Data Access Governance: Managing Risks and Enhancing Security in Snowflake, Databricks, Azure Synapse, and AWS Redshift

Theom, Copyright 2023

Introduction:

It all starts with proper data governance. AI needs data, and data needs data governance. Proper data governance can protect an enterprise against substantial risks and improve the quality of any analytics or AI. At its core, data governance focuses on security, risk, and compliance, especially at data aggregation points like data warehouses and data lakes where there are still substantial security blind spots. This e-book provides an overview of data access governance and its critical importance. It also offers easy-to-follow checklists to detect current exposures and risks while showing measures for risk mitigation.

Frequently described as the “iPhone moment for AI,” Large Language Models (LLM) have brought about an enormous focus on AI. However, AI models derive value from the underlying data they get trained on. The owners of high-quality data will benefit the most from AI, while AI models are already increasingly open source to drive transparency and faster innovation. This puts further emphasis on data and its governance.

Data governance provides the “traffic” rules, standards, and enforcement mechanisms for data collection, integration, and transformation. It ascertains that data is secure, accurate, complete, reliable, and auditable. Data governance provides the foundation for the data lifecycle spanning data collection to data exchange and AI models. Thinking through and contractually defining governance rules for the data is frequently where the data journey starts. While data governance comprises several disciplines, the most important aspects focus on security, privacy, ownership, and compliance, i.e., the twin pillars of “What data?” and “Who has access?”. Exposing sensitive data can have catastrophic consequences for an enterprise. At the same time, breaking compliance rules can lead to significant financial fines (Data regulatory violations can be fined up to 4% of annual revenue) or even jail for responsible individuals.

At the macro level, platforms for data governance ideally cover all data platforms under one pane of glass so that rules can be universally & consistently applied/updated and risk assessed holistically while operating synergies are realized. In comparison, the dedicated data governance offerings embedded in specific data warehouses, - lakes, etc., will always be at a disadvantage.

The best data governance will only mean a little without proper enforcement. Data governance platforms will need to ingest the complete framework of existing governance rules and an exhaustive inventory of all data assets, identities, access privileges, and their relationships. The unauthorized data replication for development, testing, or acceleration purposes seems too familiar. If not detected, this shadow/dark data can fly under the data governance radar, creating substantial risk. The scope of this discovery is typically not limited to intra-enterprise but also frequently extends to inter-enterprise as the exchange of data with trusted partners in so-called clean rooms becomes standard. The discovery process should run without agents or proxies to avoid additional latency or performance bottlenecks. It is also important that the discovery data is subjected to the same governance rules as the underlying data. Exporting data or metadata to external systems outside a customer’s jurisdiction usually raises concerns.

Not all data is created equal. The value of data is ultimately determined by the business cases they feed into. However, independent of business cases, certain data enjoys special protection by law/compliance rules, given their sensitive nature. Frequently, the sensitivity is more than just a one-dimensional construct, as in the case of GDPR, where the location of the data is of critical importance. Modern data governance platforms automatically classify data and assign values based on deep reads and NLP while also showing the lineage of the data.

Theom, Copyright 2023

Conflating data sources with their lineage, value, and risk profile will highlight critical risks and guide remediation efforts.

The questions “What data?” and “Who has access?” become even more complicated when the aperture is narrowed from the holistic to the atomic level, as the unit of labor for data access governance is rarely an entire data warehouse but individual cells or columns. Defining access rules down to individual cells has become necessary (dynamic masking), not just in enterprises but inter enterprises where hundreds of millions of records are exchanged in clean rooms by allowing customers unique and fine-grained access to pieces of the same data set. Access rules are rarely static: employees and customers turn over, organizational charts change, and business processes get reorganized - updating systems is frequently challenging. Additionally, access rights often get defined rather loosely, i.e., employees receive access based on a role or their position in the organization chart and not based on the data itself and the business outcomes it drives. Every overprovisioned account is a substantial security risk as it increases the attack surface. However, sophisticated data access governance platforms go significantly beyond a simple overview of overprovisioned accounts: failed or unauthorized logins get detected. At the same time, weak authentication methods and infrequent and privileged access also trigger alerts. In addition, the scope of data access governance should also get extended into a review of the source code and APIs. Seeing passwords in the clear embedded in source code is hardly surprising anymore and is a clear manifestation of a lack of governance. Scanning APIs will also expose potential abuses of access privileges, for instance, when humans try to trick systems into believing they are machines to get easier access.

