CSCI E-103

Data Engineering for Analytics to Solve Business Challenges

Data & AI Governance

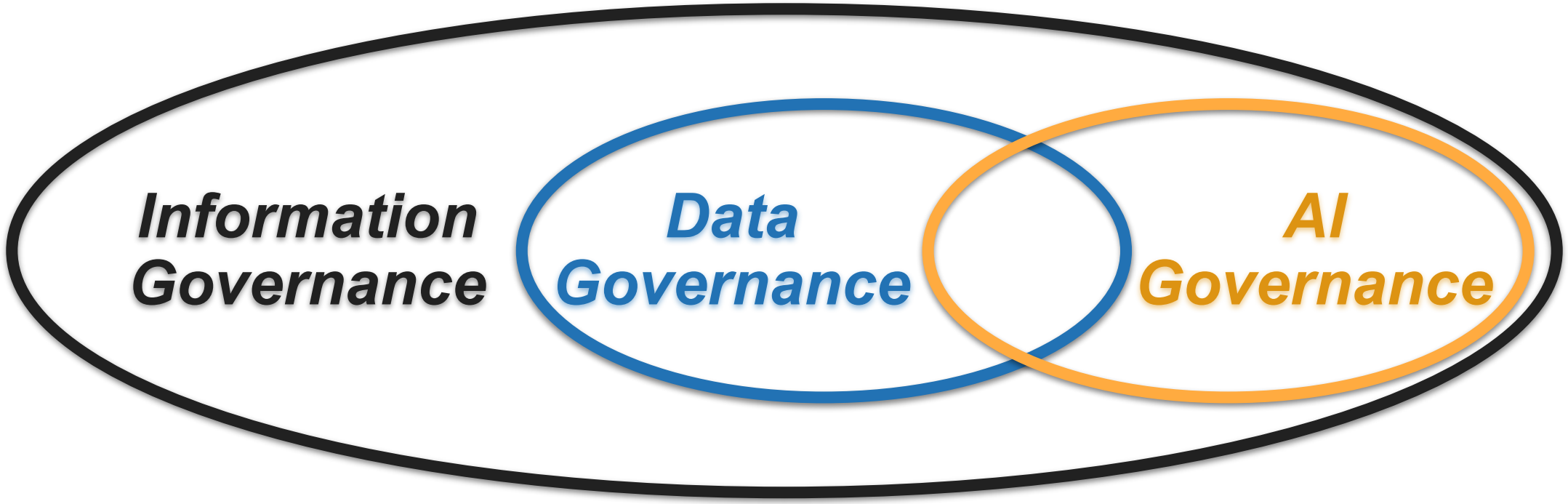
Lecture 12

Mohan

Harvard Extension, Fall 2025

Agenda

- Information Governance [Theory]
- Data & AI Governance [Theory]
- Well Architected Lakehouse
- Data & AI Governance on Databricks
  - Unity Catalog
  - Row Level Filtering and Column Level Masking
  - ABAC - Scaling Governance with Policies
  - Governed Tags
  - Data Classification
  - Data Quality Monitoring
  - Data + AI Lineage
  - Lakehouse Federation
  - Delta Sharing


Information Governance
Data Governance
AI Governance

Information Governance

Information

… in the context of Information Governance

In the context of Information Governance, "information" refers to any data or knowledge that is created, collected, stored, processed, shared, or used by an organization.

Typical types of information

- Physical information
  - Hard-copy documents or records stored in physical formats
- Digital information
  - Electronically stored structured, semi-structured, and unstructured data
  - Ranges from commonly known data to proprietary and highly sensitive data
  - Includes metadata
- Knowledge-based information
  - Insights or expertise derived from analyzing and interpreting raw data

Information Governance

Information Governance is a framework of policies, procedures, standards, and controls * that organizations use

- to manage their information assets effectively,
- ensuring that data is secure, accurate, accessible,
- and compliant with legal and regulatory requirements.

It involves the strategic oversight of how information is created, stored, shared, used, and disposed of to support business objectives while mitigating risks.

*

Policies	High-level guiding principles
Procedures	Step-by-step operational instructions
Standards	Benchmarks defining how processes should be performed
Controls	Specific mechanisms enforcing policies and standards

Information Governance Goals

Maximize Information Value

- Ensure data quality and accuracy
- Improve decision-making processes by providing access to reliable information
- Streamline information management processes for operational efficiency
- Enable the effective use of information to support business objectives

Risk Mitigation and Compliance

- Protect sensitive information from unauthorized access and breaches
- Ensure compliance with data protection laws like GDPR and HIPAA
- Mitigate legal risks associated with improper information management
- Reduce the likelihood and costs of legal discovery and regulatory penalties

Enhance Information Security

Implement robust data security measures
Protect against cyber threats and data breaches
Establish clear policies for information access and use
Ensure the integrity and confidentiality of sensitive data

Optimize Information Lifecycle Management

Implement proper data creation, storage, and deletion processes

Ensure appropriate retention and disposition of information

Optimize storage costs and reducing unnecessary data accumulation

Improve data accessibility while maintaining security

Promote Transparency and Accountability

Establish clear roles and responsibilities for information management

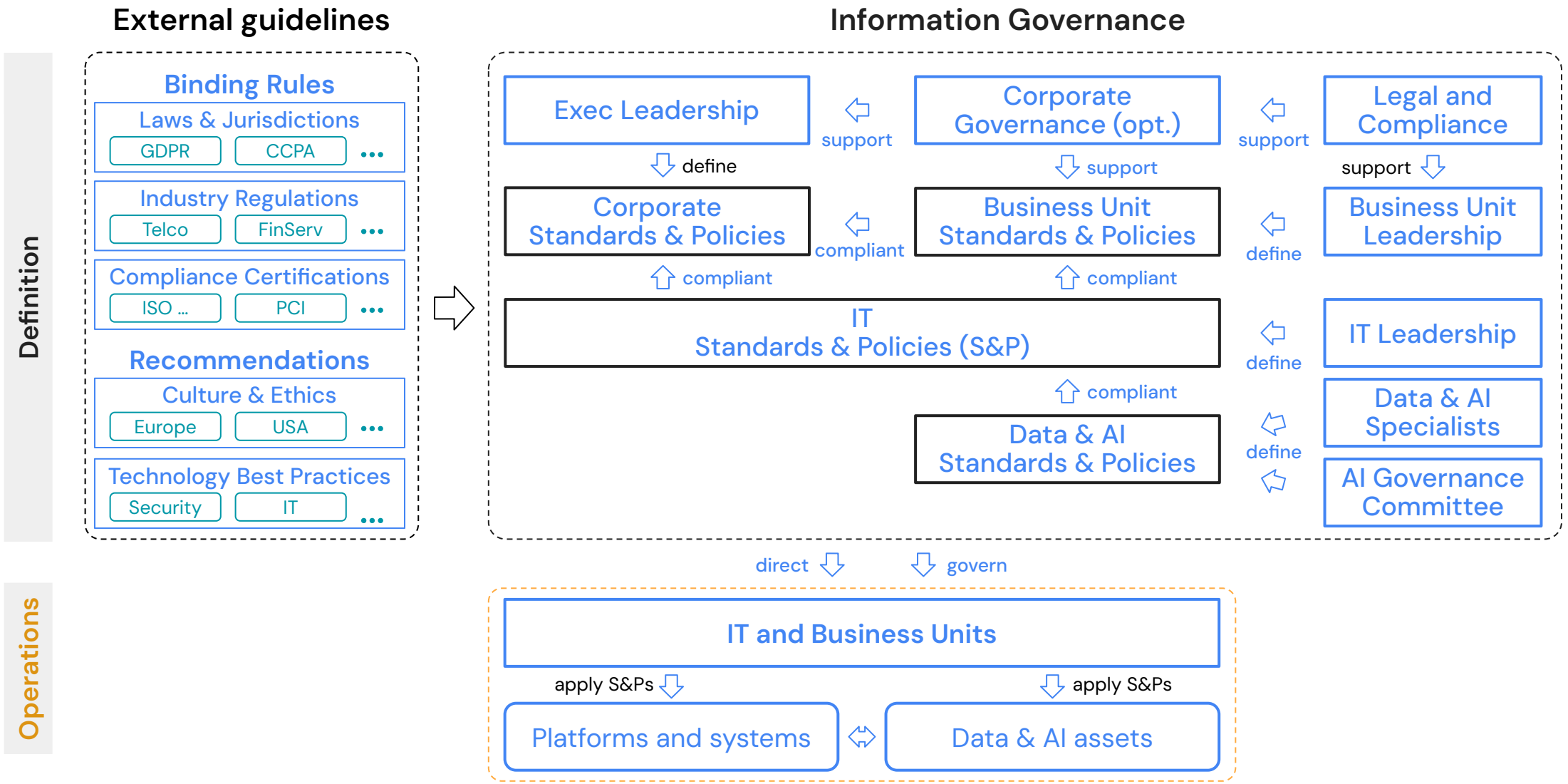
Create transparent policies and procedures for information handling

Implement audit trails and reporting mechanisms

Enhance stakeholder trust through responsible data management practices

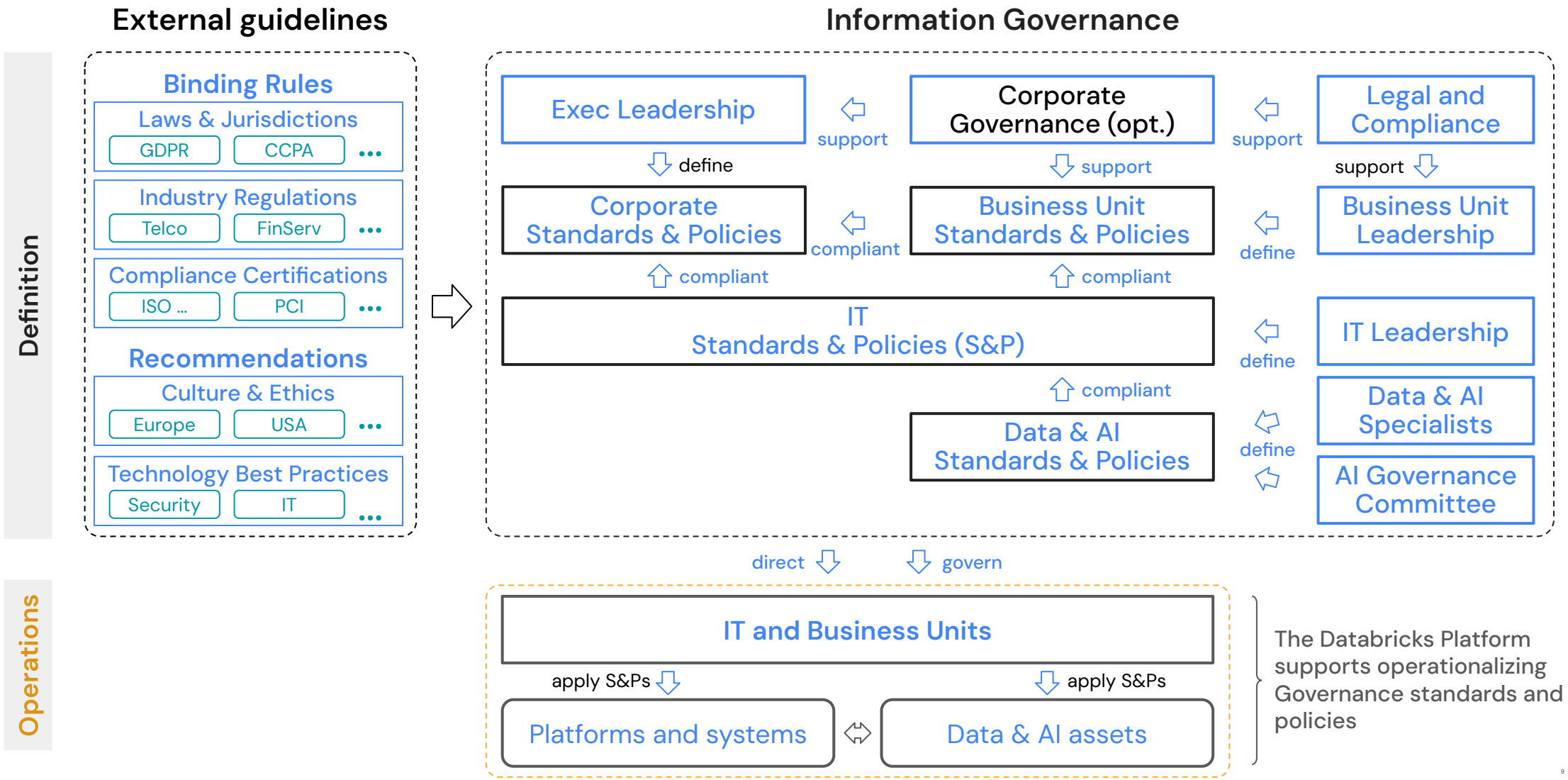
Information Governance Overview

Governance needs a strong *Definition* and a strong *Operations* practice



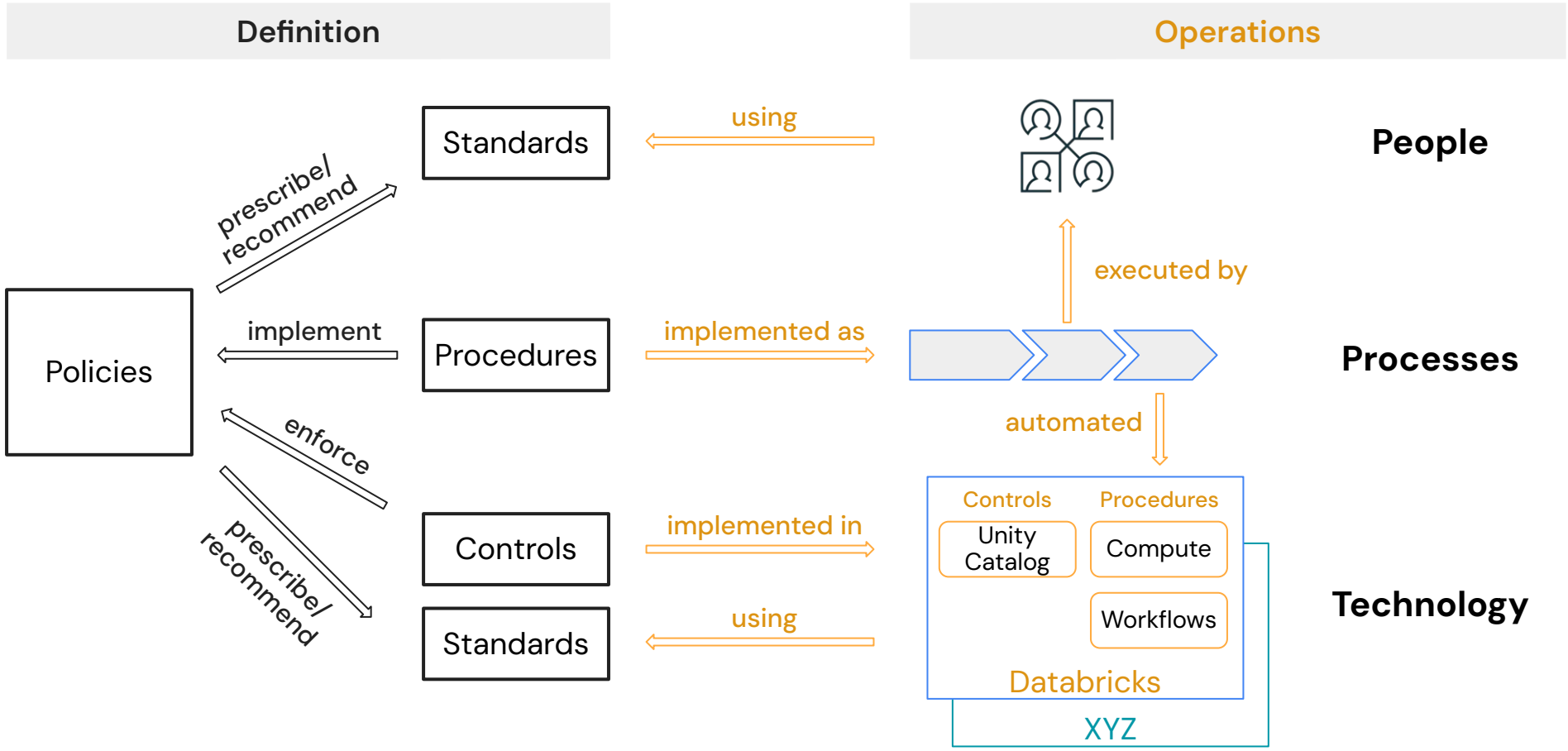
Information Governance with Databricks

Databricks offers an open framework for managing data and AI assets



Governance is a holistic approach

It provides guidelines for people, processes and technology



Policies, Procedures, Standards, and Controls

By example of GDPR

Policies are high-level, formal statements that outline an organization's commitment to GDPR compliance and its approach to managing personal data. They serve as guiding principles for employees and stakeholders. Examples:

- **Data Protection Policy:** Describes how the organization protects personal data, including its commitment to GDPR principles such as lawfulness, fairness, transparency, and accountability.
- **Privacy Policy:** A document shared with data subjects explaining how their personal data is collected, processed, stored, and shared (e.g., on a website or app).
- **Retention Policy:** Defines how long personal data will be retained and the criteria for securely disposing of it once it is no longer needed.

