


Leveraging Databricks engineering to build a robust data governance framework

Arun Khandelwal, Enterprise Data Solution Architect,
Databricks Partner Solution Architect Champion at
DXC Luxoft

Table of contents

Introduction	3
Challenges of data governance in traditional data architectures	4
Solution overview	6
Creating a data governance framework	6
Key components of a data governance framework	7
Value proposition of using Databricks engineering for data governance	9
Technical details	12
Solution approach	12
Reference architecture	12
Solution components & tools	13
Bringing it to life	15
Best practices	16
Security considerations	18
Benefits	20
Technical benefits	20
Business benefits	21
User benefits	21
References	22
Case study: Implementing data governance at an asset management company	23
Background	23
Implementation	23
Outcomes	23
Conclusion	24
Key benefits	25
Strategic roadmap	26
About the author	27



This whitepaper explores leveraging Databricks engineering to build a robust data governance framework, addressing the challenges of traditional models and ensuring data quality, security, and accessibility. By utilizing Databricks' data intelligence platform, organizations can transform data governance into a strategic advantage.

Introduction

In today's data-driven world, organizations collect massive amounts of data from diverse sources such as customer interactions, financial transactions, social media, and IoT devices. While this data offers immense opportunities for insights and innovation, it also presents significant challenges. Without proper governance, data can become a burden, making it difficult to access, manage, and trust, leading to inefficiencies, increased risks, and missed strategic opportunities.

Data has become a critical asset in the modern enterprise landscape, driving decision-making and providing competitive advantages. Effective data governance ensures data integrity, security, and compliance with regulatory requirements, thereby enhancing customer experiences, optimizing operations, and generating new revenue streams.

Databricks & Unity Catalog, a data intelligence platform, provides a powerful environment to create a comprehensive data governance framework. It integrates data engineering, data science, machine learning,

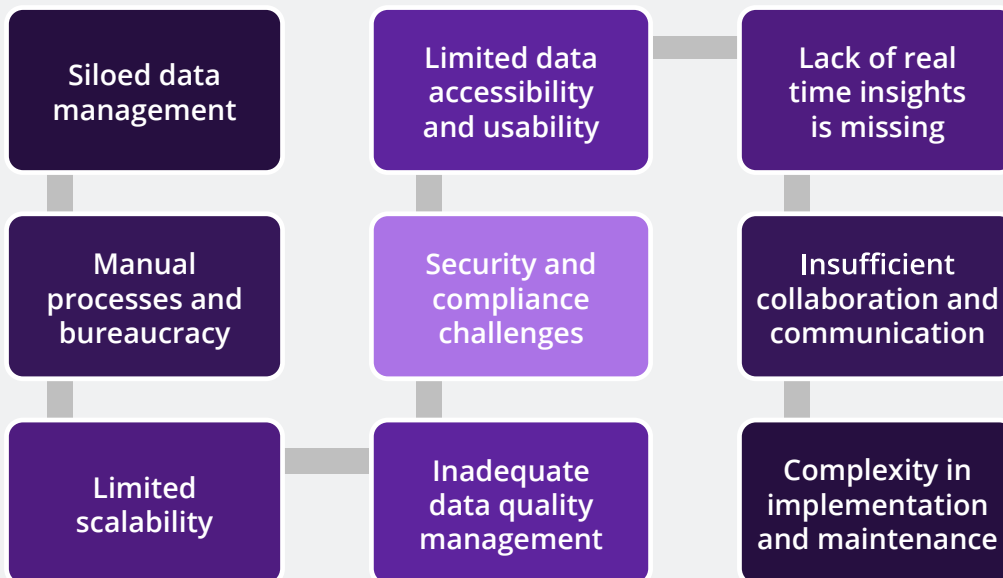
and analytics, enabling organizations to manage their data lifecycle efficiently. This whitepaper explores how to leverage Databricks & Unity Catalog to build and implement a robust data governance framework, outlining key components, capabilities, and practical strategies for ensuring data quality, security, and accessibility.

Through detailed case studies, best practices, and a solution approach, this whitepaper equips data professionals and business leaders with the tools to establish a solid data governance framework approach using Databricks & Unity Catalog. This approach mitigates data management risks and unlocks the full potential of data assets, driving innovation and sustaining competitive advantage in a data-centric world.

Challenges with data governance in traditional data architectures

Successful data governance programs are characterized by two key attributes: a robust, scalable process framework and an architectural data management foundation which is open, scalable, and flexible.

Companies struggle with their data governance efforts because of traditional, central process approaches and a data foundation which exacerbates the problems, rather than supporting the governance efforts. As a result, organizations often struggle to keep pace with the dynamic needs of modern data-driven organizations. Here are some key challenges faced:



1. Siloed data management

- **Fragmented data sources:** Traditional governance frameworks often have to deal with data spread across various departments and systems, leading to fragmented data sources that are difficult to manage cohesively
- **Lack of integration:** Integrating data from multiple sources is challenging, resulting in inconsistent data formats, standards, and quality. This hinders the ability to gain a unified view of organizational data

2. Manual processes and bureaucracy

- **Time-consuming processes:** Traditional data governance relies heavily on manual processes for data classification, validation, and reporting. This makes data management time-consuming and prone to errors
- **Bureaucratic overhead:** Rigid governance structures with multiple layers of approvals can slow down data access and usage, reducing agility and responsiveness to business needs

3. Limited scalability

- **Inability to handle big data:** Traditional approaches often struggle to handle the volume, velocity, and variety of big data generated in modern enterprises. Scaling governance processes and tooling to accommodate large datasets is challenging
- **Resource constraints:** Limited resources and outdated technologies can impede the ability to scale data governance efforts effectively

4. Inadequate data quality management

- **Inconsistent data quality standards:** Ensuring consistent data quality across disparate systems and departments is difficult. Inadequate data quality management can lead to inaccurate insights and poor decision-making
- **Reactive approach:** Traditional data governance often takes a reactive approach to data quality issues, addressing problems after they occur rather than proactively preventing them

5. Security and compliance challenges

- **Evolving regulatory:** Keeping up with rapidly evolving data privacy regulations (e.g., GDPR, CCPA, EU AI Act) is challenging and time consuming. Traditional approaches may not be flexible enough to adapt quickly to new compliance requirements
- **Insufficient security measures:** Traditional data governance implementations may lack robust security measures to protect sensitive data, increasing the risk of data breaches and unauthorized access

