5] L'Esteve, R. (2022). Databricks. In The Azure Data Lakehouse Toolkit: Building and Scaling Data Lakehouses on Azure with Delta Lake, Apache Spark, Databricks, Synapse Analytics, and Snowflake (pp. 83-139). Berkeley, CA: Apress.
[6] Shabani, M., Thorogood, A., Murtagh, M., Laurie, G., Dove, E., & Ganguli-Mitra, A. (2021). Data access governance. Biopreservation and Biobanking, 14(3), 231-240.
[7] Shah, S. T. U. (2024). Optimizing Data Warehouse Implementation on Azure: A Comparative Analysis of Efficient Data Warehousing Strategies on Azure.
[8] Vinnikainen, O. (2023). Data mesh: a holistic examination of its principles, practices, and potential.
[9] Yulchiev, K. (2024). The importance of data security today. Texas Journal of Multidisciplinary Studies, 33, 9-14.
[10] Zhang, K. (2024). Incorporating Deep Learning Model Development with an End-to-End Data Pipeline. IEEE Access.