Procedures are detailed step-by-step instructions that describe how policies are implemented in practice. They provide operational guidance for employees on specific tasks related to GDPR compliance. Examples:

- **Data Subject Access Request (DSAR) Procedure**: Step-by-step instructions for handling requests from individuals who want access to their personal data under GDPR Article 15.
- **Incident Response Procedure:** Instructions for responding to a data breach or security incident as required by GDPR Article 33.
- **Consent Management Procedure:** Steps for obtaining, recording, and managing consent from data subjects (GDPR Article 7).

Standards are specific benchmarks or criteria that define how processes should be performed in alignment with GDPR requirements. They ensure consistency and quality across the organization. Examples:

- **Encryption Standards:** Specify the type of encryption algorithms (e.g., AES-256) required to secure personal data during storage or transmission (GDPR Article 32 on security).
- **Data Minimization Standard:** Define rules for collecting only the minimum amount of personal data necessary for a specific purpose (GDPR Article 5(1)(c)).
- **Access Control Standard:** Establish guidelines for restricting access to personal data based on role-based access control (RBAC) principles.

Controls are specific measures or mechanisms put in place to enforce policies, implement procedures, and meet standards. They can be technical, administrative, or physical safeguards designed to ensure compliance. Examples:

- **Technical Controls**, e.g. implementing multi-factor authn (MFA) for systems storing personal data, or using automated tools to anonymize sensitive data.
- **Administrative Controls**, e.g. conducting regular employee training on GDPR, or performing periodic audits of data processing activities to ensure compliance.
- **Physical Controls**, e.g. restricting access to servers storing personal data through locked rooms or biometric authentication.

Data & AI Governance

Data & AI Governance

Data Governance

A formalized framework of processes, policies, and standards that guide how data is collected, stored, protected, and utilized within an organization.

It aims to ensure the availability, usability, integrity, and security of data while adhering to regulations and internal policies.

Effective data governance transforms data into a valuable enterprise asset by ensuring its accuracy, consistency, trustworthiness, and security

AI Governance*

Frameworks, principles, and tools used to manage the development, deployment, and oversight of AI systems.

It ensures that AI is ethically designed, transparent, secure, and aligned with organizational values and societal norms.

Key components include risk management, regulatory compliance, fairness in AI models, explainability of decisions, and data privacy protection

Data & AI Governance

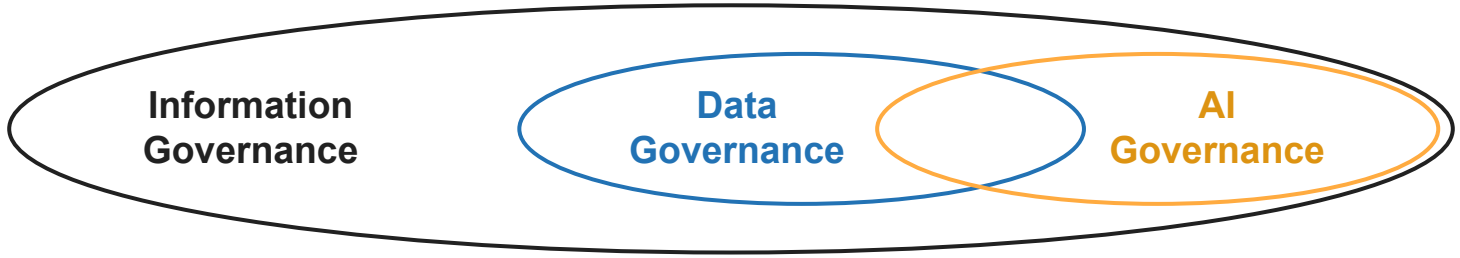
Integrates the principles of both frameworks to manage the lifecycle of data and AI systems cohesively.

While data governance ensures high-quality data for responsible use in AI processes, AI governance focuses on ethical AI development and deployment.

Together, they address challenges such as bias detection in AI models through robust data standards and ensure compliance with regulations across both domains.

* Including ML, AI Agents, deep learning, ...

Information, Data and AI Governance



	Information Governance	Data Governance	AI Governance
Focus	Meaning & context of information	Technical management of data (structured, unstructured, ...)	Ethical use & oversight of AI systems
Scope	Broad (includes physical & digital information)	Narrower (data lifecycle & quality)	AI models, algorithms & risks
Key Concerns	Compliance, business value	Data security, integrity	Bias mitigation, transparency
Interconnections	Provides context for governed data	Supplies quality data for business (incl. AI systems)	Relies on governed data; impacts corporate ethics

Areas covered by data and AI governance

Data Governance topics

- Metadata Management & Data Discovery
- Master Data Management
- Data Classification
- Data Handling Standards
- Data Quality Management
- Data Lineage & Impact Analysis
- Data Stewardship & Workflow
- Measurement & Optimization
- Integration & Interoperability
- Policy & Compliance Management
- AI & Automation Support

AI Governance topics

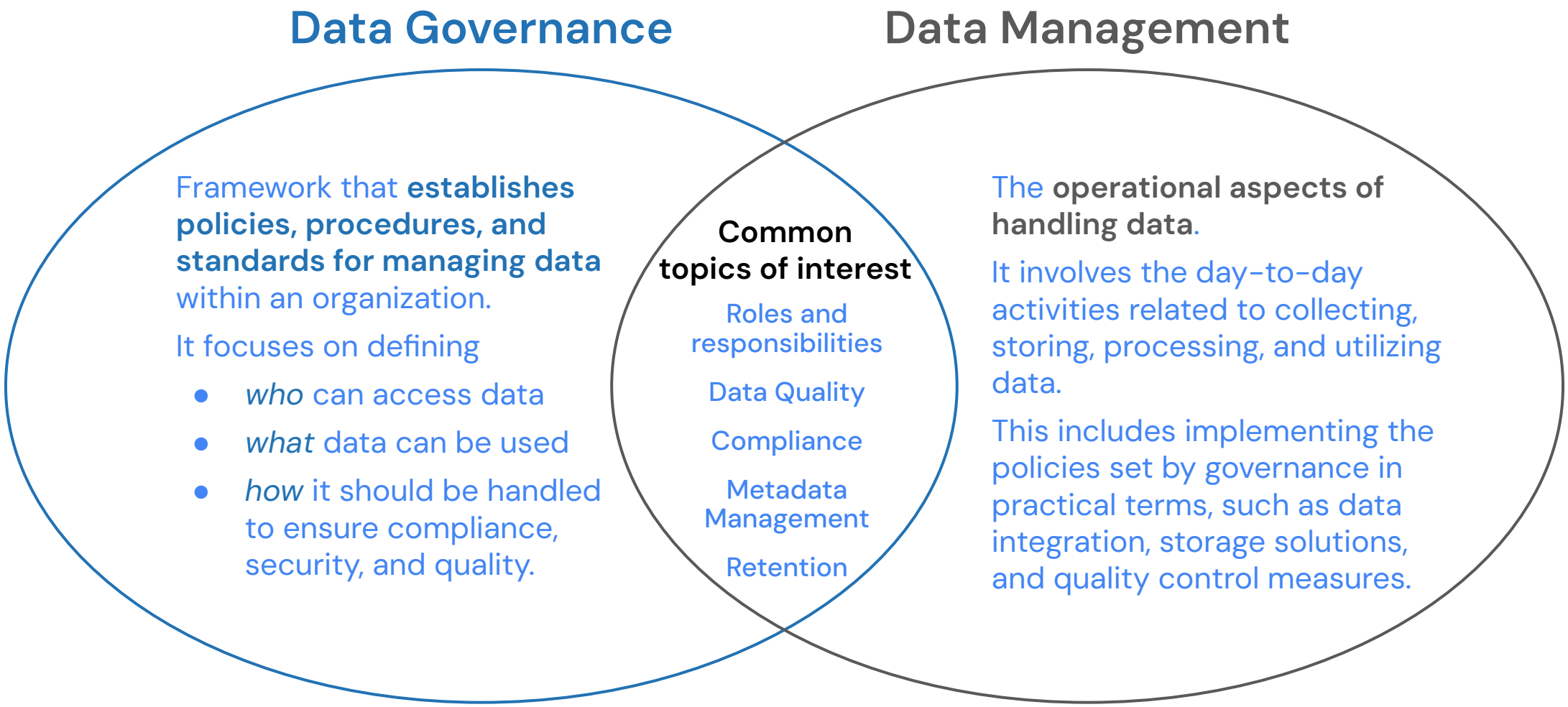
- Data Governance Integration

- Model Management & Lifecycle
- Monitoring & Validation
- Security & Privacy
- Audit & Reporting
- Collaboration & Workflow
- Risk Management & Assessment
- Transparency & Explainability
- AI Education and Training
- Fairness & Ethical AI
- Compliance & Regulatory Management
- Stakeholder Engagement & Impact Assessment

For more details see the capabilities section of the deck

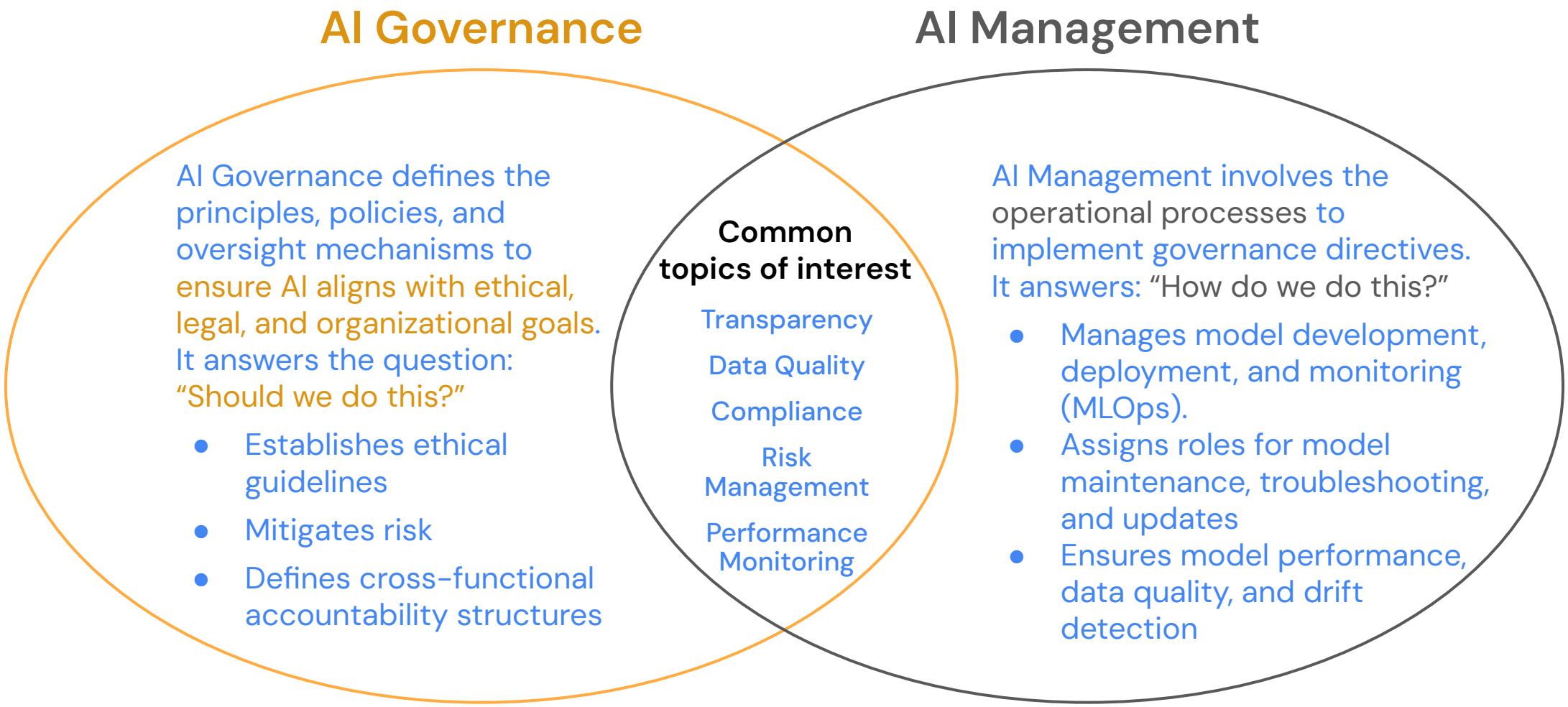
Data Governance and Data Management

A strong data governance needs a strong data management



AI Governance and AI Management

A strong AI governance needs a strong AI management



Opportunities of Data & AI Governance

A strong data governance framework can turn complexity into a strategic advantage, ensuring data remains a trusted, high-value asset

Unified Data Ecosystem	Breaking down silos and ensuring consistency across systems enables a single source of truth for better decision-making (it also leads to more efficient change management).
Standardized Data Practices	Clear definitions, formats, and policies enhance interoperability and streamline operations.
Robust Security & Compliance	Proactive governance safeguards data privacy, ensures regulatory adherence and strengthens resilience against threats.
Seamless Data Accessibility	Well-structured governance empowers teams with the right data at the right time, driving efficiency and innovation.
Clear Ownership & Accountability	Defined roles and responsibilities foster strong data stewardship, alignment with business goals, and sustained value creation.
Increased data literacy	Leading to more people being able to understand data and reasonably work with data.
Reduce duplication	Reduce duplication of data and processes reducing costs or enabling faster time to market.
Simplified collaboration	Consistent standards and policies allow easy collaboration across system and business boundaries.

Data & AI Governance Maturity

Maturity models

Refers to how advanced, structured, and effective an organization is

Maturity models are structured frameworks that help organizations assess, benchmark, and systematically improve their processes, capabilities, and overall performance.

Companies apply maturity models for different reasons:

Identify strengths/weaknesses	Pinpoint areas for improvement and leverage existing capabilities
Roadmap for improvement	Structured progression from basic to optimized practices
Benchmarking	Compare performance internally and against industry standards
Informed decision-making	Data-driven resource allocation and strategic planning
Team alignment	Unifies teams around shared goals and improvement criteria
Practical guidance	Provides actionable steps for advancing maturity
Risk management	Enhances ability to foresee and mitigate operational risks

Data and AI Governance Maturity models

Summary of characteristics collected from different maturity models

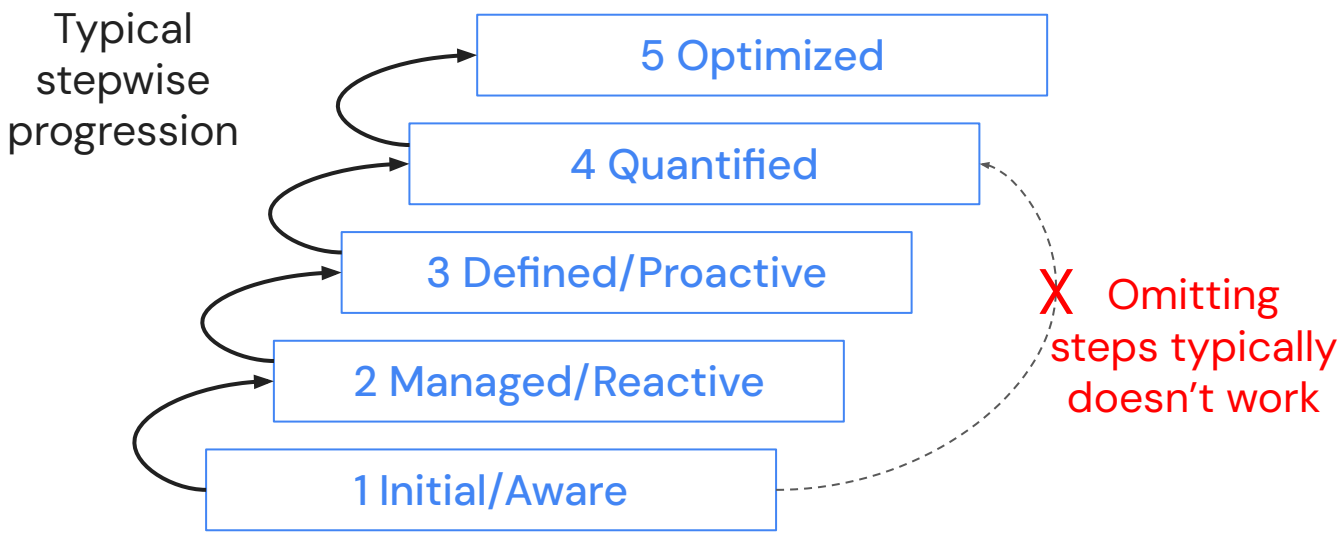
	1 Initial/Aware	2 Managed/Reactive	3 Defined/Proactive	4 Quantified	5 Optimized
DATA	• Ad-hoc and reactive data management practices • Individual efforts, uncoordinated and undocumented activities	• Value of data recognized • Governance awareness with repeatable processes • Formalized efforts remain siloed	• Enterprise-wide data governance framework is established • Consistent technology implementation across the organization	• Quantitative measuring of data governance • Outcomes, impacts, and risks can be predicted based on data patterns	• Fully embedded in the organizational culture • AI-driven automation • Governance processes evolve proactively to address changing needs