6. Limited data accessibility and usability

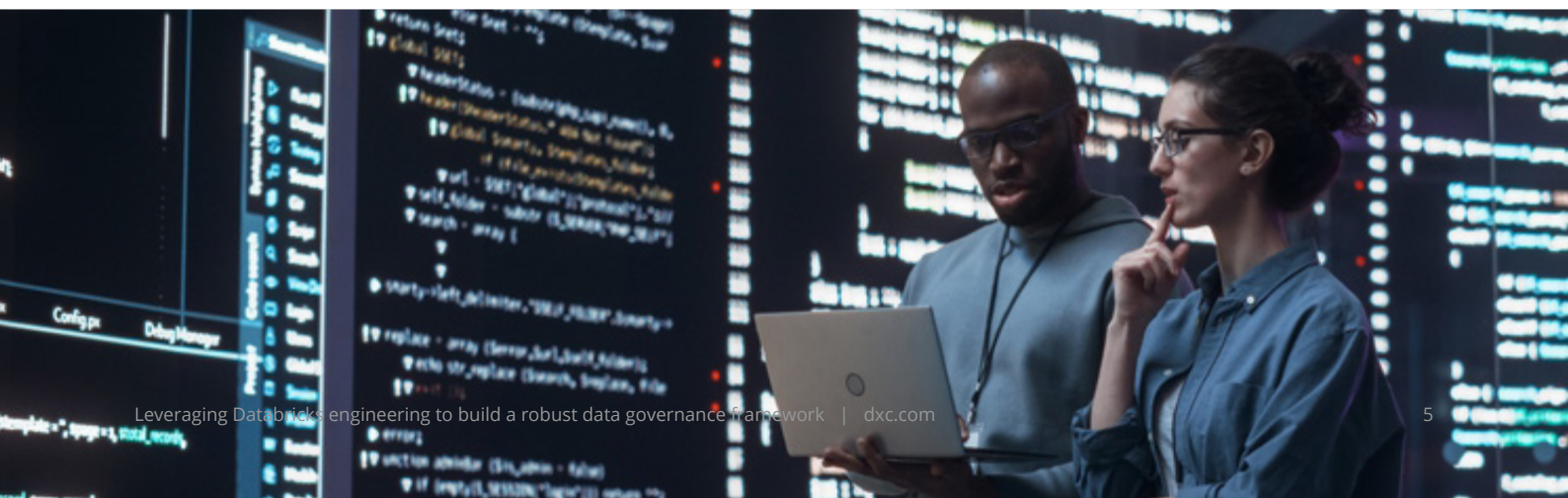
- **Access restrictions:** Traditional governance models often impose stringent access controls that can restrict legitimate data usage by authorized users, hindering innovation and data-driven decision-making. Often this is due to the complexity and efforts related to managing the access itself
- **Complex user interfaces:** Complex and non-intuitive user interfaces can make it difficult for business users to access and utilize data effectively

7. Insufficient collaboration and communication

- **Isolated teams:** Data governance responsibilities are often siloed within IT or specific departments, leading to poor communication and collaboration between data stewards, analysts, and business users
- **Lack of shared understanding:** Differing interpretations of data governance policies and standards can create confusion and inconsistencies in data management practices

8. Complexity in implementation and maintenance

- **High implementation costs:** Implementing traditional data governance frameworks can be costly and resource-intensive, requiring significant investment in technology, processes, and personnel.
- **Ongoing maintenance:** Maintaining and updating data governance policies, standards, and procedures can be cumbersome and time-consuming, especially as data environments evolve



Solution overview

A robust data governance framework is essential for ensuring the integrity, security, and accessibility of data across an organization. It also forms the foundation of any successful AI program that will deliver consistent business value.

As we discussed, the most problems with data governance today can be attributed to either insufficient capabilities of the underlying data engineering capabilities and a robust process framework that is sufficiently automated and/or simplified to be executed on a consistent basis.

Creating a data governance framework

DXC provides a flexible Data Governance Framework (DGF) from which individual governance process can be derived while Databricks, a data intelligence platform, provides the necessary tools and capabilities to implement the framework and fill it with life.

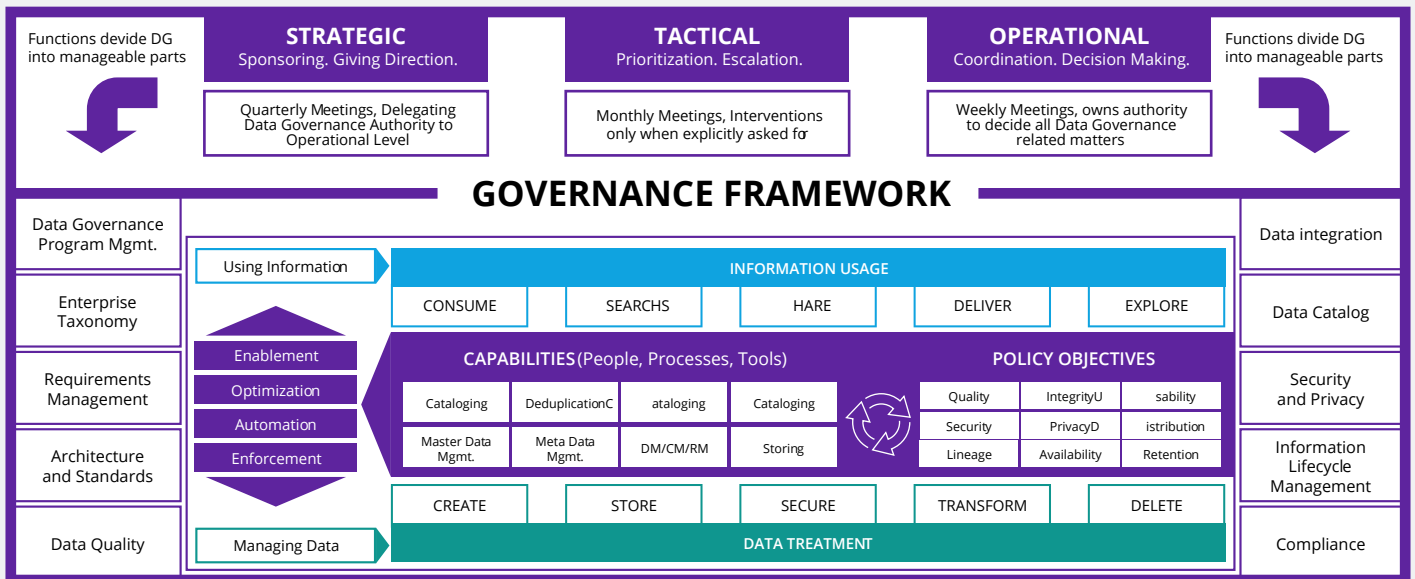


Figure 1 - DXC Data Governance Framework

The DGF is discussed with clients and then evolved into a client specific version of it. DXC's Data Governance Framework helps to structure these efforts and ensures that all data initiatives follow a common, repeatable method and structure. The framework comprises

Governance Levels and Governance Functions, Policies, Processes, Capabilities, Roles & Responsibilities. It supports breaking down this complex undertaking into manageable packages.

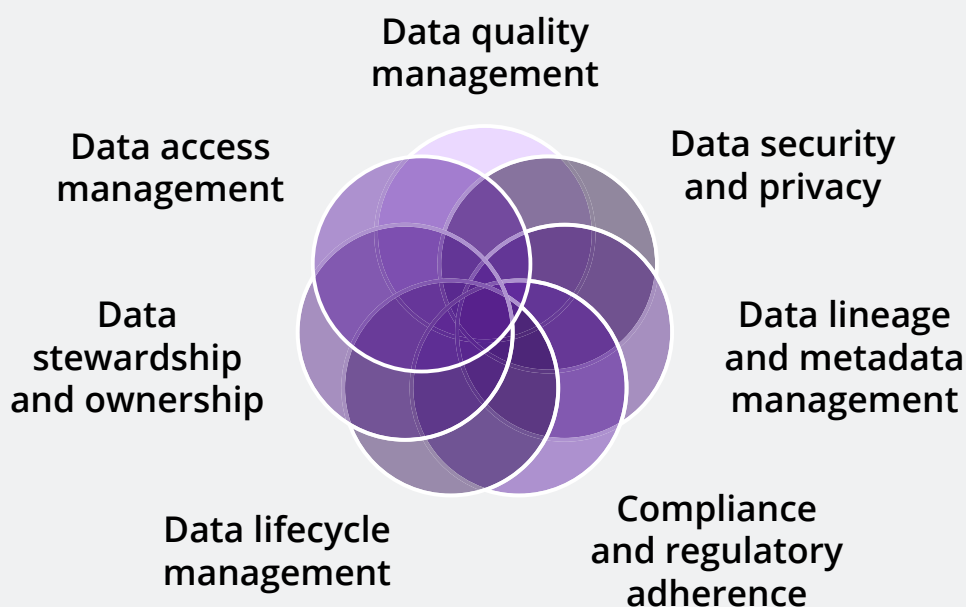
It is important to understand, that successful data governance has strategic, tactical, and operational components and is definitely not limited to technology choices and implementations. Just finding the best Data Governance tooling will not be good enough. On the other hand, just having good processes, roles and

ownership defined, is also not sufficient. Successful organizations balance both sides of the coin.