How can we defend our data against insider threats? What if the Who in “Who has access?” starts to exhibit abnormal behavior trying to cause damage? It could be an insider or an outsider who finds ways to hack into somebody’s account and is now out to exploit the assumed identity. Both cases will require extensive behavior baselining to establish what is normal so that deviations can be spotted immediately and shut down. AI technology typically plays a major role in establishing these baselines. Additionally, certain standard patterns prepare for data exfiltration. Even before an attempt to exfil is made, monitoring and instant detection of these patterns are critical. Furthermore, sophisticated data governance will detect any data loss immediately while alerting on the placement of ‘sleepers’ and ransomware or the sequential testing of access privileges by assuming different roles.

Every major enterprise has dozens of security products, and consolidating technologies into increasingly comprehensive platforms offered by a handful of vendors is a significant trend. What are the current offers to address all the above-referenced problems in the seemingly unending labyrinth of security products that all come with sophisticated acronyms? The surprising answer is that there aren’t any solutions, as traditional security has been focused on strengthening the perimeter, i.e., keeping bad actors out. However, as illustrated above, data governance will require an insider’s view. As data governance is closely attached to the data, it is only logical to have data governance and enforcement sit next to it. Co-location with the data will also avoid data transfers (including metadata, increasing the overall

system's security, being part of the overall solution, and not a problem (wherein you will need one more tool to understand the posture of egress).

Satisfying regulatory audit and compliance requirements has become very time-consuming, especially in highly regulated industries like financial services, where a single company typically has to report to several regulators. Reporting requirements can also change with the size and scope of the business. Newer efforts like DORA in Europe will further highlight the importance of data governance and its documentation. Audit and compliance use cases offer significant potential for cost savings by introducing automation. Most data access governance platforms offer varying degrees of audit and compliance reporting capabilities out of the box. The significant benefits of automating these processes will easily offset the effort to customize and expand existing capabilities. Another use case that has been becoming more popular is the so-called “legal hold,” i.e., keeping/preserving data for the satisfaction of legal holding periods while the data is no longer used for business purposes, only for potential evidence. This is a use case that most data governance platforms should accommodate, especially when the data is kept in “cold” storage for cost purposes.

One final consideration should also be the cost and the mode of operating the data governance system. Algorithms and their implementation vary significantly across vendors. We have seen scenarios where one vendor could accomplish more with one smallest Snowflake instance than another running on 6 XL instances. Using the PoC to also get a good sense of the operating costs of any data governance platforms under consideration pays off. Similarly, data governance platforms differ significantly regarding operating costs caused by manual labor. Not every platform is fully automated, 24/7, with automated alerts and remediation. Some still require lots of manual intervention, especially when the focus is more on monitoring and alerts vs. prevention and auto-remediation. The amount and quality of intelligence (AI and rule-based) embedded in a governance platform to propose and take remedial actions will be inversely correlated to the time and money spent on war rooms and MTTR. Ideally, a data platform learns from the data at rest and the data flows of each customer to progress towards higher degrees of automation. The vision is clearly to have all data governance run invisibly in the background enabling businesses to run securely and in compliance with all requirements.

Chapter 1: Understanding Data Security Blindspots

1.1 Inadequate Data Classification

Data classification is a fundamental aspect of Data Access Governance. Security teams must identify the most sensitive data and classify it according to its sensitivity. Applying appropriate access controls based on this classification enhances security awareness and protection. Custom taxonomies, specific to enterprises and industries, must be well understood. Remember, good security starts with good visibility.

1.2 Beyond Authentication and Authorization

While robust authentication methods like Single Sign-On (SSO) and Multi-Factor Authentication (MFA) are essential, Data Access Governance goes further. Organizations should enforce the principle of least privilege, ensuring users have access ONLY to the necessary resources. This simple step can help to avoid data abuse.

1.3 Insecure Data Sharing

Data sharing is a significant aspect of the new data economy. However, companies should exercise caution when sharing data, Python jobs, notebooks, or other resources. It is crucial to understand the contextual aspects of what data is being shared, its significance, and whether the share is or can be time-bound. Incorrect data sharing can devastate a company's business and reputation.