	1 Initial/Aware	2 Managed/Reactive	3 Defined/Proactive	4 Quantified	5 Optimized
AI	• Basic awareness of AI governance needs, but no structured approach • Ad-hoc AI usage with minimal or non-existent governance	• Experimenting with AI governance through informal, localized efforts • Basic policies start emerging, but inconsistent across departments	• Formalized, standardized AI governance processes are established across the enterprise • Clear roles, policies, and procedures for AI projects	• Quantitative measuring of AI governance • Governance becomes aligned with strategic objectives, and embedded in business processes	• Continuous self-optimizing improvement • Strategic advantage via responsible innovation • Proactively adapting to emerging risks and changes

Maturity models

An organization cannot omit steps

Increasing maturity typically is a step-by-step process



Data & AI Governance Capabilities

Data Governance Capabilities (1/2)

Metadata Management & Data Discovery

2 Data cataloging
Repository for all data assets, enabling easy discovery and understanding of available data

2 Business glossary
Standardized business terminology with clear ownership

2 Data modeling
Creating and managing data models to ensure consistency

3 Metadata automation
Automated capture of technical and business metadata

4 Knowledge graphs
Advanced relationship modeling and visualization with metrics

5 Semantic search
AI-powered discovery with self-improving algorithms

Master Data Management

3 Data consolidation
Standardized integration of master data

3 Entity resolution
Consistent matching algorithms for entities

3 Version control
Capabilities to track changes in data assets and governance artifacts over time

4 Hierarchy management
Advanced relationship management with metrics

Policy & Compliance Management

3 Policy creation and management
Document and manage consistent governance rules

3 Compliance tracking
Capabilities to generate documentation demonstrating regulatory adherence

3 Privacy controls
Standardized management of personal information

3 Data Retention and Archiving
Manage data lifecycle, including retention and secure disposal

3 5* Automated policy enforcement
Develop, implement, and auto-matically enforce data policies

4 Compliance reporting
Metrics-based compliance documentation

Security & Access Management

2 Identity and access management
Basic authentication controls

2 Data encryption
Standard protection for sensitive data at rest and in transit

3 RBAC / ABAC
Standardized role and attribute based permission model

3 5* Data masking
Protect sensitive data modifying its original letters and numbers

3 Audit Trails
Mechanisms to track and record all data-related activities for accountability

4 Sensitive data identification
Automated discovery with classification metrics

Data Quality Management

2 Data profiling
Analyze and understand the structure, content, and quality of data sets

3 5* Automated quality checks
Standardized rules for validation across domains

3 Data cleansing
Consistent processes for data remediation

3 Quality scoring
Metric-based assessment of quality over time

4 Monitoring & alerting
Proactive notification based on thresholds

3 5* Root cause analysis
Identify and address underlying causes of data quality issues

Data Stewardship & Workflow

2 Data Stewardship
Clear assignment of data responsibilities

3 Governance workflows
Standardized processes for governance activities

2 Collaboration tools
Structured stakeholder engagement

3 Task Management
Assign, track, and monitor data governance activities effectively

4 Change management
Measured impact assessment for data changes

1 Initial 2 Managed 3 Defined 4 Quantified 5 Optimized

* if AI driven

Data Governance Capabilities (2/2)

Data Lineage & Impact Analysis

3 Automated end-to-end lineage capture
Comprehensive tracking from source to consumption

3 Visualization tools
Standard representation of data flows

4 Impact analysis
Assess how changes in data or policies affect downstream processes and applications

3 / 5* Root cause investigation
Tracing of data related issues

Measurement & Optimization

4 Governance metrics
Comprehensive KPIs for governance effectiveness

4 Value tracking
Quantified business impact measurement

4 Customizable dashboards
Visualize and monitor key data governance metrics and KPIs

4 Maturity assessment
Benchmark current state and identify areas for improvement in data governance

5 Predictive Analytics
Anticipate potential data governance issues and take proactive measures

5 Continuous improvement
Self-optimizing governance processes

Integration & Interoperability

2 API connectivity
Basic integration interfaces

2 Data Sharing
Tools to facilitate secure data sharing within and outside the organization

3 Cross-platform governance
Consistent governance across environments (e.g. clouds)

3 Metadata exchange
Secure and consistent metadata sharing between systems

AI & Automation Support

5 AI-driven metadata management
AI-driven automated metadata capture and enrichment

5 Algorithm transparency
Advanced explainability capabilities

5 Intelligent Data Quality
AI-driven automated data quality checks and improvements

5 Anomaly detection
Predictive identification of governance issues

1 Initial 2 Managed 3 Defined 4 Quantified 5 Optimized

* if AI driven

AI Governance Capabilities (1/2)

Model Management & Lifecycle

- 2 **AI inventory & registration** — Centralized repository of all AI systems, models, and use cases
- 2 **Model documentation** — Comprehensive record of model specifications, training data, and intended uses
- 3 **Version control** — Tracking of all model iterations, updates, and changes
- 3 **Lifecycle management** — End-to-end governance from conception to retirement
- 3 **Development standards enforcement** — Implementation of consistent protocols for model building

Monitoring & Validation

- 3 **Continuous performance monitoring** — Real-time tracking of model behavior and outputs
- 3 **Drift detection** — Identification of changes in model performance over time
- 4 **Threshold alerting** — Automated notifications when metrics fall outside acceptable ranges
- 4 **Validation against benchmarks** — Comparison of performance against established standards
- 4 **Output sampling & validation** — Testing of model outputs for quality and appropriateness

Data Governance Integration

- 2 **Data quality assessment** — Evaluation of training data for completeness, accuracy
- 3 **Data lineage tracking** — Visibility into data sources and transformations
- 3 **Metadata management** — Structured organization of data about AI systems
- 4 **Master data integration** — Connection to authoritative data sources
- 4 **Data ethics evaluation** — Assessment of data collection and usage practices

Security & Privacy

- 2 **Access controls** — Fine-grained permissions for accessing AI systems and data
- 3 **Data protection mechanisms** — Safeguards for sensitive data used in AI systems
- 3 **Privacy-Preserving AI Techniques** — Tools for privacy-enhancing technologies in AI systems
- 3 **Security vulnerability scanning** — Identification of potential security weaknesses
- 4 **Secure deployment frameworks** — Protected pipelines for model deployment
- 4 **AI Security Measures** — Tools to protect AI systems from adversarial attacks and data poisoning

Audit & Reporting

- 2 **Audit trails** — Immutable records of model development, deployments, and changes
- 3 **Automated reporting** — Generation of standard and custom compliance reports
- 3 **Evidence collection** — Gathering of artifacts needed for internal and external audits
- 3 **Incident documentation** — Recording of issues and resolution actions
- 4 **Board-level reporting** — Executive dashboards summarizing governance status

Collaboration & Workflow

- 2 **Cross-functional collaboration tools** — Platforms for governance stakeholders to coordinate
- 3 **Workflow automation** — Streamlined processes for governance activities
- 3 **Role-based interfaces** — Tailored views based on user responsibilities
- 3 **Approval workflows** — Structured sign-offs for model development stages
- 4 **Knowledge sharing mechanisms** — Capabilities to share governance insights

1 Initial 2 Managed 3 Defined 4 Quantified 5 Optimized

AI Governance Capabilities (2/2)

Risk Management & Assessment

2 Model Validation and Testing
Capabilities to rigorously test AI models before deployment

3 Risk identification
Automated detection of potential risks in AI systems

3 AI risk impact analysis
Tools to determine potential business and ethical impacts of AI deployments

3 AI Incident Response
Protocols and tools for managing and mitigating AI-related incidents

4 Risk scoring
Quantitative evaluation of risk levels for prioritization

Transparency & Explainability

3 Explainability tools
Capabilities to make AI decision-making understandable

3 Interpretability visualizations
Visual representations of how AI models arrive at conclusions

3 Audit Trail for AI Decisions
Track and record AI-driven decisions for accountability

4 Decision pathway mapping
Tracing of factors influencing AI outputs

5 Technical documentation automation
Generation of model explanation artifacts

5 Stakeholder-appropriate explanations
Self-improving explanations by audience

AI Education and Training

3 AI Literacy Programs
Tools and resources to improve AI understanding across the organization

3 AI Ethics Training
Platforms for delivering and tracking completion of AI ethics education

3 AI Skills Assessment
Capabilities to evaluate and monitor AI-related skills within the organization

Fairness & Ethical AI

2 Bias detection and mitigation
Tools to identify/address potential biases in models, data

2 Fairness monitoring
Quantitative measures of model fairness across different demographics

3 Ethical guidelines implementation
Frameworks for embedding ethical principles in AI

3 Ethical AI Assessment
Tools to evaluate AI systems for alignment with ethical principles and organizational values

4 Fairness monitoring
Quantitative measures of model fairness across different demographics

Compliance & Regulatory Management

2 Compliance assessment
Evaluation of AI systems against regulatory requirements

3 Compliance documentation
Generation of evidence for regulatory submissions

3 Compliance reporting
Capabilities to generate reports demonstrating adherence to AI-specific regulations

4 AI Regulatory Tracking
Tools to monitor and interpret evolving AI regulations globally

5 Automated policy enforcement
Automated implementation of regulatory requirements

Stakeholder Engagement & Impact Assessment

2 Stakeholder Feedback Mechanisms
Tools for gathering & analyzing feedback on AI systems

4 Public Trust Monitoring
Tools to gauge and track public perception and trust in the organization's AI initiatives

4 AI Impact Assessment
Frameworks for evaluating the societal and environmental impact of AI systems

1 Initial 2 Managed 3 Defined 4 Quantified 5 Optimized

The trade-off of governance and agility

Applying the right level of governance

Risk-based approach to trade off accuracy and compliance vs. agility

High risk data and applications

For high-risk data (such as sensitive personal information, financial data, or data subject to strict regulatory requirements) where violations can lead to e.g. severe fines or reputation damage, a more stringent governance framework is necessary.
This includes:

Data

- Rigorous data quality and governance standards
- Stringent access controls
- Comprehensive audit trails
- Regular compliance checks

AI

- Rigorous risk management processes
- Stringent data quality and governance standards
- Consistent recordkeeping
- Transparency and provision of information to users
- Guaranteed human oversight
- Robust accuracy, security, and cybersecurity measures

⇔

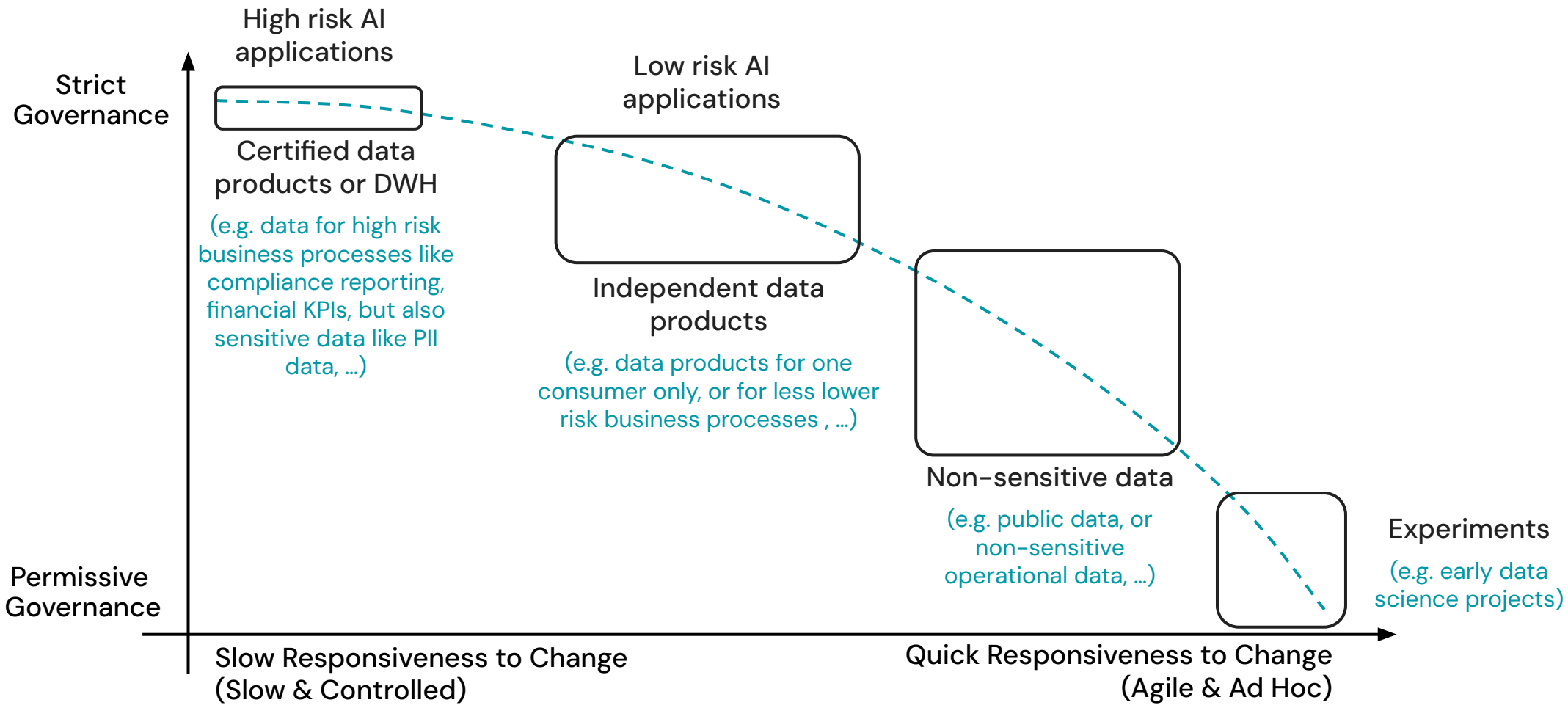
Low risk data and applications

For low-risk data and applications, such as publicly available information, or non-sensitive operational data/applications, a looser governance approach can be applied

Typical factors determining the risk of an application

- Industry-specific regulatory requirements
- Potential impact on individuals and the organization
- Scale of data usage and number of users affected
- Sensitivity and criticality of the data

Applying the right level of governance



General Considerations for Data and AI Governance

- Data & AI governance is not a tactical solution to a problem, it **is a strategy** that the business pursues.
- You **cannot buy data & AI governance**, it is something you have to implement across people, processes and technology.
- While the data & AI governance strategy may be the responsibility of a single team, the implementation and **success depends on all teams**.
- Also **focus on the opportunities** (business value) when building a data & AI governance strategy
- **Start simple** and implement the requirements that are immediately needed, adding advanced topics over time. Use the data & AI strategy to guide this process.
- **Carefully choose the level** of data & AI governance for each business unit, data set, and AI application (loose & agile vs. strict & slow).

Data & AI Governance Roles and Tasks

Typical tasks involved in data and AI governance

Separated by data/AI and by definition/operations*

	Data Governance	AI Governance
Definition	Develop and approve data governance policies, standards, and frameworks. Define data quality metrics, data definitions, and data classification schemes. Align data governance strategy with business goals and regulatory requirements. Oversee the creation and maintenance of data governance documentation (policies, procedures, standards). Set strategic direction and provide oversight for the data governance program. Coordinate with legal, compliance, and IT to ensure policies address all relevant risks and regulations.	Develop and approve AI governance frameworks, policies, and ethical guidelines. Define standards for responsible AI use, risk management, transparency, and accountability. Establish criteria for AI system approval, risk assessment, and regulatory compliance. Oversee the alignment of AI initiatives with organizational values and legal requirements. Set up processes for ongoing review and adaptation of AI governance policies in response to technological and regulatory changes. Provide cross-functional oversight, drawing on expertise from technical, legal, business, and ethical domains.
Operations	Implement and enforce data governance policies and standards at the operational level. Monitor data quality, resolve data discrepancies, and maintain data integrity. Manage data access controls and ensure data security and privacy compliance. Oversee data lifecycle management, including data retention, archiving, and disposal. Support business users in understanding and applying data governance rules. Document and report on data governance compliance and issues	Implement AI governance policies in the development, deployment, and monitoring of AI systems. Conduct risk assessments, document AI models, and maintain audit trails for AI systems. Monitor AI system performance, fairness, and compliance with ethical and legal standards. Perform independent reviews and validations of AI models, especially for high-risk applications. Respond to incidents, update models, and manage the lifecycle of AI systems in accordance with governance requirements. Communicate AI governance requirements and outcomes to stakeholders, including reporting to senior leadership and regulatory bodies.