For the remainder of this paper let's focus on how Databricks provides all the functionalities necessary to realize a DGF and bring it to life.

Key components of a data governance framework

A robust data governance framework typically encompasses approaches for the following key components.



1. Data quality management

- **Data profiling and cleansing:** Data teams who utilize Databricks' powerful processing capabilities to profile, clean, and transform data at scale working on a cluster running DBR 9.1 or newer have two ways to generate data profiles in the Notebook: via the cell output UI and via the dbutils library. This involves detecting anomalies, validating data against predefined rules, and ensuring consistency across datasets
- **Automated quality checks:** Implement automated data quality checks using Databricks' Delta Lake, which supports ACID transactions and enables time travel for auditing changes in data.



2. Data security and privacy

- **Encryption and access controls:** Leverage Databricks' integration with cloud providers to ensure end-to-end encryption of data at rest and in transit. Implement fine-grained access controls using role-based access control (RBAC) and attribute-based access control (ABAC) features.
- **Data masking and anonymization:** Implement data masking techniques and anonymize sensitive data using built-in functions in Databricks, ensuring compliance with data privacy regulations like GDPR and CCPA.

3. Data lineage and metadata management

- **Data lineage tracking:** Use Databricks to automatically track data lineage from source to destination. This includes documenting data transformations, aggregations, and movements to ensure transparency and traceability.
- **Unified metadata repository:** Leverage Databricks' integration with tools like unity catalog to maintain a unified metadata repository. This helps in maintaining a comprehensive catalog of data assets, their schemas, and usage patterns.

4. Compliance and regulatory adherence

- **Audit trails:** Implement audit trails in Databricks (Delta-Lake) to log all data access and modification activities. This helps in demonstrating compliance during audits and investigations.
- **Compliance monitoring:** Use Databricks' lake house monitoring and alerting capabilities to continuously monitor data for compliance with industry-specific regulations and internal policies.

5. Data lifecycle management

- **Data retention policies:** Define and enforce data retention policies using Databricks' unity catalog, ensuring data is retained and purged in compliance with organizational and regulatory requirements.
- **Automated archival:** Automate the archival process for outdated or infrequently accessed data, reducing storage costs while maintaining data availability when needed.

6. Data stewardship and ownership

- **Data stewardship roles:** Establish clear roles and responsibilities for data stewards within Databricks, ensuring accountability for data quality, security, and compliance.
- **Collaboration and documentation:** Use Databricks' collaborative features to document data stewardship practices, data dictionaries, and governance policies, fostering a culture of data ownership and accountability.

7. Data access management

- **Granular access controls:** Implement granular access controls in Databricks' **access control lists (ACLs)** to restrict data access based on user roles and responsibilities. This ensures only authorized personnel can access sensitive data.
- **Self-service data access:** Enable self-service data access for business users through Databricks, reducing bottlenecks while maintaining control over data access and usage.

Value proposition of using Databricks engineering & Unity Catalog for data governance

Leveraging Databricks engineering for building a robust data governance framework offers numerous benefits that address the complexities and demands of modern data environments. Below are some of the key benefits

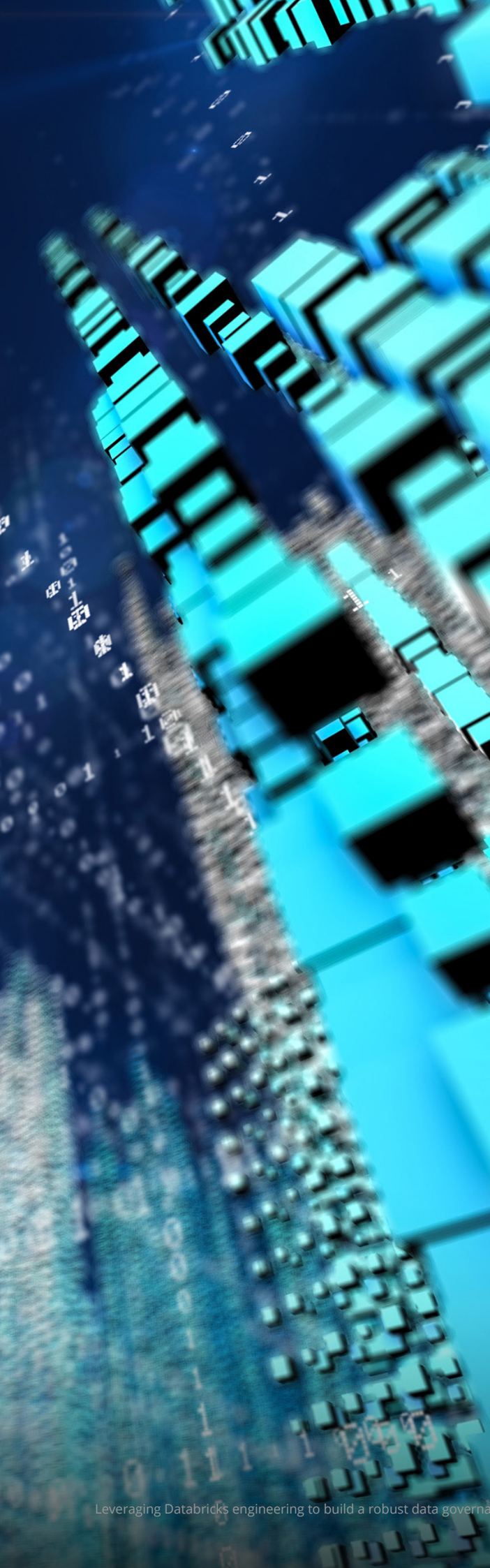
1. Data intelligence platform

- **Integrated environment:** Databricks provides a Unity Catalog that integrates data engineering, data science, and machine learning (ml-flow). This cohesion ensures that data governance policies are applied consistently across all data-related activities.
- **Streamlined workflows:** Databricks Workflows offers a unified and streamlined approach to managing Data, BI, and AI workloads. It can define data workflows through the user interface or programmatically – making it accessible to technical and non-technical teams. The platform's simplicity, reliability, ease of authoring, and price point (it is FREE!) empowers organizations of all sizes to tackle the challenges of data orchestration efficiently.

2. Scalability and performance

- **Elastic scaling:** Databricks' cloud-native architecture allows for elastic scaling, accommodating large volumes of data without compromising performance. This ensures that governance processes are efficient and can handle varying data loads.
- **High performance:** Optimized for high performance, Databricks enables real-time processing and analysis, crucial for maintaining up-to-date data governance.





3. Automated data quality and compliance

- **Automated data quality checks:** Databricks facilitates the implementation of automated data quality checks, ensuring that data remains accurate, complete, and reliable. These checks can be integrated into data pipelines, providing continuous monitoring and validation.
- **Regulatory compliance:** Databricks offers features that help organizations comply with various data regulations (e.g., GDPR, CCPA). By automating compliance checks, Databricks ensures that data handling practices are in line with legal requirements.
- **Data Lineage and Audit Logs:** Tracks every transformation, access request, and change to datasets, ensuring complete transparency for compliance and auditing purposes.
- **Role-Based Access Control (RBAC):** Ensures that only authorized users have access to specific datasets, preventing unauthorized access.

4. Enhanced security

- **Robust security measures:** Databricks provides robust security features, including encryption, access controls, and auditing. These measures help protect sensitive data and ensure that only authorized users can access it.
- **Fine-grained access control:** With Databricks, organizations can implement fine-grained access controls, defining specific permissions for different users and roles. This enhances data security and governance.