1.4 Insider Threats

Tools and processes to address insider threats are vital. Questions to consider include: Have users been phished? Are privileged users dumping out data? If so, why and where is the data going? Can you specifically detect exfiltration activities? What steps are in place to address and prevent such attacks?

1.5 Insufficient Monitoring and Auditing

Regular monitoring and auditing of your cloud data warehouse environment are necessary to detect suspicious activities and potential security issues. Classifying which behaviors are typical and which are abnormal is a crucial aspect of Data Access Governance.

1.6 Misconfigurations

Misconfigurations can expose your data and environment. Validating correct configuration and continuously monitoring for deviations is necessary. Misconfiguration doesn't only happen in the access or rights settings, and misconfigured notifications are also common.

1.7 Insecure Access Practices and APIs

It's common to see data validation scripts, Python code/Notebooks/Client SQL Scripts that contain sensitive information. Practicing good hygiene with code that handles sensitive data is essential. Similarly, when using APIs, ensure that proper authentication and authorization are in place and validate all incoming data.

Chapter 2: Data Exfiltration from Cloud Data Warehouses

2.1 Cloning

Cloning allows a user to create a complete copy of a database, schema, or table. This feature can extract data from warehouses by cloning a specific database or schema and then using that copy to extract the required data. Zero-copy cloning makes it even easier because it does not require deep copies of the original data tables.

2.2 Data Marketplace

A data marketplace is a platform feature that allows companies to “publish” and customers to discover and access third-party data sets that have been curated and optimized for use within the data warehouse or with data marketplace applications. This data can be published natively in the data marketplace.

Theom provides secure, clean rooms and data shares, allowing enterprises to share and innovate with partners and third parties confidently. Theom secures clean rooms and data shares so enterprises can confidently share and innovate on their data with first parties and vendors.

Securing clean rooms and data shares, particularly in the context of data lakes for collaborative analysis, involves implementing several measures to prevent data abuse and maintain security. Here are some key considerations that Theom addresses in securing clean rooms:

2.2.1 Shares/Direct Shares governance

With Theom, enterprises can understand what shares are being used, what data is in those shares, and who is creating the shares. Governing shares is a significant gap that Theom addresses.

2.2.2 Access Controls

Theom implements strong access controls to ensure that only authorized individuals have access to the clean room environment. Using Theom, enterprises can review and understand the reality of authorization privileges: who is accessing the data and who is not.

2.2.3 Data Masking

If fields have to be accessed in a masked or anonymized manner, using Theom, enterprises can ensure that no clear text accesses happen. Anonymization techniques and masking

remove personally identifiable information (PII) or other sensitive data elements. If the clean rooms have to adhere to masked access, Theom ensures that all access complies.

2.2.4 Data Loss Prevention (DLP)

Theom employs DLP measures to monitor and prevent the unauthorized exfiltration of sensitive data from the clean room environment. This includes techniques such as data leakage monitoring, data classification, and data loss prevention policies.

2.2.5 Data access audit/monitoring

Enterprises can monitor all accesses and track suspicious activity for any data shares or data designated for clean rooms.

2.2.6 Compliance and Regulations

Theom ensures that the clean room environment adheres to relevant compliance standards and regulations, such as HITRUST, HIPAA, NIST, or industry-specific regulations. Enterprise can regularly review and update security controls to maintain compliance using Theom's access governance rules engine.

Theom supports data shares, clean rooms on Snowflake, and delta shares on Databricks. With sharing of data bring secured, enterprises can collaborate on building new business models and confidently partner within their organizations and 3rd parties.

2.3 "COPY INTO" Function or Similar

This function allows you to export data from a cloud warehouse to a file in a cloud storage location, such as Amazon S3, Azure Blob Storage, or Google Cloud Storage.

2.4 "UNLOAD" Function or Similar

This function allows you to export data from cloud data warehouses to an external stage, such as Amazon S3, Azure Blob Storage, or Google Cloud Storage.

2.5 JDBC/ODBC Drivers

You can use these drivers to connect to the cloud warehouse and extract data using SQL queries. You can then write the results to a local file or any other destination.

2.6 REST API

You can use a cloud data warehouse REST API to extract data in a JSON format via simple API execution or scripting.