* This list is not comprehensive and not every organisation needs every task accomplished

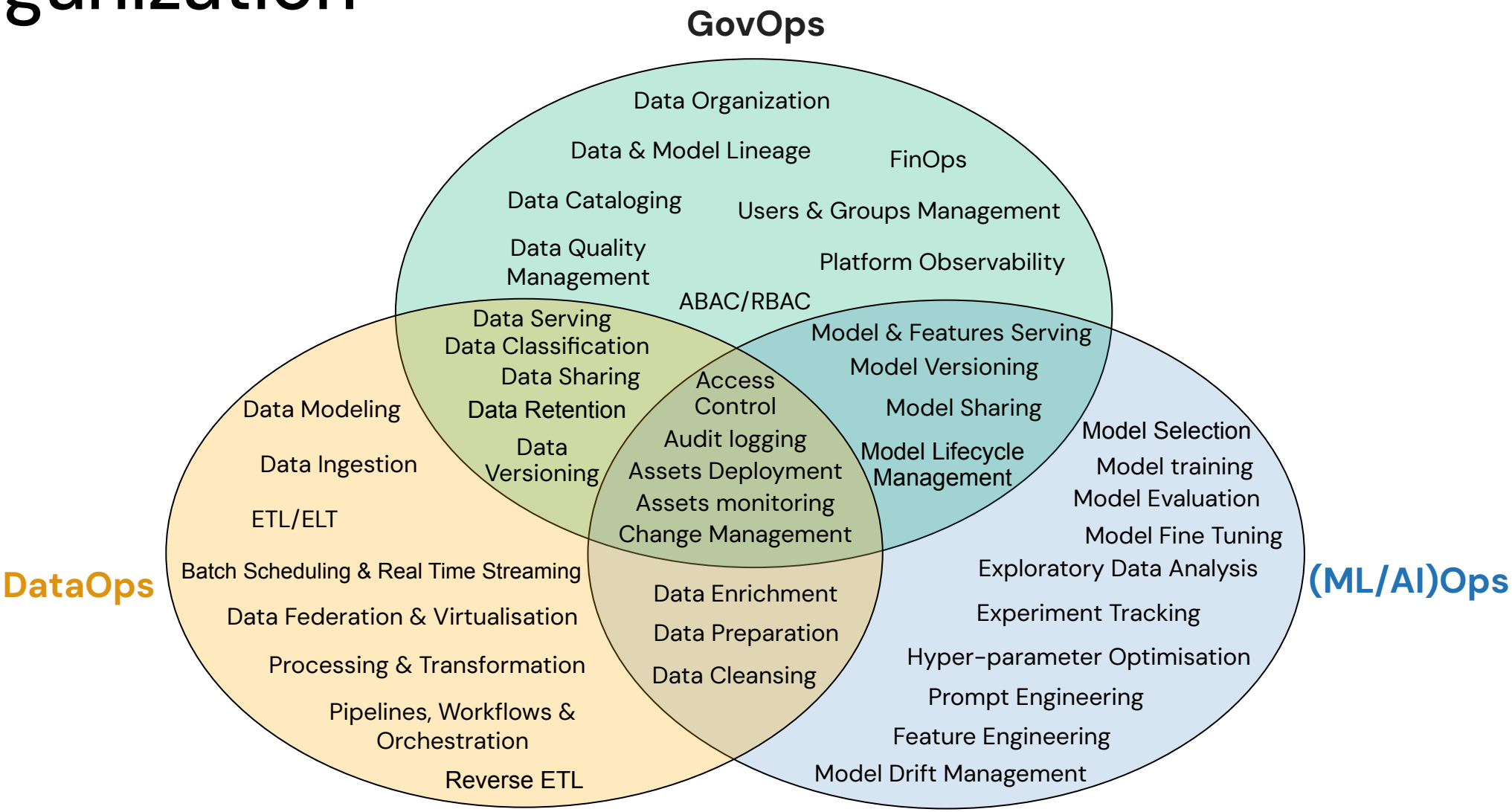
Roles involved in data and AI governance

Separated by data/AI and by definition/operations*

	Data Governance	AI Governance
Definition	**Chief Data Officer (CDO):** Senior executive responsible for overall data strategy, governance, and value realization. **Data Governance Council/Committee:** Cross-functional group that sets and oversees data governance policies and priorities. **Data Governance Manager/Director:** Leads the design and implementation of data governance frameworks and initiatives. **Data Architect:** Designs and maintains the organization's data structures, ensuring alignment with governance standards.	**Head of AI Governance / AI Governance Officer:** Leads the development and oversight of AI governance strategy and policies. **AI Governance Board/Committee:** Multi-disciplinary group that defines and reviews AI governance frameworks and standards. **Governing Body (e.g., Board, Executive Committee):** Provides executive oversight and approval for AI governance initiatives. **Legal and Compliance Officers:** Ensure AI governance policies comply with laws and regulations. **Ethics Officer or AI Ethics Committee:** Guides the ethical development and deployment of AI systems.
Operations	**Data Steward:** Manages data quality, integrity, and usage within specific domains or business units. **Data Custodian:** Handles the technical management and safeguarding of data assets. **Data Owner:** Has ultimate accountability for a specific data set, including its access and quality. **Data Administrator:** Maintains databases and ensures data is stored, organized, and accessible as per governance rules. **IT/Data Security Officer:** Ensures data security policies are implemented and regulatory requirements are met.	**AI Product/Application Owner:** Responsible for the lifecycle and compliance of specific AI products or applications. **AI System/Model Developer (Data Scientist, ML Engineer):** Designs, builds, and maintains AI models in accordance with governance policies. **Application Reviewer/Model Validator:** Independently assesses AI models for accuracy, fairness, and regulatory compliance. **Compliance/Legal Specialist:** Monitors and enforces adherence to legal and regulatory requirements in AI operations. **AI Operations Manager:** Oversees the deployment, monitoring, and maintenance of AI systems in production. **Technical Specialist:** Provides technical expertise to ensure AI systems meet governance, security, and performance standards.

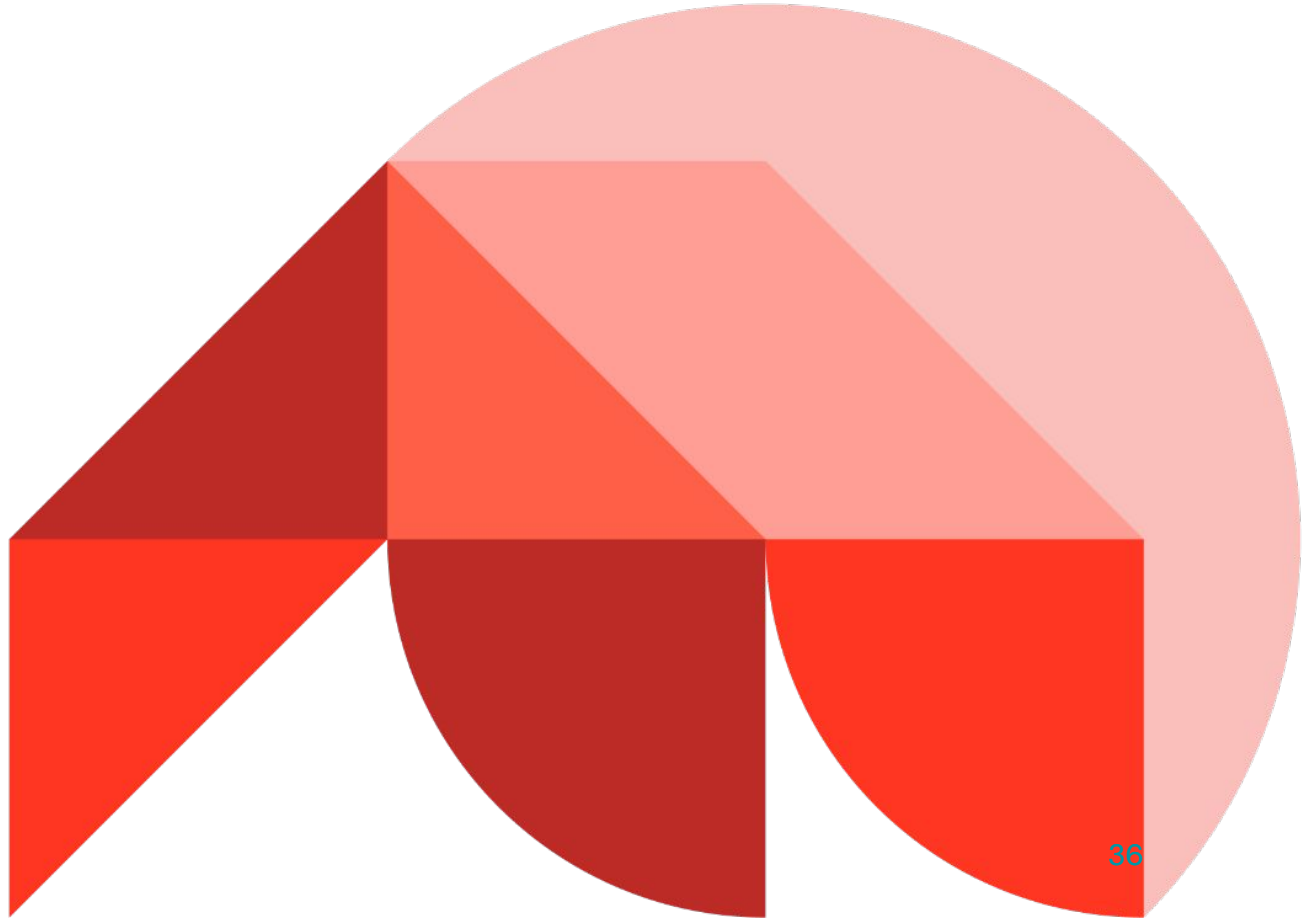
* This list is not comprehensive and not every organisation needs every role

Successful Data & AI governance operations depends on other Ops practices in the organization



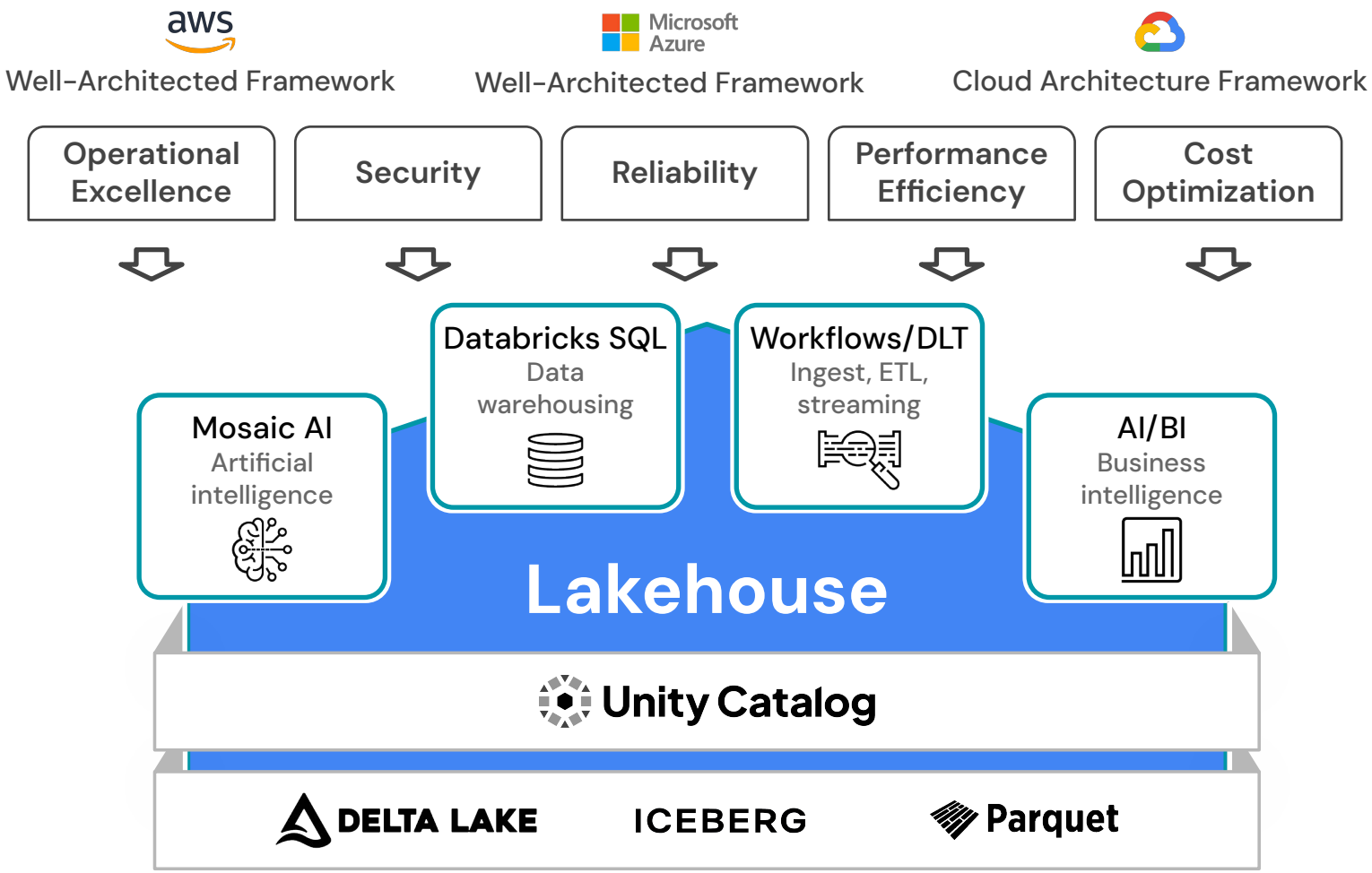
databricks

Well-Architected Framework

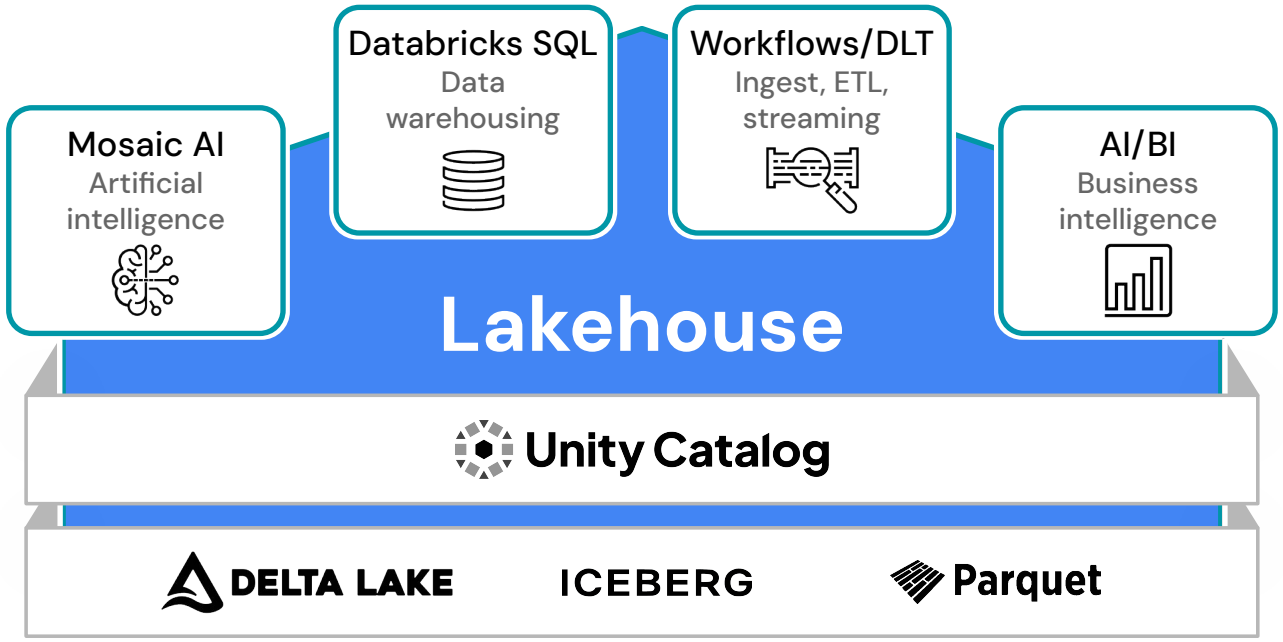


The Well-Architected Lakehouse

Extends the Cloud Well-Architected Frameworks to the Lakehouse



The Well-Architected Lakehouse



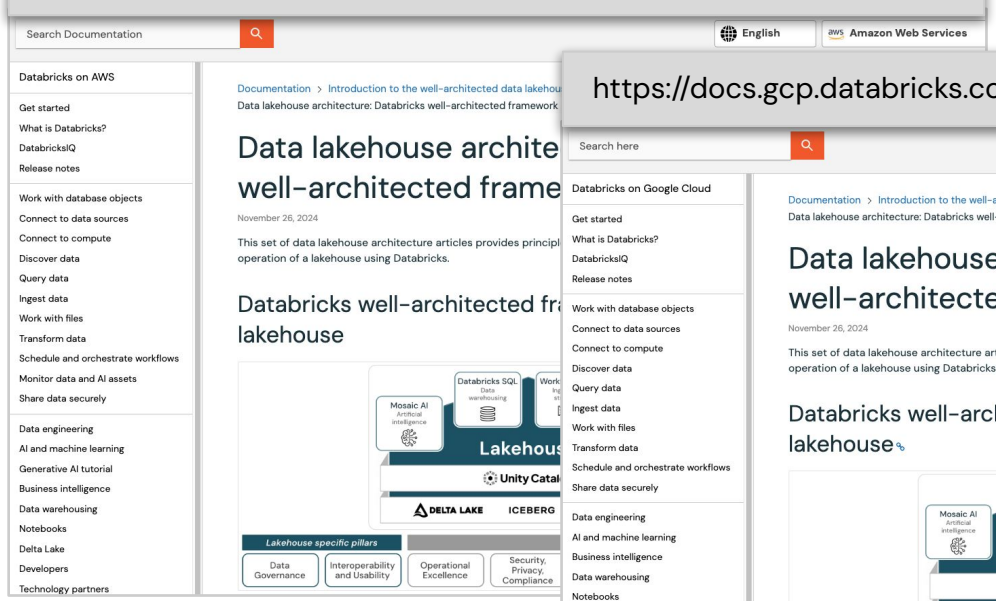
	Lakehouse specific pillars		*Common pillars*				
Pillars	Data Governance	Interoperability and Usability	Operational Excellence	Security, Privacy, Compliance	Reliability	Performance Efficiency	Cost Optimization
Scope	Ensure that data brings value and supports your business strategy	The ability of the Lakehouse to interact with users and other systems	All operations processes that keep the Lakehouse running in production	Protect Databricks application, customer workloads and customer data from threats	The ability of a system to recover from failures and continue to function	The ability of a system to adapt to changes in load	Managing costs to maximize the value delivered