5. Real-time monitoring and alerts

- **Real-time monitoring:** Databricks enable real-time monitoring of data pipelines and processes. This allows for immediate detection and resolution of issues, ensuring the integrity and reliability of data.
- **Alerting mechanisms:** Customizable alerts can be set up to notify stakeholders of any anomalies or compliance breaches, facilitating proactive governance.

6. Collaboration and data stewardship

- **Collaborative environment:** Databricks promotes collaboration among data engineers, data scientists, and business analysts. This collaborative environment ensures that data governance policies are well-informed and effectively implemented.
- **Enhanced data stewardship:** By providing tools for better data management and documentation, Databricks enhances data stewardship, ensuring that data assets are well-governed and maintained throughout their lifecycle.

7. Cost efficiency

- **Optimized resource utilization:** Databricks' scalable infrastructure allows for optimized resource utilization, reducing the costs associated with data processing and storage.
- **Cost transparency:** Built-in cost management features help organizations track and manage expenses related to data governance, ensuring cost-effective operations.

8. Advanced analytics integration

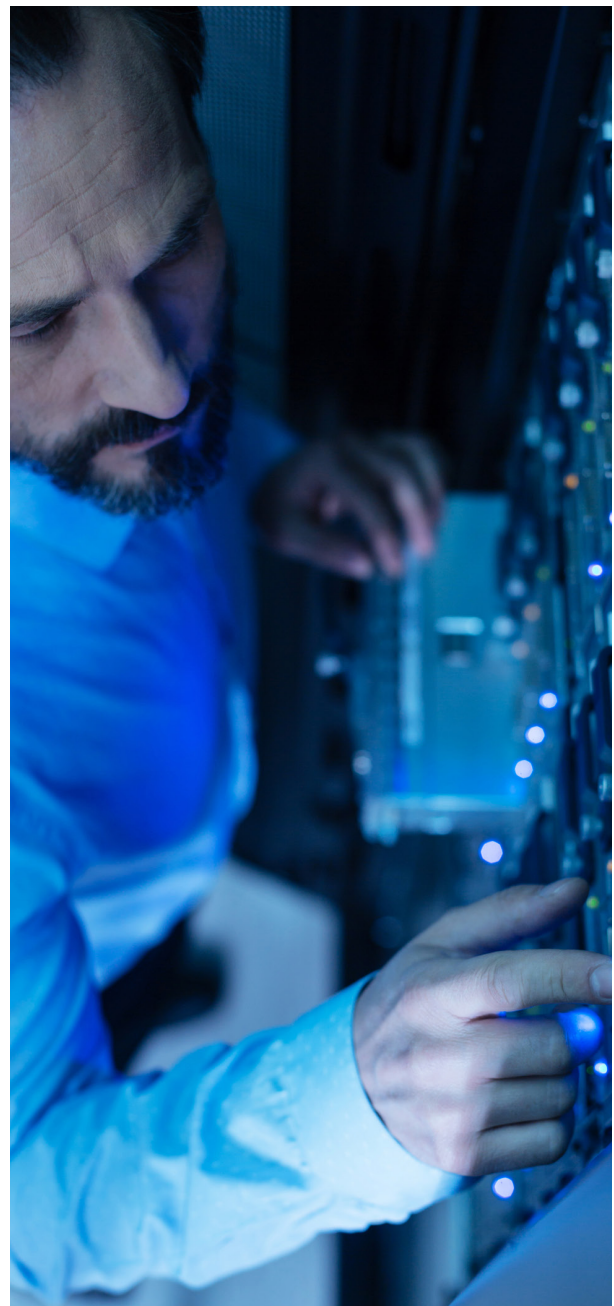
- **Seamless integration with advanced analytics:** Databricks seamlessly integrates with advanced analytics and machine learning tools, enabling organizations to derive more value from their data governance efforts.
- **Predictive analytics:** Organizations can leverage predictive analytics to anticipate and mitigate data governance challenges, ensuring proactive management.

9. Improved decision-making

- **Data-driven insights:** By ensuring data quality, accessibility, and security, Databricks enables better decision-making based on accurate and reliable data.
- **Strategic advantage:** Robust data governance enhances the organization's ability to leverage data strategically, gaining a competitive edge in the market.

10. Future-proofing data governance

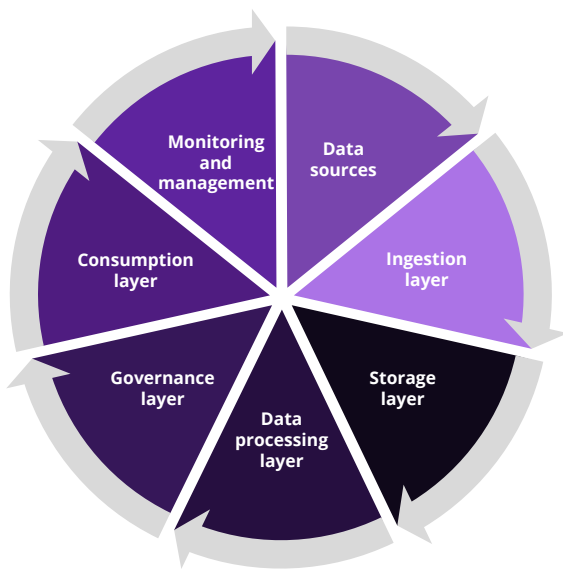
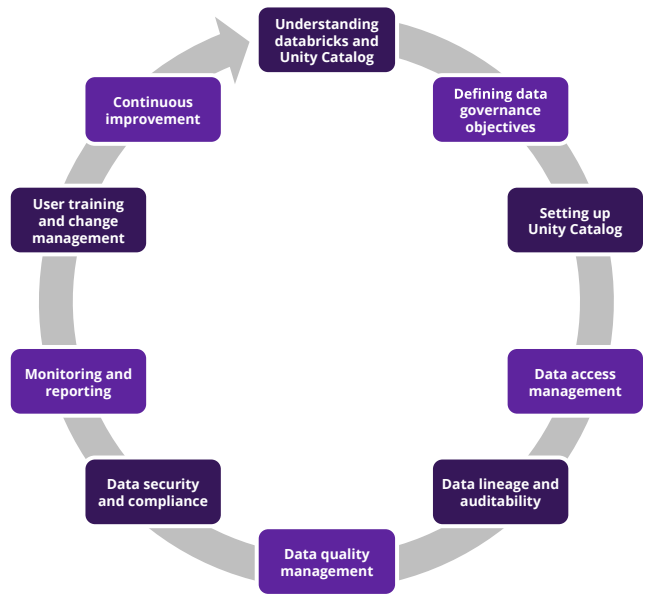
- **Adaptability to evolving needs:** Databricks' flexible architecture allows organizations to adapt their data governance frameworks to evolving business and regulatory needs.
- **Continuous improvement:** The platform supports continuous improvement of data governance practices through iterative enhancements and updates.



Technical details

Solution approach

Implementing a robust data governance framework around Databricks and Unity Catalog involves a combination of technology, policies, and processes to ensure data integrity, security, and compliance. Below is a structured approach to designing and implementing such a framework.



Reference architecture

This reference architecture outlines a comprehensive data governance framework using Unity Catalog and Databricks. It includes key components and best practices for managing, securing, and ensuring the quality of your data.

Solution components & tools:

1. Data sources

Data is ingested from various sources such as:

- **Databases:** Relational databases (PostgreSQL, MySQL), NoSQL databases (MongoDB, Cassandra).
- **Files:** CSV, JSON, Parquet files stored in cloud storage (AWS S3, Azure Blob Storage).
- **Streaming Data:** Real-time data from Apache Kafka, Kinesis, or Event Hubs.
- **APIs:** Data from RESTful APIs.