2.7 Third-Party Tools

Many third-party tools support cloud warehouses and can be used to extract data. For example, you can use tools like Apache NiFi, Apache Airflow, or Talend to extract data.

2.8 Partner Ecosystem

Partner ecosystem that includes data integration and ETL tools that can be used to extract data. Examples of such tools include Informatica, Matillion, Fivetran, and many others.

Chapter 3: Managing Risks Associated with Data Exfiltration

With so many ways to take data out of your cloud data warehouse, it's essential to understand the risks associated with these methods. Users of even the most critical data stores can be phished, which allows attackers to use any of the above methods to exfiltrate data. SQL injection attacks can also exploit applications built on cloud data stores. It’s also not uncommon for misconfiguration of user rights, roles assignment, or privileges to allow for increased risk exposure in these vast data warehouses.

3.1 Awareness of Data Existence and Value

Security teams should be aware of what data exists and what are the crown jewels from a data perspective. They should also be aware of the value of the data and the related economic risks.

3.2 Understanding Data Access

Understand who accesses these data tables and the context involved in that access. Users' behaviors are a good giveaway of risks.

3.3 Insider Risk Management with Theom

Mitigate insider risks before they become threats with Theom’s native attack detection modules for Snowflake, Databricks, and Azure. Theom is the industry’s only holistic data-centric access governance platform designed to continuously evaluate all access of your trusted workforce and applications on cloud data stores.

Using machine learning, the MITRE ATT&CK framework, Theom natively maps abnormal access activity into cloud data warehouses and lakehouses. Theom enables you to gain data visibility,

operationalize best practices, intervene early before an incident occurs, and advance your insider risk management program alongside your access governance and security initiatives.

Theom seamlessly integrates contextual, human behavioral, and security monitoring to provide a single pane of glass—through which you can examine insider risks from every possible angle. Theom unifies insider risk management across clouds, mapping into MITRE ATT&CK framework and detecting malicious activity natively on data stores.

Theom detects malicious use of native Snowflake capabilities, including load, query, clone, share, and export, to prevent data breaches. Understand how unity catalog and data sharing from Databricks can be protected natively. Abnormal accesses on your data pipeline identities or atypical user behaviors querying parquet with SparkSQL are protected to prevent data breaches.

Theom also enables access governance on Azure data stores. Integrated with Azure AD and Azure Sentinel, with Theom, you can detect and respond to data breach attempts working within the extensive Azure SIEM/SOAR/Identity infrastructure.

Theom is embedded within your data lakes and warehouses and does not transfer any data out. Theom does not use a proxy or cross-account roles to mitigate insider risks. By operationalizing best practices in the context of understanding data, you can now manage insider risk and avoid large-scale incidents.

Chapter 4: Mitigating Data Security Blindspots with Theom

The easiest way to gain contextual knowledge and visibility is to deploy Theom onto your cloud data warehouse instance. Theom deploys in minutes, requires no agents or proxies, and has no impact on data warehouse performance. Theom supports Snowflake, Databricks, AWS, and Azure data stores.

Staying up-to-date with your cloud warehouse vendor’s recommendations and best practices is always good practice. Continuously reviewing and improving your organization's security posture can eliminate many security blind spots.

Remember, recovering from a data breach event is costly and damaging for any organization. It is better to avoid and mitigate the risk of fines, reputation loss, and trust erosion rather than always waiting for something to happen and reacting to the situation.

In the end, effective Data Access Governance is about proactive management of data access and security, ensuring that your organization's data remains secure and accessible only to those who need it.

Theom offers an efficient solution with low operational costs and no need for operational staff retraining. It provides efficient and cost-capped data classification, eliminating hidden data transfer costs or egress charges. Theom offers value within 2-3 hours with a fast and fully automated setup process. It also integrates seamlessly with existing tools, offering supported modules for IT Service Management (ITSM) platforms such as ServiceNow and Jira and Security Information and Event Management (SIEM)/Security Orchestration, Automation, and Response (SOAR) systems like Splunk and Azure Sentinel.

Theom is embedded within your data lakes and warehouses and does not transfer any data. Theom does not use proxy or cross-account roles to secure cloud warehouses. By deploying Theom, you can operationalize best practices in the context of data, access governance, manage insider risks, and avoid large-scale incidents.