Documented principles and best practices

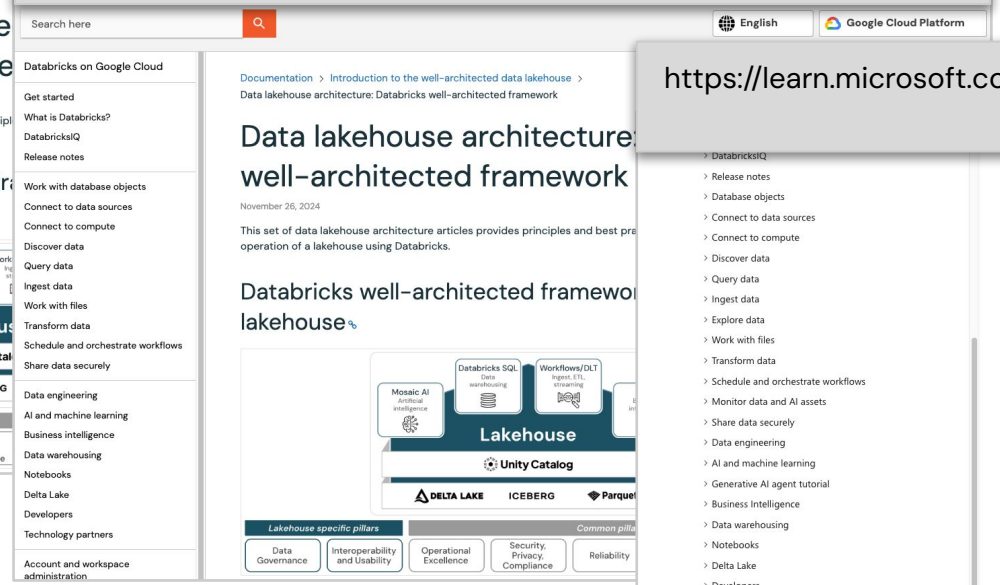
Data Governance | Interoperability and Usability | Operational Excellence | Security, Privacy, Compliance | Reliability | Performance Efficiency | Cost Optimization

Principles and Best Practices

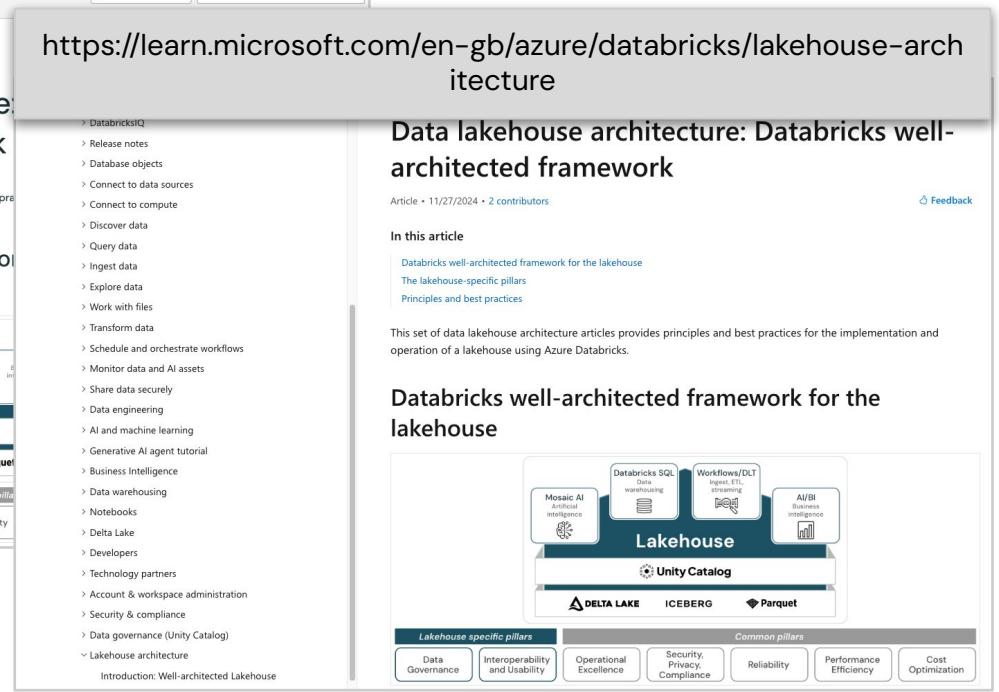
https://docs.databricks.com/lakehouse-architecture/index.html



https://docs.gcp.databricks.com/lakehouse-architecture/index.html

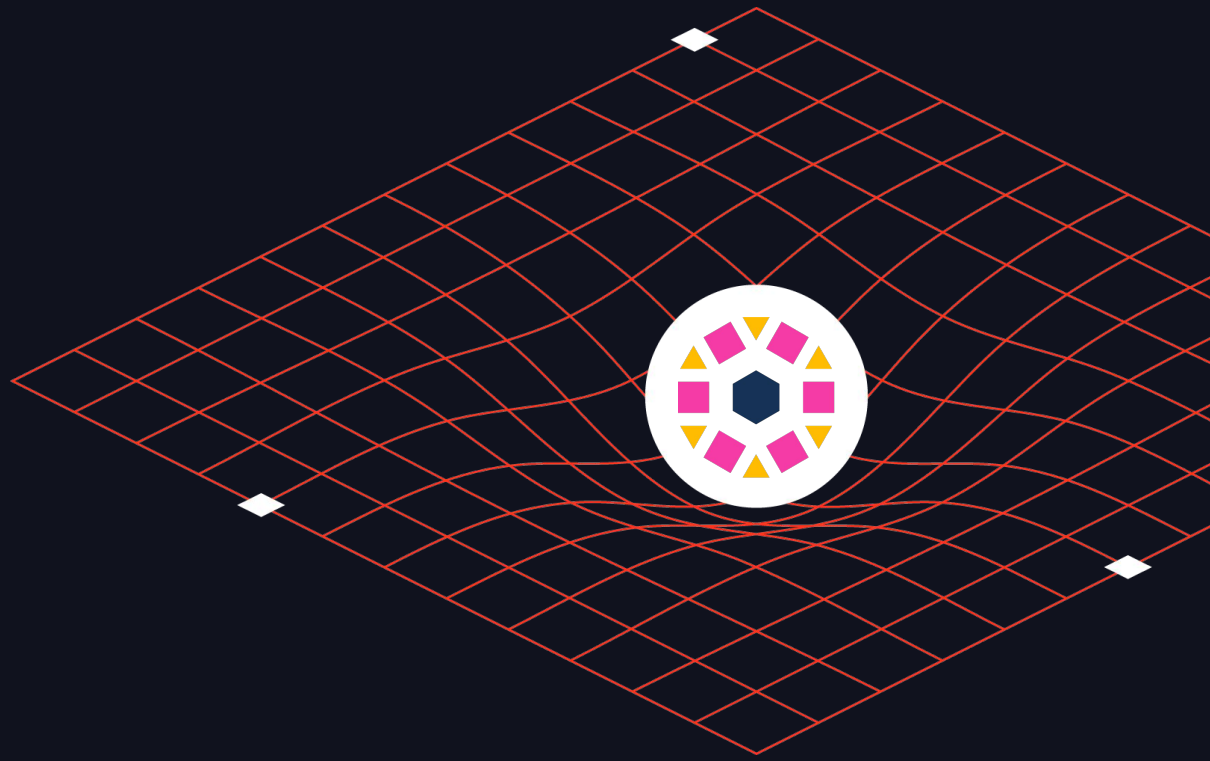


https://learn.microsoft.com/en-gb/azure/databricks/lakehouse-architecture

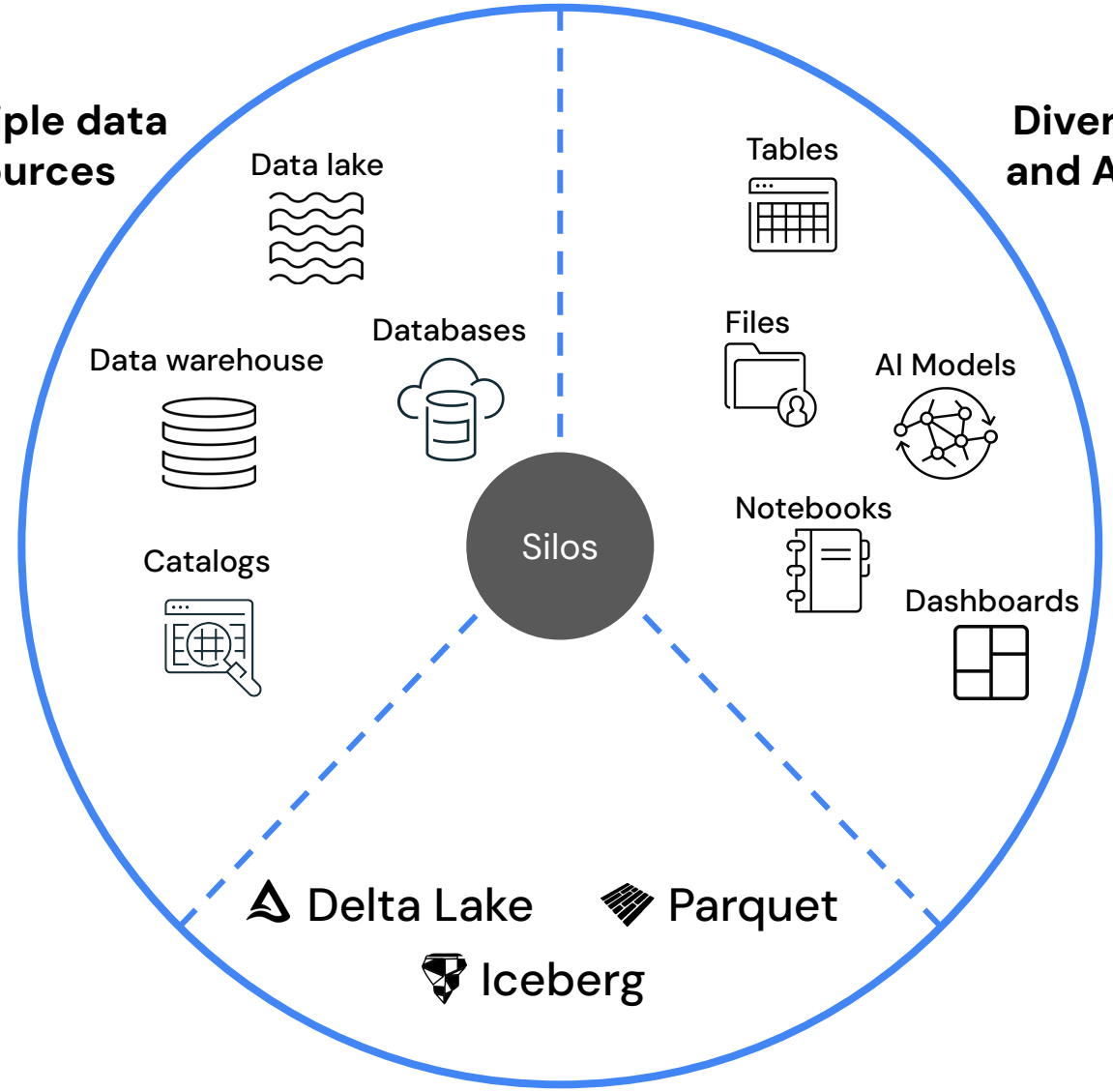


Unity Catalog

The foundation of the
Data Intelligence Platform



But governance of the entire data estate is hard



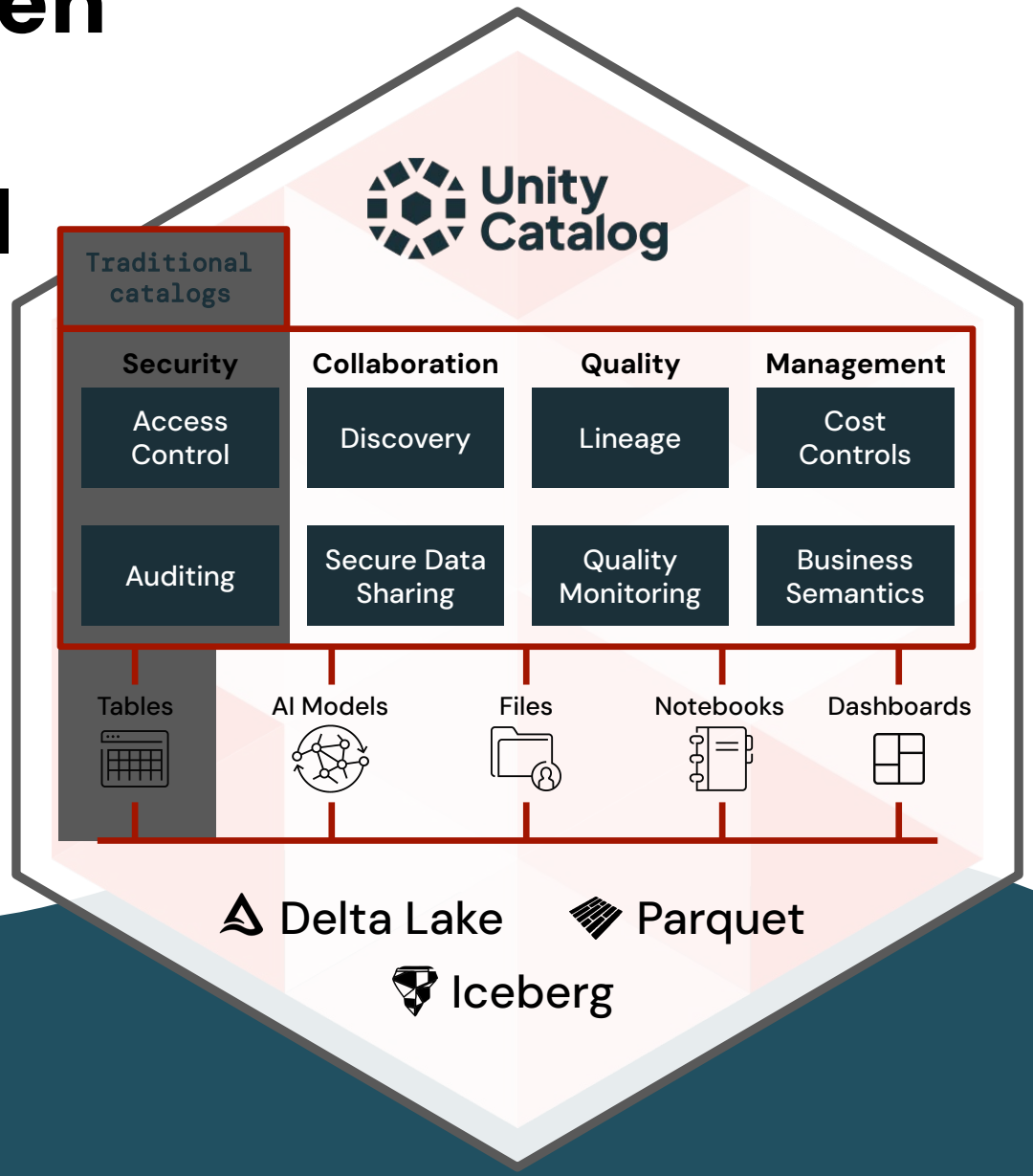
Fragmented governance for data and AI
Across access management, auditing, monitoring, observability, and lineage

Lack of open connectivity
For access and cross-platform collaboration

Lack of built-in intelligence
For data discovery, understanding, and gaining domain insights