Tools:

- **Databricks Connect:** For integrating with on-premises databases.
- **Delta Live Tables:** For real-time ingestion from streaming sources like Kafka or Kinesis.
- **DBFS:** For storing static files and data.

2. Ingestion layer

This layer is responsible for ingesting data from various sources into the data lake house:

- **Databricks Auto Loader:** Automates the process of loading new data files as they arrive in the data lake.
- **Streaming ingestion:** Uses Apache Kafka, Kinesis, or Event Hubs for real-time data ingestion.
- **Batch ingestion:** Scheduled jobs that run ETL processes to load data in batches.

Tools:

- **Databricks notebooks:** For custom ETL pipelines using Python, Scala, or SQL.
- **Delta live tables:** For real-time ingestion and processing.
- **Apache Spark:** For batch and streaming data processing.
- **Databricks Autoloader:** For automated ingestion of files from cloud storage.

3. Databricks Lakeflow

Data is stored in a unified data lake house architecture, leveraging Databricks and Delta Lake:

- **Delta Lake:** Provides ACID transactions, scalable metadata handling, and data versioning.
- **Data Lake Storage:** AWS S3, Azure Data Lake Storage (ADLS), or Google Cloud Storage (GCS).

Tools:

- **Databricks Delta Lake:** The primary storage format for structured and semi-structured data.
- **Cloud Object Storage** (e.g., AWS S3, Azure Blob Storage, GCP Cloud Storage): For storing raw data and backups.

4. Data Processing Layer

Data processing is handled using Databricks:

- **ETL/ELT processes:** Transform and clean data using Apache Spark and Databricks notebooks.
- **Batch processing:** Scheduled jobs for regular data processing tasks.
- **Stream processing:** Real-time processing with Spark Structured Streaming.

Tools:

- **Databricks Notebooks:** For exploratory data analysis, feature engineering, and model development.
- **Apache Spark:** For large-scale data processing and machine learning workloads.
- **Delta Live Tables:** For real-time data processing and transformation.



5. Governance layer

The governance layer ensures data integrity, security, and compliance using Unity Catalog and other tools:

- Unity Catalog: Centralized metadata management, fine-grained access controls, and data lineage.
- Role-Based Access Control (RBAC): Manage permissions based on user roles and responsibilities.
- Data Lineage: Track the flow of data from source to consumption to ensure transparency and auditability.
- Data Quality: Implement data quality rules and validation checks to maintain high data standards.
- Data Masking and Encryption: Protect sensitive data through masking and encryption techniques.

Tools:

- Databricks Unity Catalog: For centralized metadata management, access control, and data governance policies.
- Apache Atlas: For enterprise-wide metadata management (optional, for large-scale environments).
- Databricks ACLs: For fine-grained access control.
- Databricks Workflows: For automating data governance tasks.

6. Consumption layer

The consumption layer provides various tools and interfaces for data access and analysis:

- Data analysts and scientists: Use Databricks notebooks, SQL Analytics, and BI tools (e.g., Tableau, Power BI) for data exploration and analysis.
- Business users: Access curated datasets through dashboards and reports.
- APIs and data services: Expose data through RESTful APIs for application integration.

Tools:

- DB SQL Warehouse: For SQL-based data exploration and analysis.
- Databricks Notebooks: For interactive data exploration and visualization.

- BI Tools (e.g., Tableau, Power BI): For creating dashboards and reports.
- Databricks offers SQL Serverless capabilities, where SQL queries are executed in a fully managed, scalable environment. This allows for high concurrency in queries, making it suitable for BI tools, dashboards, and SQL workloads that need quick query execution without manual scaling.

7. Monitoring and management

Continuous monitoring and management to ensure the health and performance of the data governance framework:

- Monitoring: Use Prometheus, Grafana, or Databricks native monitoring tools to track system performance and health.
- Logging: Centralized logging using the ELK stack (Elasticsearch, Logstash, Kibana) or Databricks logging features.
- Alerts: Set up alerts for critical events such as unauthorized access, data quality issues, and system failures.

Tools:

- Databricks monitoring: For monitoring cluster and job performance.
- Databricks SQL analytics: For analyzing query performance and resource utilization.
- Cloud Provider monitoring (e.g., AWS CloudWatch, Azure Monitor, GCP Cloud Monitoring): For overall platform monitoring.

Additional considerations

- **Data quality:** Use Databricks built-in functions and libraries for data profiling and validation
- **Security:** Implement encryption, network security, and access controls as outlined in previous responses.
- **Metadata management:** Leverage Unity Catalog for comprehensive metadata management and governance.
- **Collaboration:** Utilize Databricks workspaces and collaboration features for team-based projects.

Bringing it to life

In order to help clients speed up their data management and data analytics projects, DXC has developed an accelerator called “Data Integration Framework” (DIF).

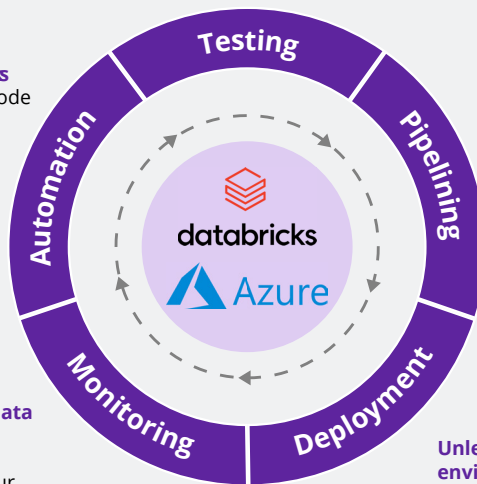
DIF is a code accelerator, meaning when DXC starts implementing a project, we come with prebuild

pipelines, ETLs, orchestration etc. (actual code that can be deployed) which can cover a large number of use cases. This code is adjusted for your specific use case for your deployment. When engineers start a new task, they start being 70% done.

The DXC quick start for DIF Available for Microsoft Azure

Free your teams from repetitive manual tasks by making use of a no-code fully metadata driven ETL/ingestion design

Get near-real-time data flow monitoring by automatic statistics collection on all of your pipelines and dataflows



Maintain high reliability by using the built in Testing suite which is enabling your developers to swiftly test and verify dataflows

Ingest from even thousands of data sources/source tables and leverage our automatic pipeline building tools allowing you to securely ingest from even hundreds of databases at the time

Unleash CI/CD and multiple environment/tenant support by seamlessly and effortlessly deploying your infrastructure across multiple locations

Figure 2 - DIF Benefits

While its original focus was on avoiding technical debt along the platform lifecycle and cutting cost for pipeline development by ~50%, it also ensures a data governance framework can be brought to life easier.

DIF, including code deployment, is fully metadata driven and hence provides a full, detailed overview of all data flows and their execution. The powerful Databricks features used and orchestrated in our framework form the ideal basis for a sophisticated data governance.

- Data Governance Foundation: Uniform definitions of data sets, ETLs, ingestions and transformations (business logic) together with data security policies and resource definitions

- Data Trust Foundation: Uniform ways of collecting QA KPIs across all the data flows
- Data Integration Framework: Multiple sets of deployment, orchestration and execution engines executing logic that is stored in DGF

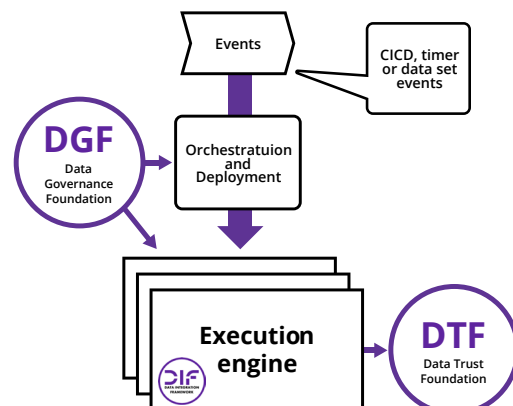


Figure 3 - DIF Component Interaction

Users add new metadata about ingestions and processing (DGF). The system picks up these changes when made deploying new tasks (DIF). The execution engine performs the tasks and stores the Quality Assurance KPIs (DTF). In case the KPIs point to a failure, anomalies are automatically detected, and incidents created for follow-up. This way Quality Assurance for instance, is not a “bolt-on” but built into the system.