Unified and open governance for all data + AI

Connect to any external data source

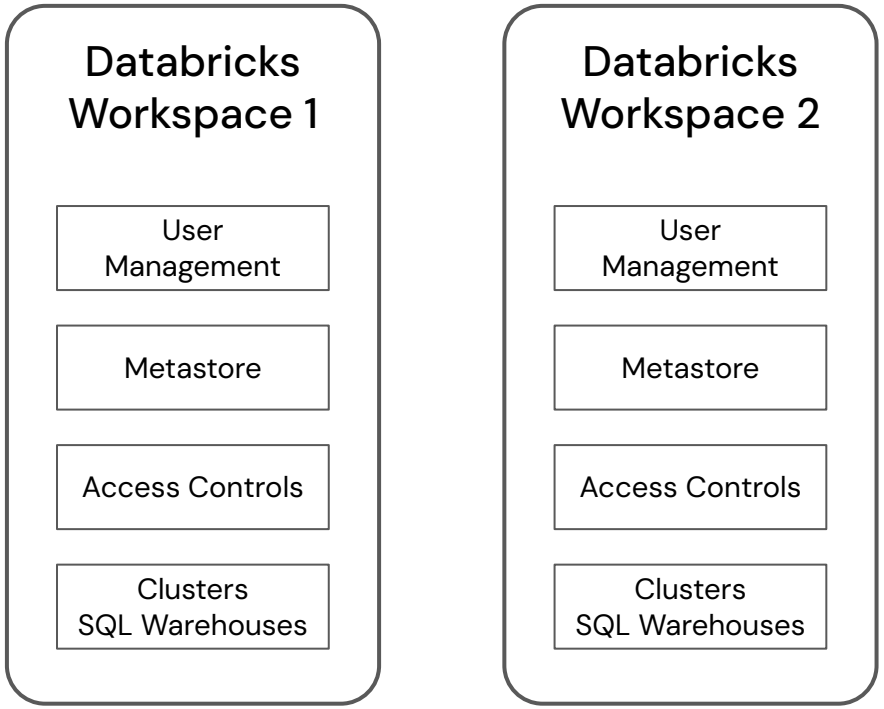


Open access and collaboration with any tool, engine or platform

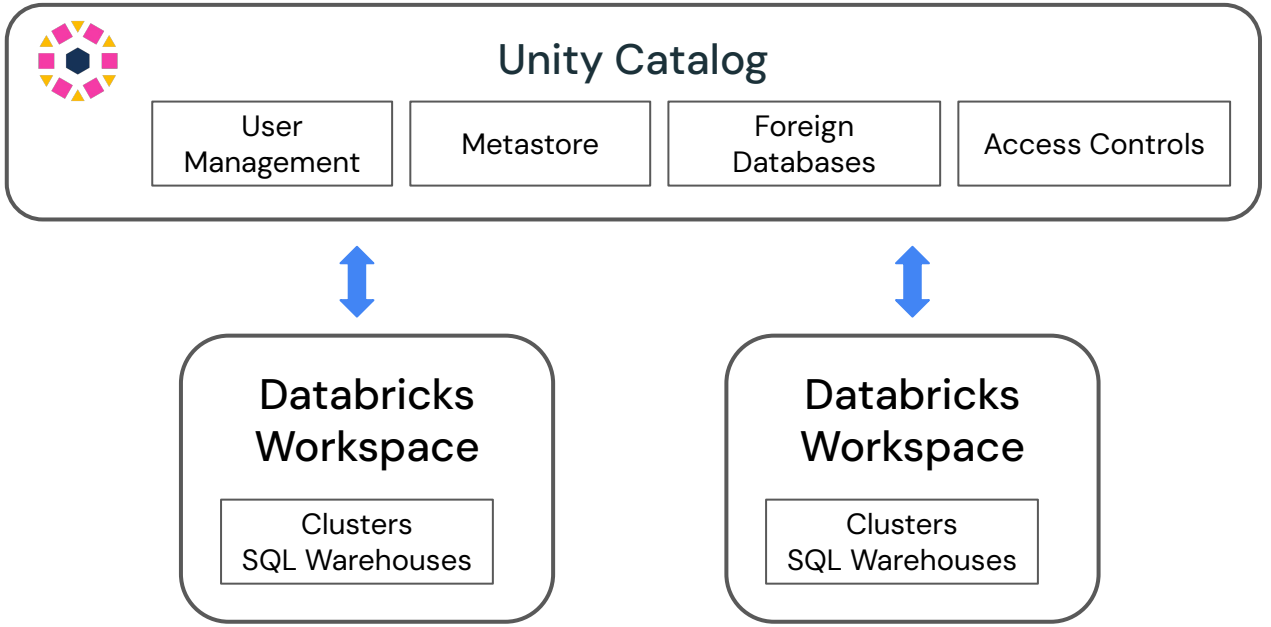
All your metadata, in one place

One metadata layer across file and database sources **superpowers** governance

Without Unity Catalog



With Unity Catalog



Fundamental Concepts

Working with file based data sources

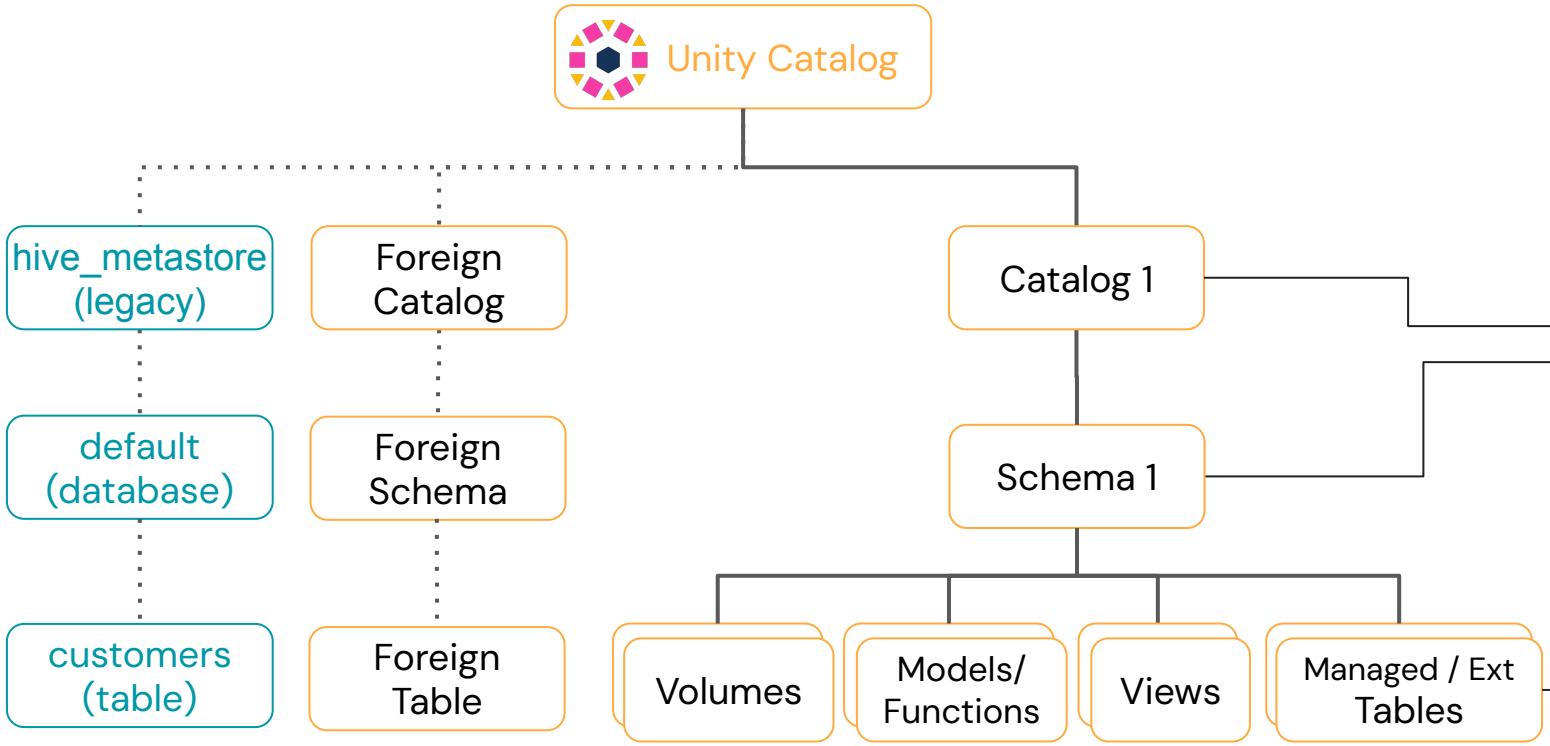
- Credentials
  - Storage Credential to connect to storage
  - Cloud Service Credential to connect native services
- External Locations
  - Storage location used for tables, files or volumes
- Tables
  - For managing tabular data
- Volumes
  - File container to govern a folder on cloud storage

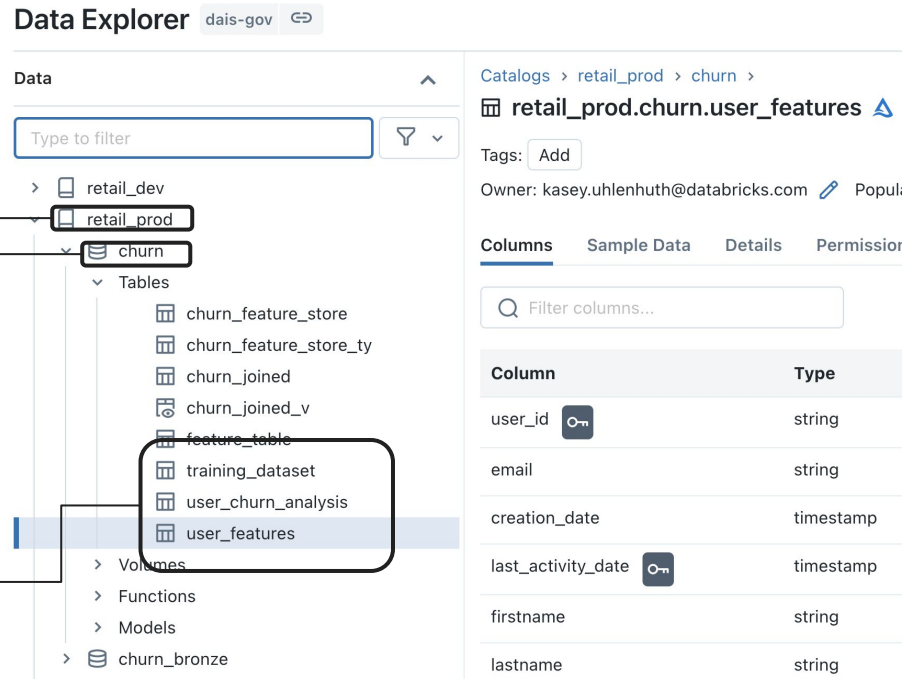
Working with databases

- Connections
  - Credential and connection information to connect to an external database
- Foreign Catalogs
  - A catalog that represents an external database in UC and can be queried alongside managed data sources and file sources

Governed namespace across file and database sources

Access legacy metastore and foreign databases powered by Lakehouse Federation






```
SELECT * FROM main.paul.red_wine;  -- <catalog>.<database>.<table>

SELECT * FROM hive_metastore.default.customers;

SELECT * FROM snowflake_warehouse.some_schema.some_table;
```

Centralized Access Controls

Centrally grant and manage access permissions across workloads

Using ANSI SQL DCL

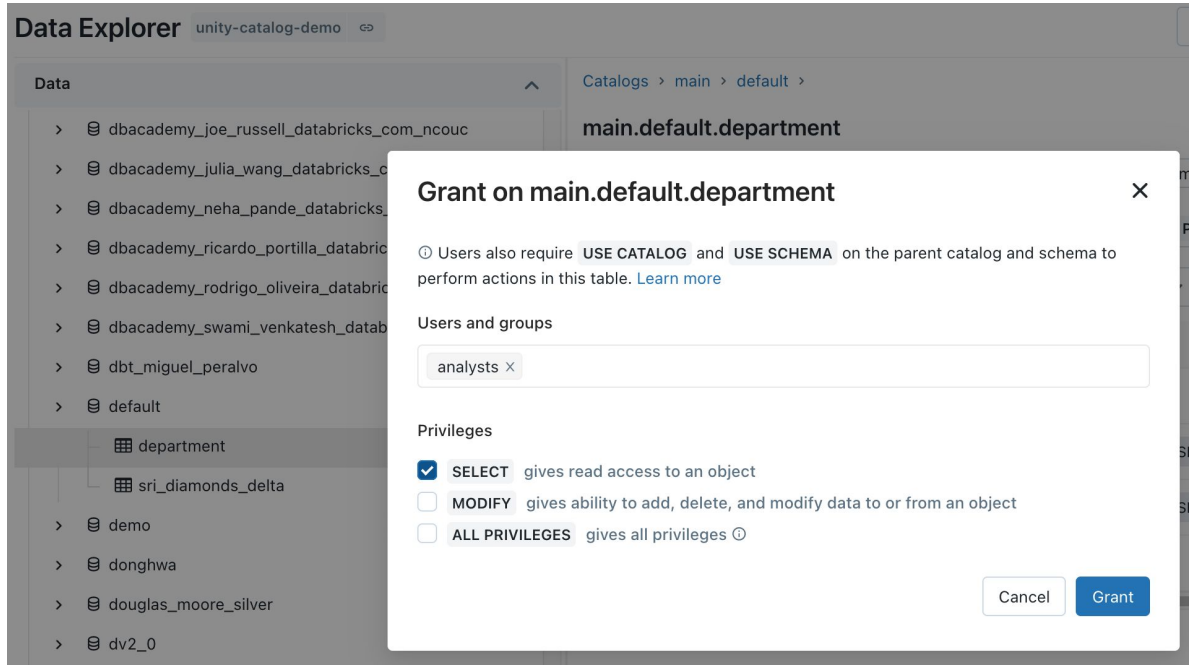
```
GRANT <privilege> ON <securable_type>
<securable_name> TO `<principal>`

GRANT SELECT ON iot.events TO engineers
```

Choose permission level

'Table'= collection of files in S3/ADLS

Sync groups from your identity provider

Using UI



Row Level Security and Column Level Masking

Provide differential fine grained access to datasets

Only show specific rows

```
CREATE FUNCTION <name> ( <parameter_name >
<parameter_type> .. )
RETURN {filter clause whose output must be a boolean}


CREATE FUNCTION us_filter(region STRING)
RETURN IF(is_account_group_member ('admin'), true,
region="US");

ALTER TABLE sales SET ROW FILTER us_filter ON region;
```

Test for group at the account level

Assign reusable filter to table

Specify filter predicates

Mask or redact sensitive columns

```
CREATE FUNCTION <name> (<parameter_name>,
<parameter_type>, [, <column>...])
RETURN {expression with the same type as the first
parameter}

CREATE FUNCTION ssn_mask(ssn STRING)
RETURN IF(is_account_group_member('admin'), ssn, "****");

ALTER TABLE users ALTER COLUMN table_ssn SET MASK
ssn_mask;
```