Best practices

Building a robust data governance framework using Databricks requires a strategic approach that encompasses various aspects of data management, security, and compliance. Here are the best practices to effectively leverage Databricks engineering for establishing a comprehensive data governance framework.

1. Establish clear data governance policies and procedures

- **Define governance objectives:** Start by clearly defining the objectives of your data governance program. This includes identifying the key goals such as data quality, compliance, security, and accessibility. It is essential, to link the objectives to business requirements and values such as “compliance,” “lower stock security levels,” “faster closing” or similar.
- **Develop policies and standards:** Create comprehensive data governance policies and standards that align with your organization’s regulatory and business requirements. These should cover data classification, data privacy, access controls, and data lifecycle management.
- **Document procedures:** Ensure that all procedures related to data handling, processing, and management are well-documented. This includes data ingestion, data transformation, and data storage practices.

2. Implement data quality management

- **Automate data quality checks:** Use Databricks to automate data quality checks during data ingestion and processing. Implement validation rules to ensure data accuracy, consistency, and completeness.
- **Monitor data quality metrics:** Continuously monitor data quality metrics using Databricks’ integrated tools (Unity catalog). Set up alerts for any deviations from the defined quality standards to enable prompt corrective actions.
- **Data cleansing pipelines:** Develop data cleansing pipelines in Databricks to systematically clean and standardize data. This helps in maintaining high-quality data that can be trusted for decision-making.

3. Ensure data security and compliance

- **Access controls and user permissions:** Leverage Databricks' security features to implement fine-grained access controls. Use role-based access control (RBAC) to ensure that users only have access to data necessary for their roles.
- **Encryption and masking:** Ensure that sensitive data is encrypted both at rest and in transit. Use data masking techniques for non-production environments to protect sensitive information.
- **Audit and compliance logging:** Enable comprehensive logging to track data access and modifications. This ensures accountability and helps in compliance with regulatory requirements such as GDPR, HIPAA, and CCPA.

4. Foster data collaboration and stewardship

- **Data cataloging:** Use Unity catalog to create and maintain a data catalog that provides a centralized repository of metadata. This helps users discover, understand, and trust the data available.
- **Data stewardship roles:** Assign data stewards responsible for overseeing data governance practices, ensuring data quality, and managing data lifecycle.
- **Collaborative platforms:** Leverage Unity catalog collaborative features to enable seamless communication and collaboration among data engineers, data scientists, and business users.



5. Develop scalable data architecture

- **Modular pipelines:** Design modular and reusable data pipelines in Databricks to streamline data processing and ensure scalability. Modular pipelines simplify maintenance and facilitate the integration of new data sources. Be cognizant of not building up technical debt. Our DIF framework may help with this task.
- **Data lake house architecture:** implement a data lake house architecture in Databricks that combines the best of data lakes and data warehouses. This allows for scalable storage and efficient analytics.
- **Performance optimization:** Optimize the performance of your Databricks environment by leveraging features like caching, optimized cluster configurations, and efficient data partitioning strategies.

6. Enable continuous improvement

- **Feedback loops:** Establish feedback loops to continuously collect input from data users and stakeholders. Use this feedback to refine data governance practices and policies.
- **Regular audits and reviews:** Conduct regular audits and reviews of your data governance framework to ensure compliance and identify areas for improvement.
- **Training and awareness:** Provide ongoing training and awareness programs for your teams to keep them updated on data governance best practices and Databricks capabilities.

7. Utilize Databricks' integrated tools and features

- **Delta lake:** Use Delta Lake for reliable data lakes with ACID transactions, scalable metadata handling, and unified streaming and batch data processing.
- **Databricks Unity Catalog:** Implement Unity Catalog for fine-grained data governance and data discovery across all your Databricks workspaces.
- **Machine learning governance:** Leverage Databricks' machine learning capabilities to be govern ML models, ensuring they are well-documented, reproducible, and compliant with regulatory standards.

Security considerations

A robust data governance framework is essential for protecting sensitive data and ensuring compliance. When leveraging Databricks and Unity Catalog, several security considerations must be addressed:

1. Data classification and access control

- **Granular access controls:** Utilize Unity Catalog's fine-grained access controls to define permissions at the object, column, and row levels, minimizing data exposure.
- **Data classification:** Implement a comprehensive data classification scheme to categorize data based on sensitivity and regulatory requirements.
- **Least privilege principle:** Adhere to the principle of least privilege, granting users only the necessary permissions to perform their tasks.
- **Regular access reviews:** Conduct periodic reviews of user access to identify and revoke unnecessary privileges.

2. Network security

- **Network isolation:** Isolate the Databricks environment from the public internet through firewalls, network segmentation, and VPNs.
- **Secure communication:** Enforce encryption for data in transit using protocols like TLS/SSL.
- **Intrusion detection and prevention systems (IDPS):** Deploy IDPS solutions to monitor network traffic for suspicious activities.

3. Data encryption

- **Data at rest encryption:** Encrypt data stored in Databricks using platform-provided encryption capabilities.
- **Data in transit encryption:** Ensure data is encrypted during transmission between clients and Databricks.
- **Key management:** Implement robust key management practices to protect encryption keys.

4. Identity and access management (IAM)

- **Centralized IAM:** Integrate Databricks with your organization's IAM system for unified user management and authentication.
- **Multi-factor authentication (MFA):** Enforce MFA for all users accessing the Databricks environment.
- **Role-based access control (RBAC):** Define roles with appropriate permissions and assign them to users based on their job functions.



5. Data loss prevention (DLP)

- **Sensitive data detection:** Implement DLP policies to identify and protect sensitive data within Databricks.
- **Data loss prevention controls:** Implement DLP controls to prevent unauthorized data exfiltration.

6. Threat detection and response

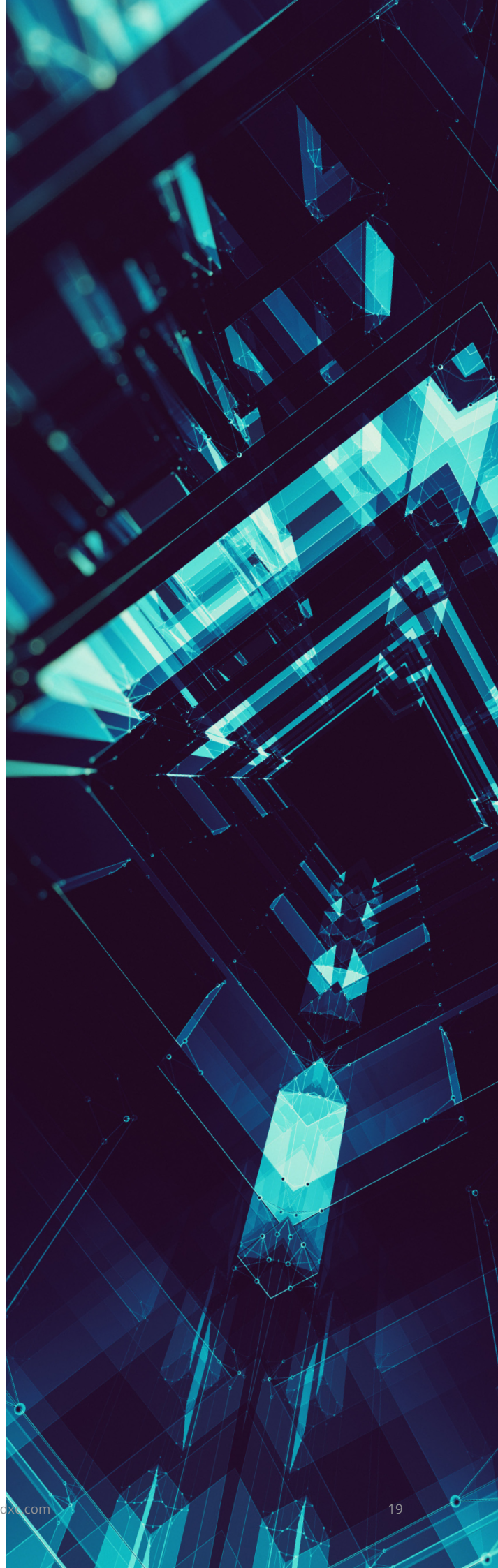
- **Anomaly detection:** Utilize Databricks' built-in monitoring and logging capabilities to detect unusual activities.
- **Incident response plan:** Develop a comprehensive incident response plan to address security breaches effectively.