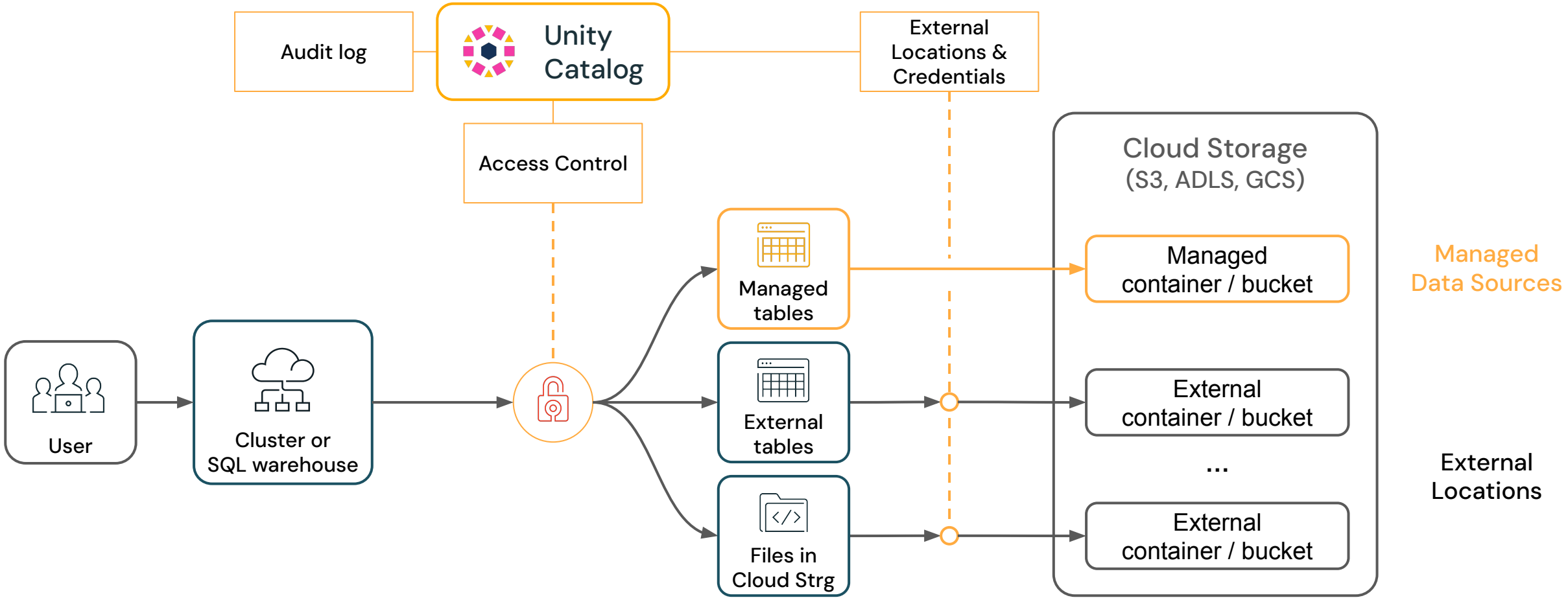
Test for group at the account level

Assign reusable mask to column

Specify mask or function to mask

Managed Data Sources & External Locations

Simplify data access management across clouds



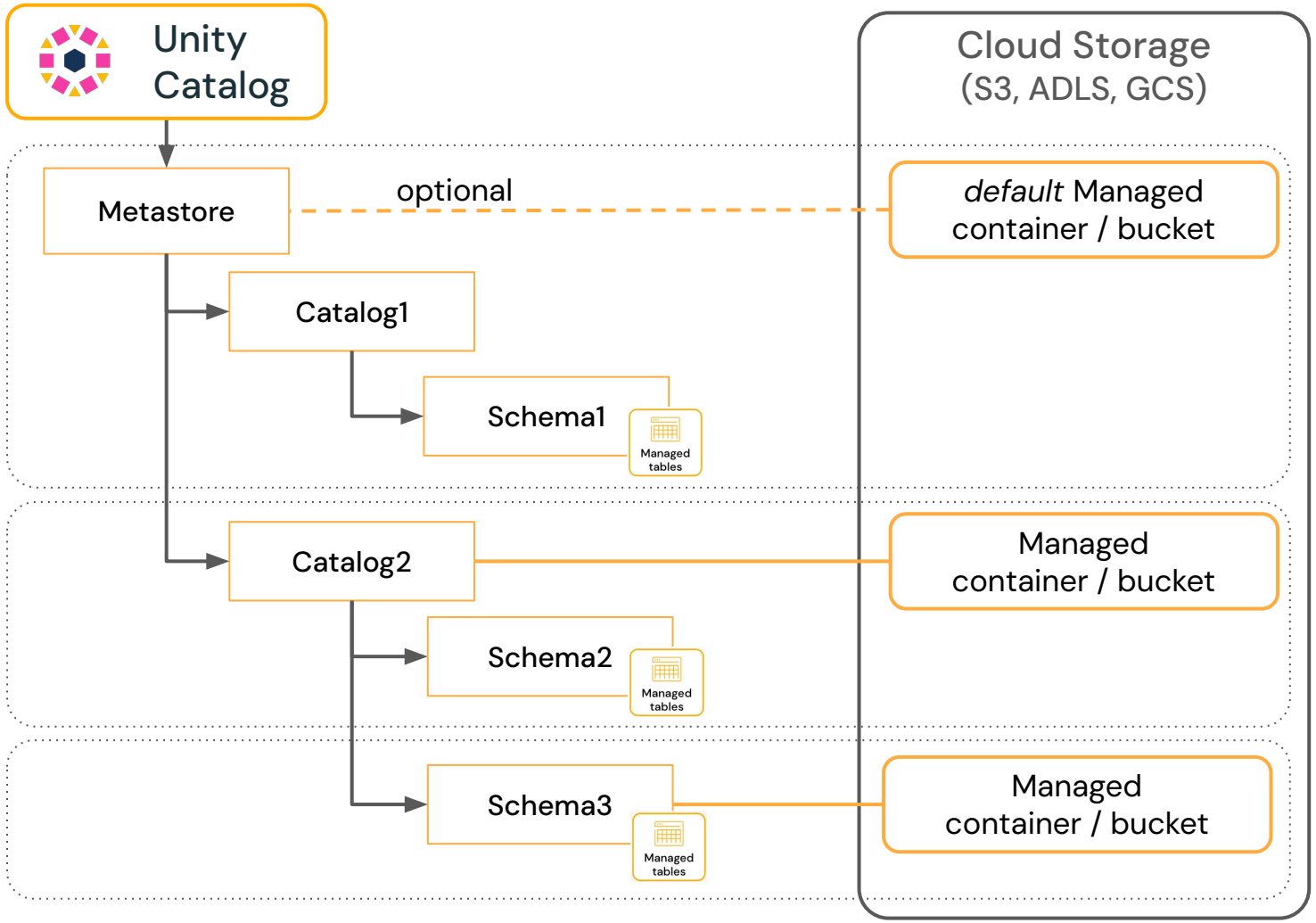
Default access to storage by catalog or schema

Use managed data sources for data isolation or cost allocation

1) default storage associated with the **metastore**

2) define at the **catalog**

3) define at the **schema**

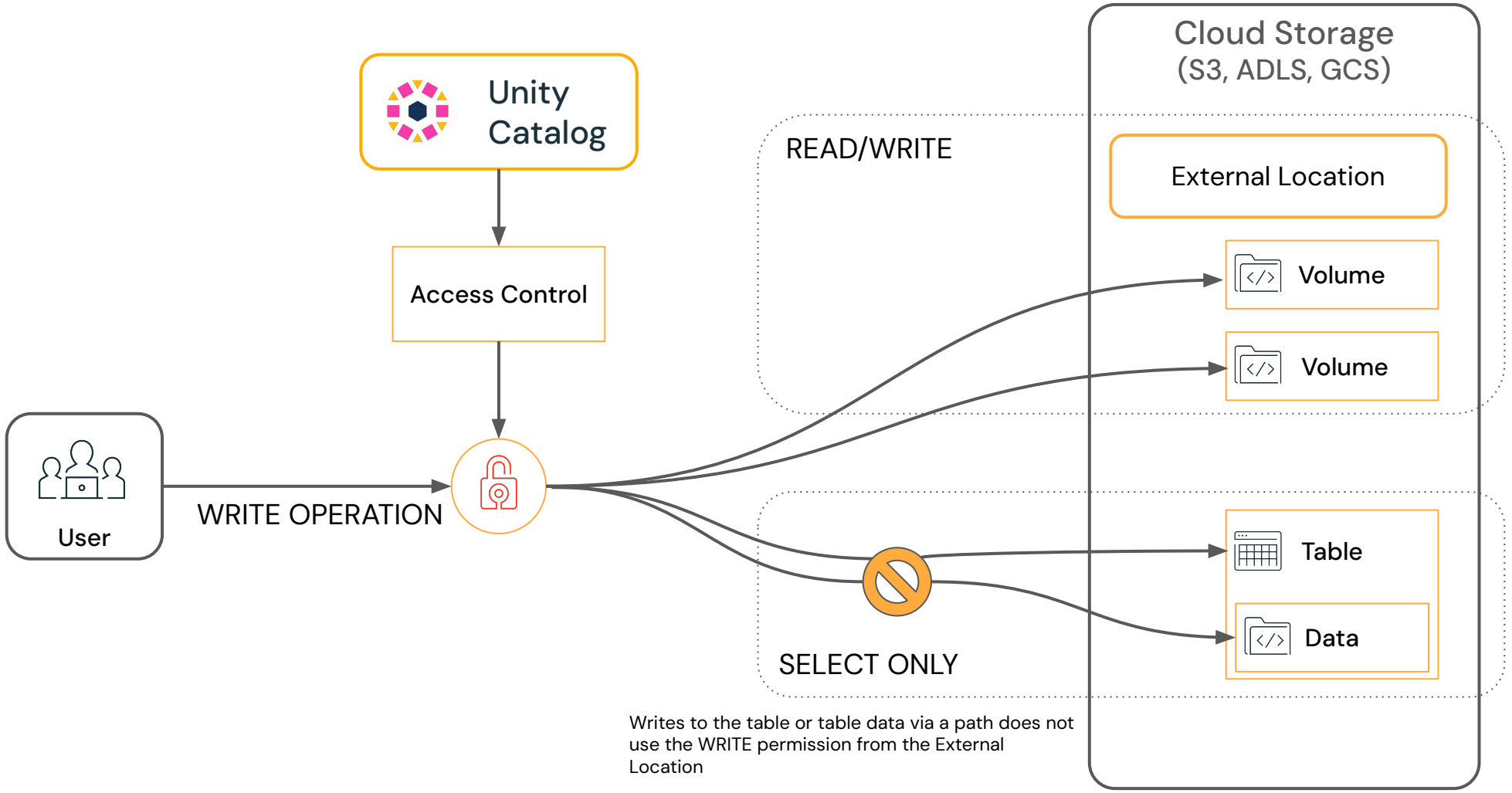


strong recommendation: do *not* store catalogs/schemas in the metastore-level default location

.
.
.

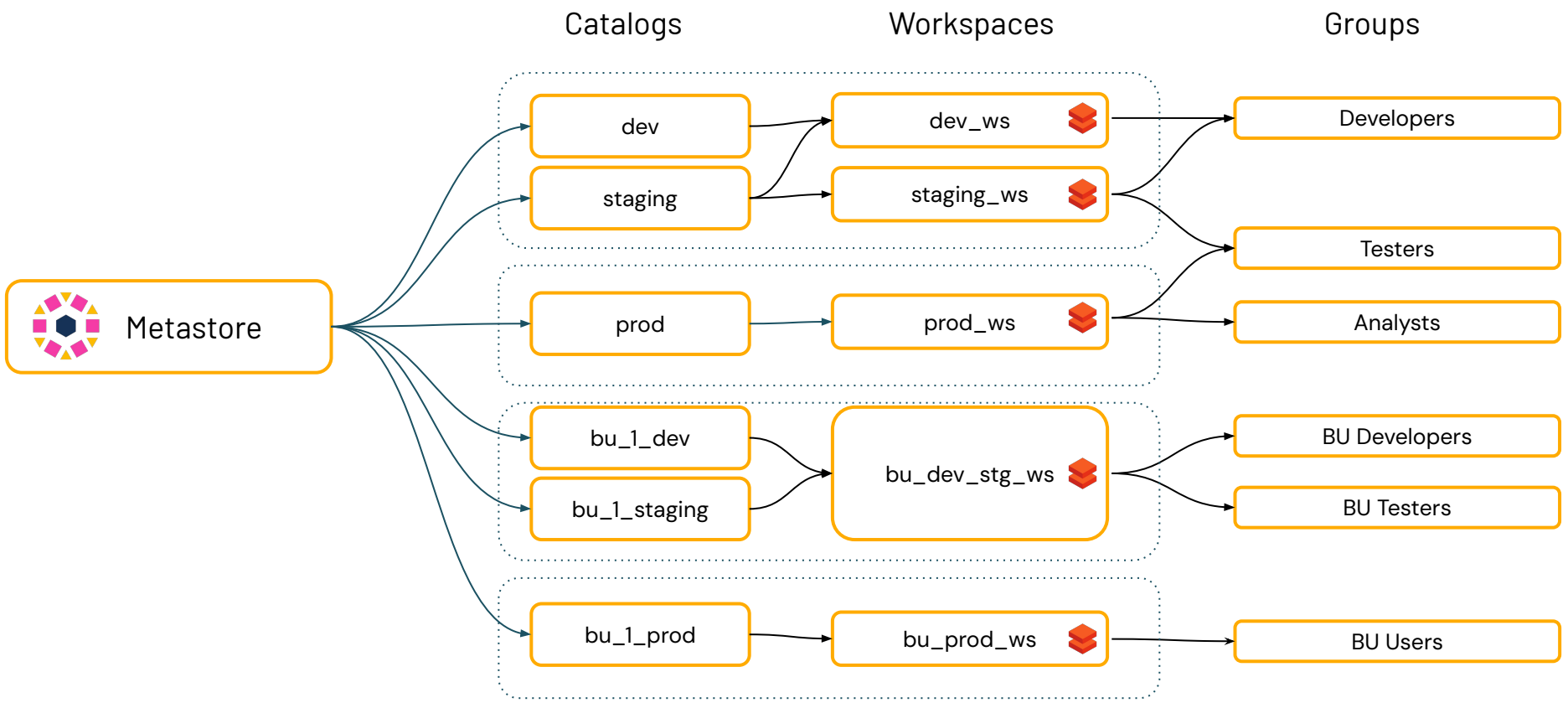
rather, specify a separate managed storage location for each catalog (which becomes the default location for schemas/tables w/in that catalog)

Govern filesystems and objects distinctly



Access data from specified environments only

Restrict data access by environment or purpose



Access to data and availability of data can be isolated across workspaces and groups

databricks

Resources:
Public Docs

Attribute Based Access Controls

Scaling Access Management

Governance at scale

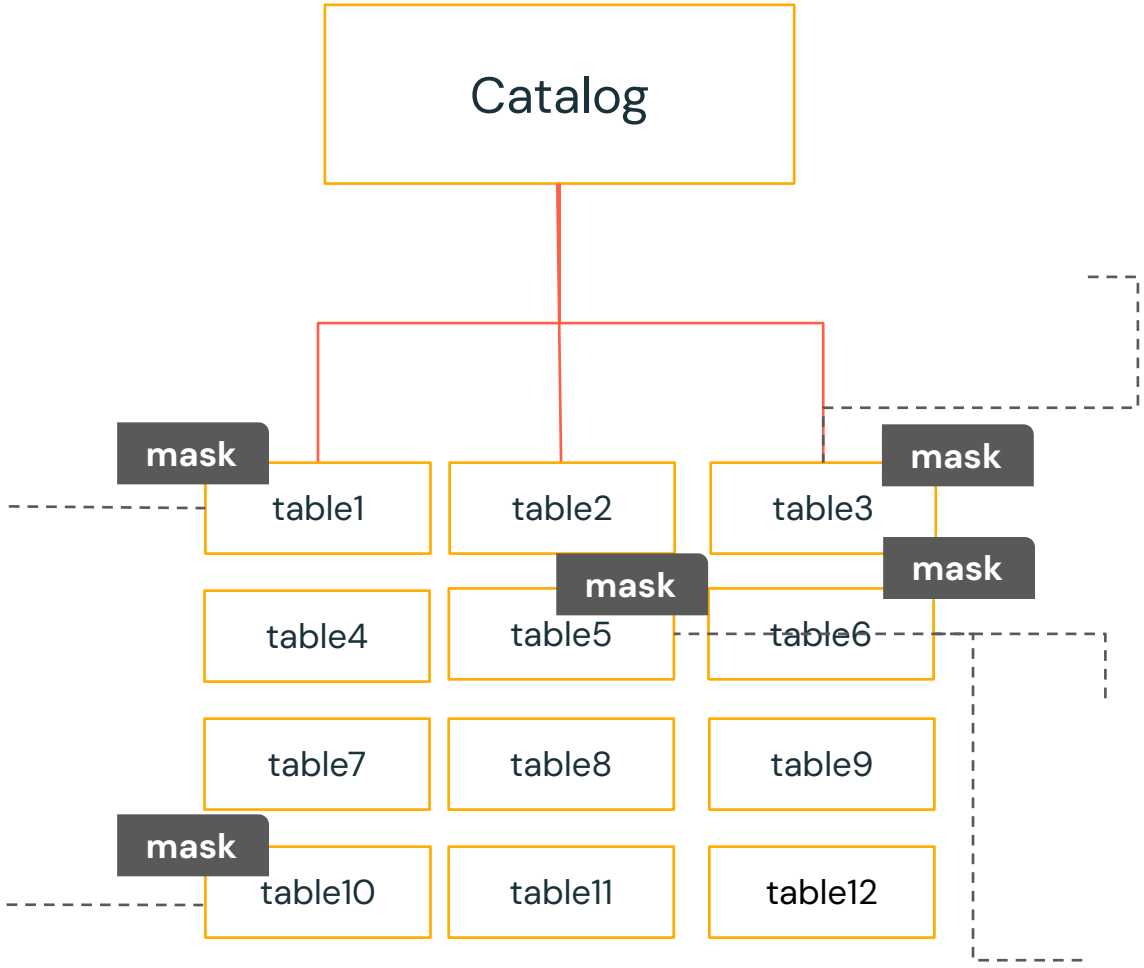
As user and data volumes grow, governance is challenged

Use Case

Mask all columns containing ***PII***

Challenges

1. Manual and slow
2. Inconsistent
3. Conflates responsibilities



Governance at scale

Our vision with Attribute Based Access Controls

1

High leverage governance.

Define rules once and apply consistently everywhere.

2

Intelligent.

Automation scales tagging and learns from usage over time, continuously improving accuracy and reducing human effort.

3

Clear roles, shared responsibility.