7. Regular security audits and assessments

- **Vulnerability assessments:** Conduct regular vulnerability assessments to identify and address security weaknesses.
- **Penetration testing:** Perform penetration testing to simulate real-world attacks and identify vulnerabilities.
- **Compliance audits:** Ensure compliance with relevant industry regulations and standards (e.g., GDPR, HIPAA, PCI DSS).

8. Additional considerations

- **Data masking:** Utilize data masking techniques to protect sensitive data during development and testing.
- **Secure data sharing:** Implement secure data sharing mechanisms (e.g., Delta Sharing) when sharing data with external parties.
- **User education and awareness:** Conduct regular security awareness training for employees to prevent social engineering attacks.



Benefits

By harnessing the power of Databricks and Unity Catalog, organizations can construct a data governance framework that is not only effective but also agile, scalable, and cost-efficient. This framework empowers businesses to create a data-driven culture by ensuring data quality, security, and accessibility, ultimately maximizing the value of their data assets.

Technical benefits

Databricks and Unity Catalog offer a powerful combination for establishing a robust data governance framework. The technical benefits of this approach include:

1. Centralized data management

- **Unified metadata store:** Unity Catalog provides a centralized repository for managing metadata, including schemas, tables, views, and machine learning models.
- **Consistent governance policies:** Enforce uniform governance policies across all data assets, reducing inconsistencies and risks
- **Improved data discovery:** Easily discover and access relevant data through a centralized catalog

2. Enhanced data security

- **Fine-grained access control:** Implement granular access controls at the object, column, and row levels to protect sensitive data.
- **Role-based access control (RBAC):** Define and manage user roles and permissions efficiently.
- **Data masking and obfuscation:** Protect sensitive data through masking and obfuscation techniques

3. Improved data quality

- **Data lineage:** Track the origin and transformation of data, enabling root cause analysis and data quality improvements
- **Data profiling and validation:** Automate data quality checks and profiling to identify inconsistencies and anomalies
- **Data cleansing and enrichment:** Easily clean and enrich data to improve its quality and reliability

4. Accelerated data engineering and analytics

- **Self-Service Data Preparation:** Empower data analysts and scientists to prepare and access data independently
- **Increased productivity:** Streamline data engineering workflows with automated tasks and reusable components
- **Faster time to insights:** Accelerate data-driven decision making with improved data quality and accessibility



5. Scalability and performance

- **Elastic scalability:** Handle increasing data volumes and workloads efficiently
- **Optimized Performance:** Leverage Databricks' distributed computing architecture for high performance
- **Cost-Effective:** Optimize resource utilization and reduce costs through efficient data management

6. Integration and interoperability

- **Open Standards:** Adherence to open standards (e.g., Delta Lake) facilitates data sharing and integration with other systems
- **Ecosystem integration:** Connect to a wide range of data sources and tools through APIs and connectors
- **Cloud agnostic:** Deploy Databricks on multiple cloud platforms for flexibility and portability

Business Benefits

A robust data governance framework built on Databricks and Unity Catalog offers significant business advantages:

- **Improved data quality:** Centralized data management and automated data quality checks enhance data accuracy, consistency, and reliability, leading to better decision-making
- **Enhanced data security:** Comprehensive access controls, data classification, and encryption protect sensitive information from unauthorized access, reducing the risk of data breaches and financial losses
- **Increased operational efficiency:** Automated workflows and streamlined data processes improve operational efficiency and reduce time-to-market for new products and services
- **Enhanced regulatory compliance:** A well-defined data governance framework helps organizations meet industry regulations (e.g., GDPR, CCPA, HIPAA) by ensuring data privacy, security, and accountability
- **Improved decision making:** High-quality, accessible, and trustworthy data empowers businesses to make informed decisions, driving growth and innovation
- **Risk mitigation:** By proactively identifying and addressing data quality and security issues, organizations can mitigate risks and protect their reputation
- **Cost reduction:** Optimized data management and reduced data redundancy can lead to significant cost savings

User benefits

A well-implemented data governance framework using Databricks and Unity Catalog provides numerous benefits to users:

- **Simplified data access:** Users can easily discover and access relevant data through a centralized catalog
- **Increased productivity:** Self-service data preparation and access enable users to focus on analysis and insights rather than data wrangling
- **Enhanced collaboration:** Improved data sharing and collaboration capabilities foster teamwork and knowledge sharing
- **Trustworthy data:** Users can rely on the accuracy and consistency of data, leading to increased confidence in insights and decisions
- **Faster time to value:** Accelerated data analysis and insights generation drive faster time-to-market for new products and services
- **Improved data literacy:** A well-governed data environment promotes data literacy and empowers users to make data-driven decisions

References:

Unity Catalog Setup: [Data governance with Unity Catalog | Databricks on AWS](#)

[Connect BI tools to Unity Catalog | Databricks on AWS](#)

Data Sharing: [Share data and AI assets securely using Delta Sharing | Databricks on AWS](#)

Databricks Marketplace: [What is Databricks Marketplace? | Databricks on AWS](#)

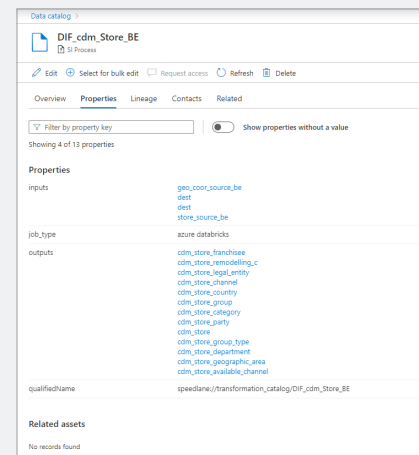
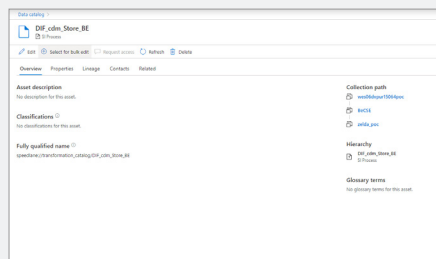
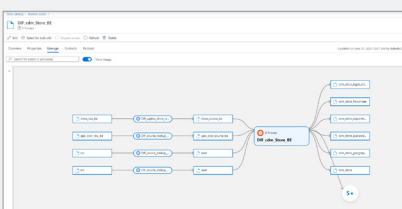
Security & Compliances: [Security and compliance guide | Databricks on AWS](#)

Unity Catalog Best Practices: [Unity Catalog best practices | Databricks on AWS](#)

Available DXC solution accelerators:

- DIF is a **code accelerator**
- Engineers have access to actual **prebuilt code that can be deployed**, such as deployment, pipelines, ETLs, orchestration...
- When engineers start a new task, they start being **60-80% done**, as they take and adjust rebuilt components
- DIF is built using either **cloud-native tools**, we do not introduce any expensive licensed components
- When you get DIF, **you get the source code** and get a non-revokable license to use and modify it for your own use

- DIF is being actively developed, once you get it you **get all the updates/bugfixes/new features** that we develop internally free of charge as long as our collaboration is ongoing










Example case study: Implementing data governance at an asset management company

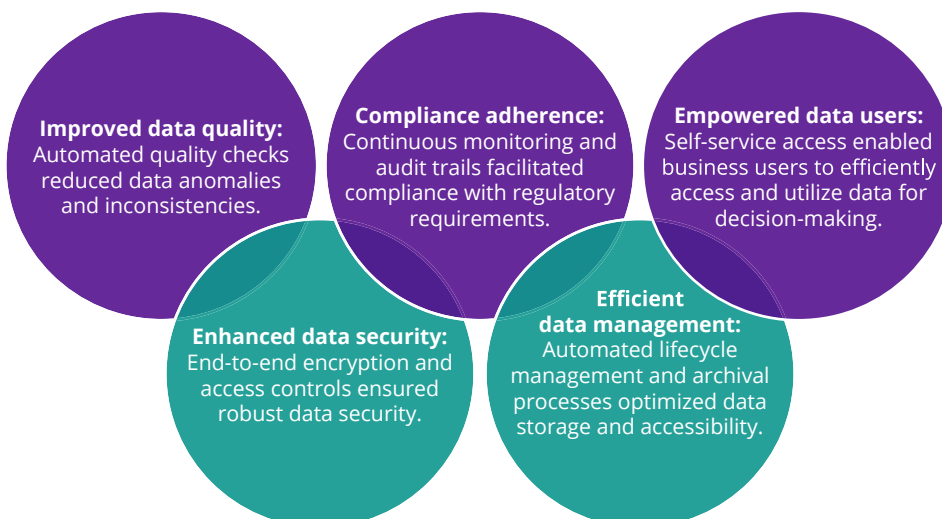
Background

Asset Management Co, a global enterprise, faced challenges in managing data quality, security, and compliance across its diverse data landscape. The company leveraged Databricks to engineer a comprehensive data governance framework.

Implementation

-  **Data quality management:** Asset Management Co used Databricks to profile and cleanse data, implementing automated quality checks to maintain data integrity.
-  **Data security:** The company implemented end-to-end encryption and fine-grained access controls in Databricks, ensuring data security and privacy.
-  **Data lineage:** Databricks' data lineage tracking capabilities enabled Asset Management Co to maintain transparency and traceability of data transformations and movements.
-  **Compliance:** By leveraging Databricks' audit trails and compliance monitoring, Asset Management Co ensured adherence to regulatory requirements and internal policies.
-  **Lifecycle management:** The company defined and enforced data retention policies in Databricks, automating the archival of outdated data.
-  **Data stewardship:** Asset Management Co established clear data stewardship roles and documented governance practices using Databricks' collaborative features.
-  **Access management:** Granular access controls and self-service data access capabilities in Databricks empowered business users while maintaining data security.

Outcomes



Conclusion

By leveraging Databricks engineering, organizations can create a robust data governance framework that enhances data quality, security, and compliance. This framework mitigates data management risks while unlocking the full potential of data assets, driving innovation, and maintaining competitive advantage. The strategies and best practices outlined in this whitepaper serve as a comprehensive guide for organizations on their data governance journey, ensuring effective, secure, and compliant data management.

A key aspect highlighted is the integration of data governance with data engineering practices. Traditionally treated as separate disciplines, combining these practices within a single platform results in more comprehensive and unified solutions. This integration reduces costs and streamlines data practices across the enterprise, promoting consistency and efficiency.

Adopting the approaches detailed in this whitepaper transforms data governance from a burdensome necessity into a strategic asset. In the modern data-driven landscape, effective data management significantly impacts business success. Databricks' advanced capabilities equip organizations to build a data governance framework that supports current needs and adapts to future challenges, ensuring sustained growth and competitive advantage.

Key benefits

- **Improved data quality:** The combination of Delta Lake's ACID transactions and schema enforcement capabilities ensures that data remains accurate, consistent, and reliable. These foster trust in the data, which is critical for making informed business decisions.
- **Enhanced security:** With fine-grained access control, end-to-end encryption, and robust authentication mechanisms, Databricks and Unity Catalog provide a secure environment for data storage and processing. These security measures protect sensitive data from unauthorized access and breaches, ensuring data confidentiality and integrity.
- **Compliance adherence:** Databricks' built-in compliance features and support for regulatory standards such as GDPR, HIPAA, and CCPA help organizations adhere to legal requirements. Automated policy enforcement, detailed audit logs, and comprehensive data lineage tracking further enhance compliance and governance efforts.
- **Scalability and performance:** Databricks' cloud-native architecture allows organizations to scale their data infrastructure seamlessly, accommodating growing data volumes and varying workloads without compromising performance. This scalability ensures that the data governance framework remains effective as data needs to evolve.
- **Unified data management:** By centralizing data management through Unity Catalog, organizations can streamline data discovery, classification, and access management. This unified approach reduces data silos and promotes a holistic view of data across the enterprise.
- **Advanced analytics integration:** Databricks' support for machine learning and AI enables organizations to leverage advanced analytics within their governed data environment. This integration facilitates innovative data-driven solutions while maintaining governance and security standards.





Strategic roadmap

To successfully leverage Databricks for data governance, organizations should follow a strategic roadmap that includes:

- **Assessment and planning:** Conduct a thorough assessment of current data governance practices and identify any gaps. Develop a detailed plan that outlines the objectives, scope, and timeline for implementing the data governance framework.
- **Policy development:** Establish comprehensive data governance policies that define data quality standards, security protocols, access controls, and compliance requirements. Ensure that these policies are aligned with organizational goals and regulatory standards.
- **Implementation and integration:** Deploy Databricks and Unity Catalog, integrating them with existing data infrastructure and workflows. Implement the defined data governance policies and configure access controls, encryption, and compliance settings.
- **Training and adoption:** Provide training to data stewards, engineers, analysts, and other stakeholders to ensure they understand the data governance framework and their roles within it. Foster a culture of data stewardship and accountability.
- **Monitoring and optimization:** Continuously monitor the data governance framework for effectiveness and compliance. Utilize audit logs, monitoring tools, and real-time alerts to identify and address issues promptly. Regularly review and update policies to adapt to changing data needs and regulatory requirements.

About the author



Arun Khandelwal

Enterprise Data Solution Architect, Databricks Partner
Solution Architect Champion at DXC Luxoft

Arun is a Senior Technology and Principal Architect with 22 years of experience in leading organizations to successfully tackle complex business challenges through innovative solutions and robust execution. He specializes in creating IT solutions that harness the power of Data Architecture, Data Governance, Data Analytics, Data Engineering, Microservices, SOA architecture, Web APIs, Data Warehousing (DW), Data/Business Modeling, Data-as-a-Service (DaaS), and Optimization solutions to drive growth and modernization. Arun has a proven track record of rearchitecting, redesigning, and migrating legacy systems to modern platforms, ensuring seamless transitions and enhanced business performance.

Special thanks to **Mario Palmer-Huke** for providing feedback and review.



DXC Technology (NYSE: DXC) helps global companies run their mission-critical systems and operations while modernizing IT, optimizing data architectures, and ensuring security and scalability across public, private and hybrid clouds. The world's largest companies and public sector organizations trust DXC to deploy services to drive new levels of performance, competitiveness, and customer experience across their IT estates. Learn more about how we deliver excellence for our customers and colleagues at [DXC.com](https://www.dxc.com).