Governance teams define policies. Data owners (w/ AI) tag assets.

What is ABAC?

ABAC 101

ABAC: Access control that is conditional based on properties of the user, resource, and the request.

Step #1

Data Steward / AI

`Tags columns & tables`

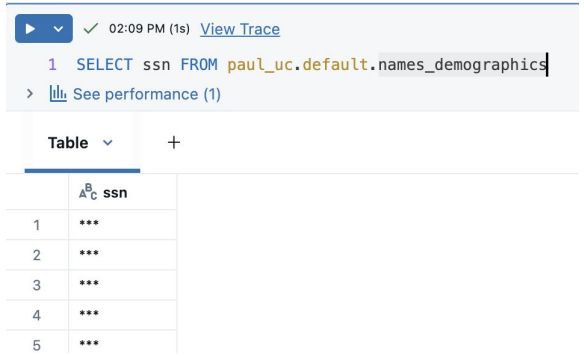
Step #2

Governance admin

`Creates a Policy to define data access rules based on tags`

Step #3

Data Analyst



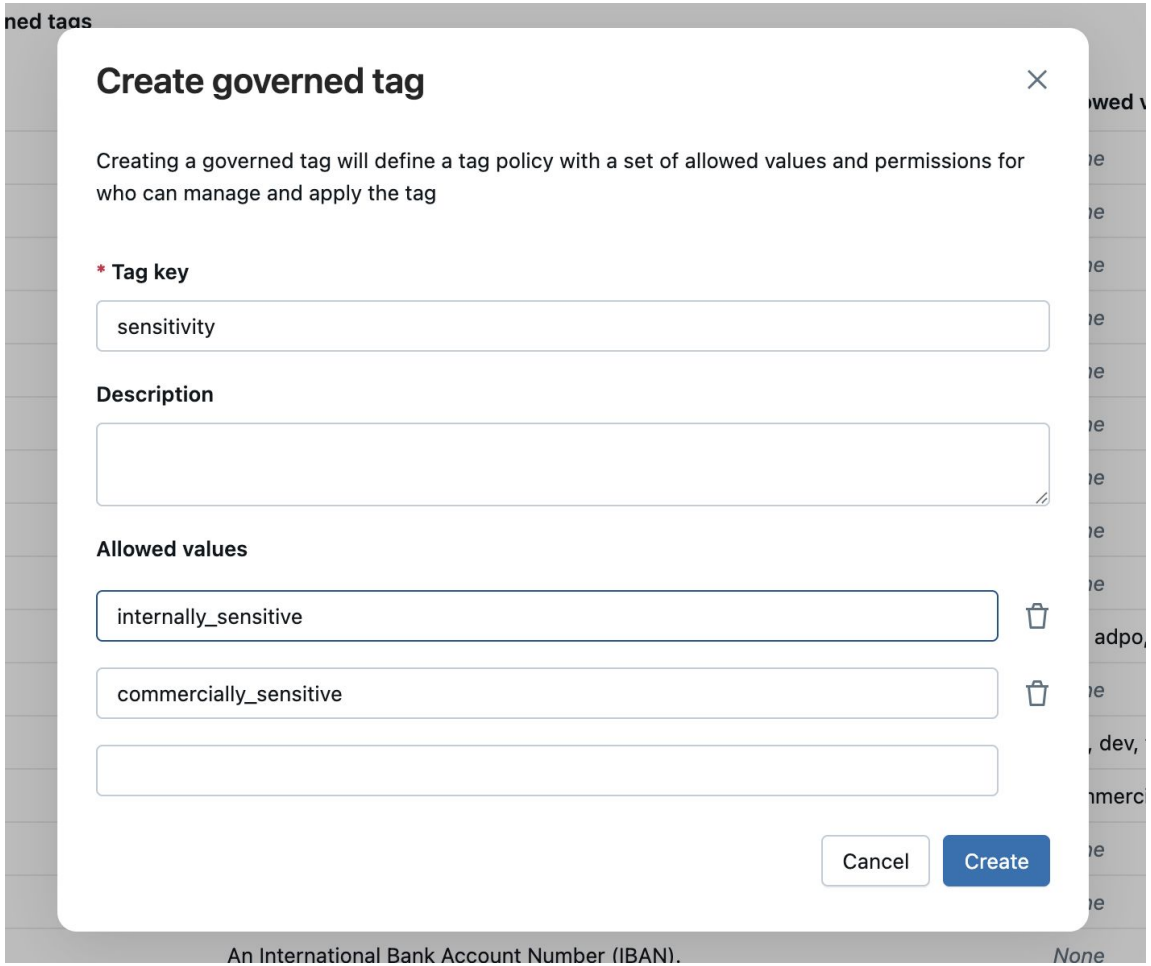
ABAC

How does it work?

Governed Tags

The foundation of ABAC

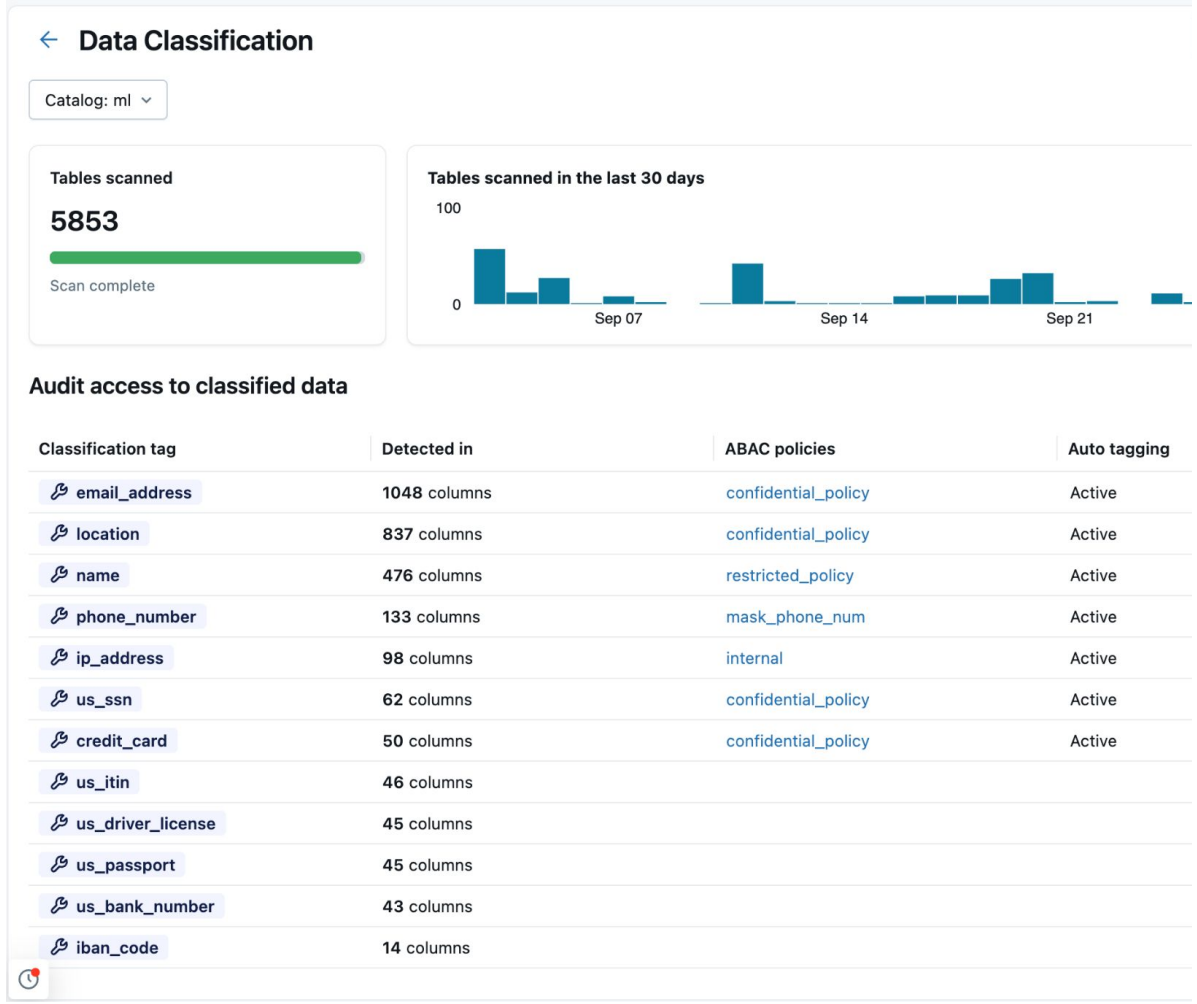
- Rules to define **tag composition**.
- **Permissions** on who can use the tags.
- Supports **many use cases**: ABAC, Discovery, Cost Observability, ...



Data Classification

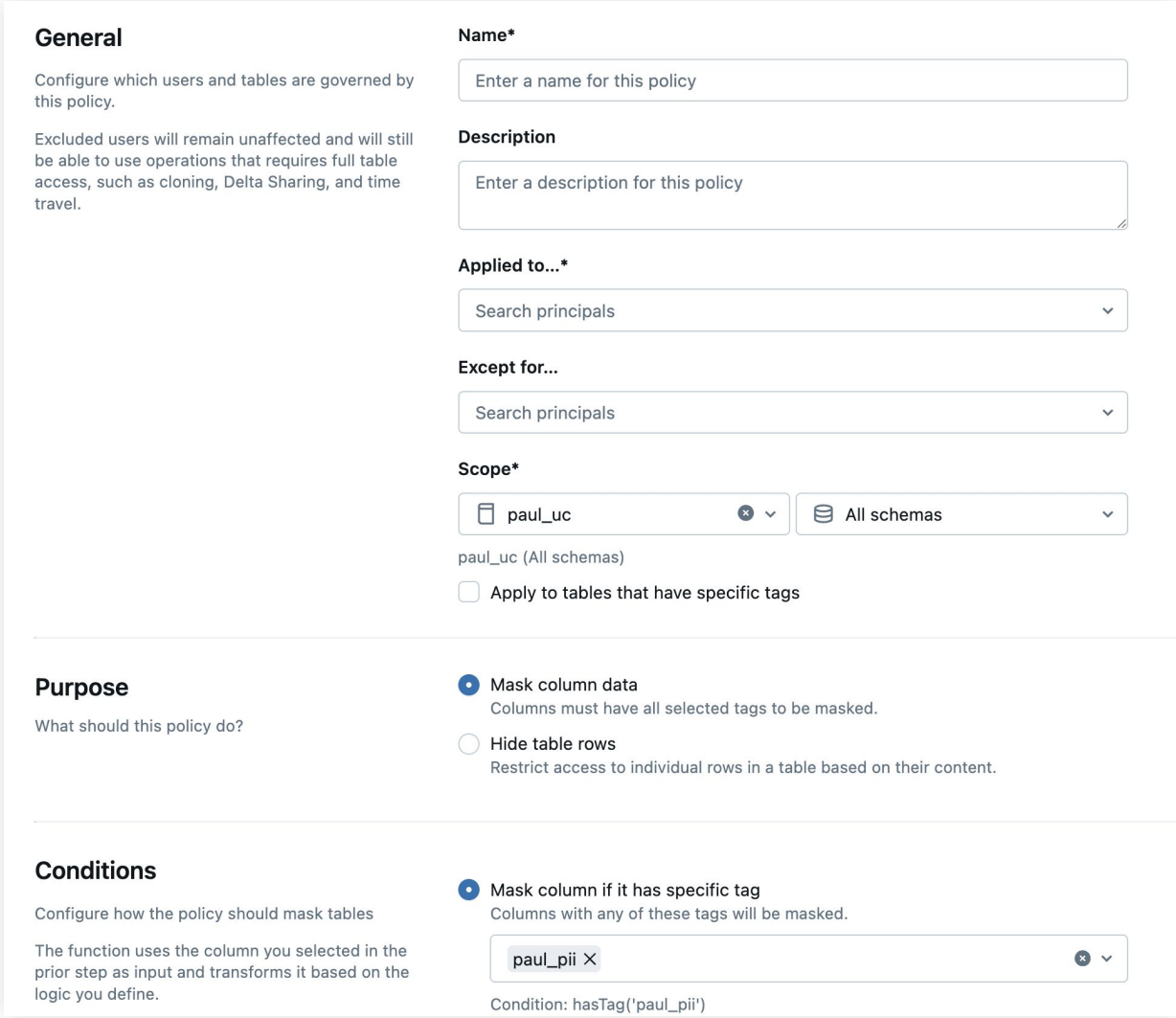
Automating Governed Tag application

- Automatically **detects PII** using an **agentic AI system** and applies as Governed Tags.
- Only scans new or changes tables and columns to ensure **efficient coverage at scale.**
- Provides **audit readiness** by logging all results to system tables



ABAC Policies

- **Policies** are rules that dictate users' access by looking at Governed Tags.
- Today, can dictate **column masking** or **row filtering**.
- Policies are attached to catalogs or schemas and apply downwards to **protect thousands of assets.**



Putting it all together

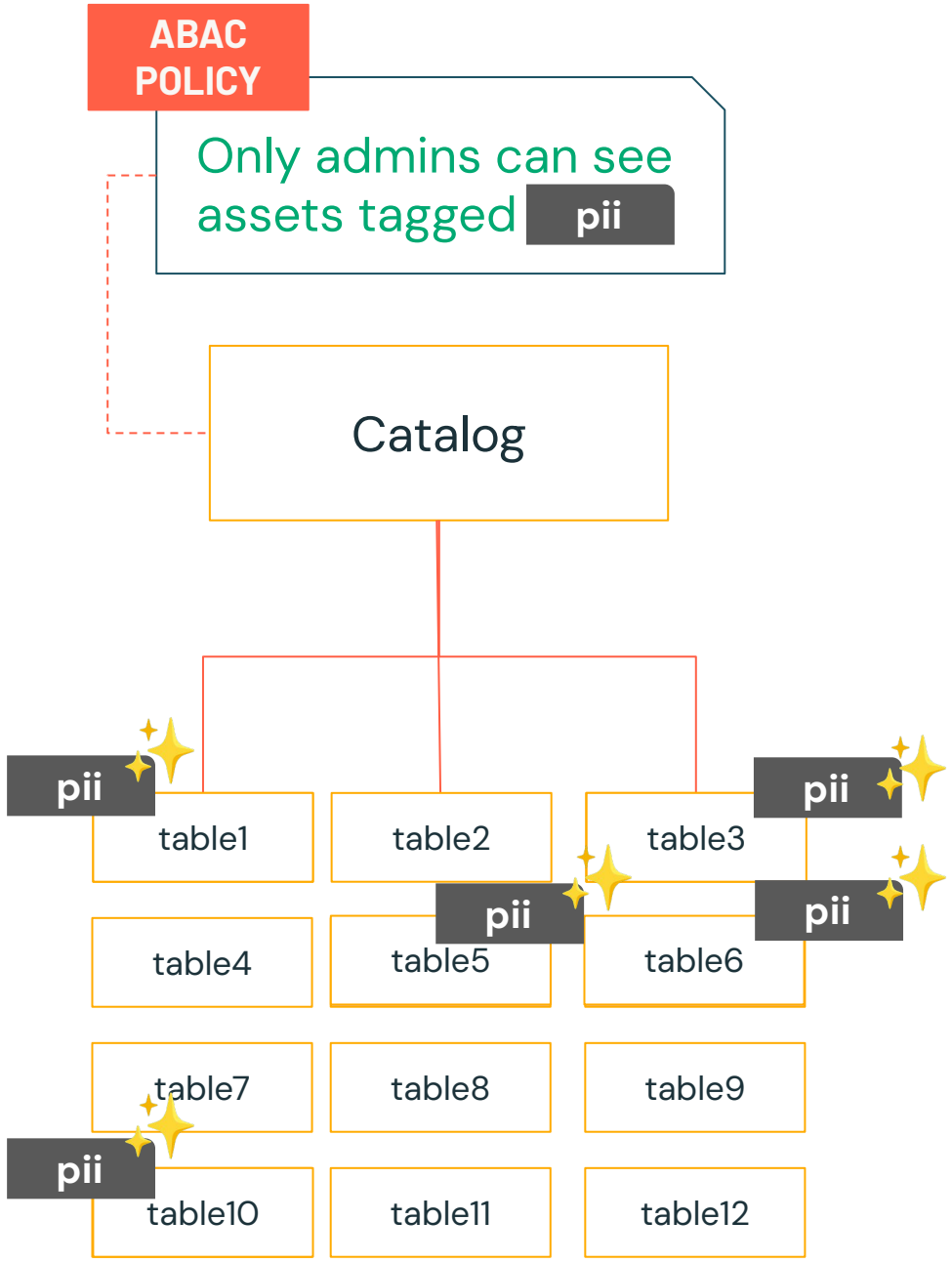
(Column Masking Use Case)

Use Case

Mask all columns containing ***PII***

With ABAC + Data Classification

1. **Data classification** automatically finds PII and applies Governed Tags
2. **ABAC** consistently applies governance policies
3. Clear **separation of responsibilities**



Governed Tags

Resources:

Public Docs

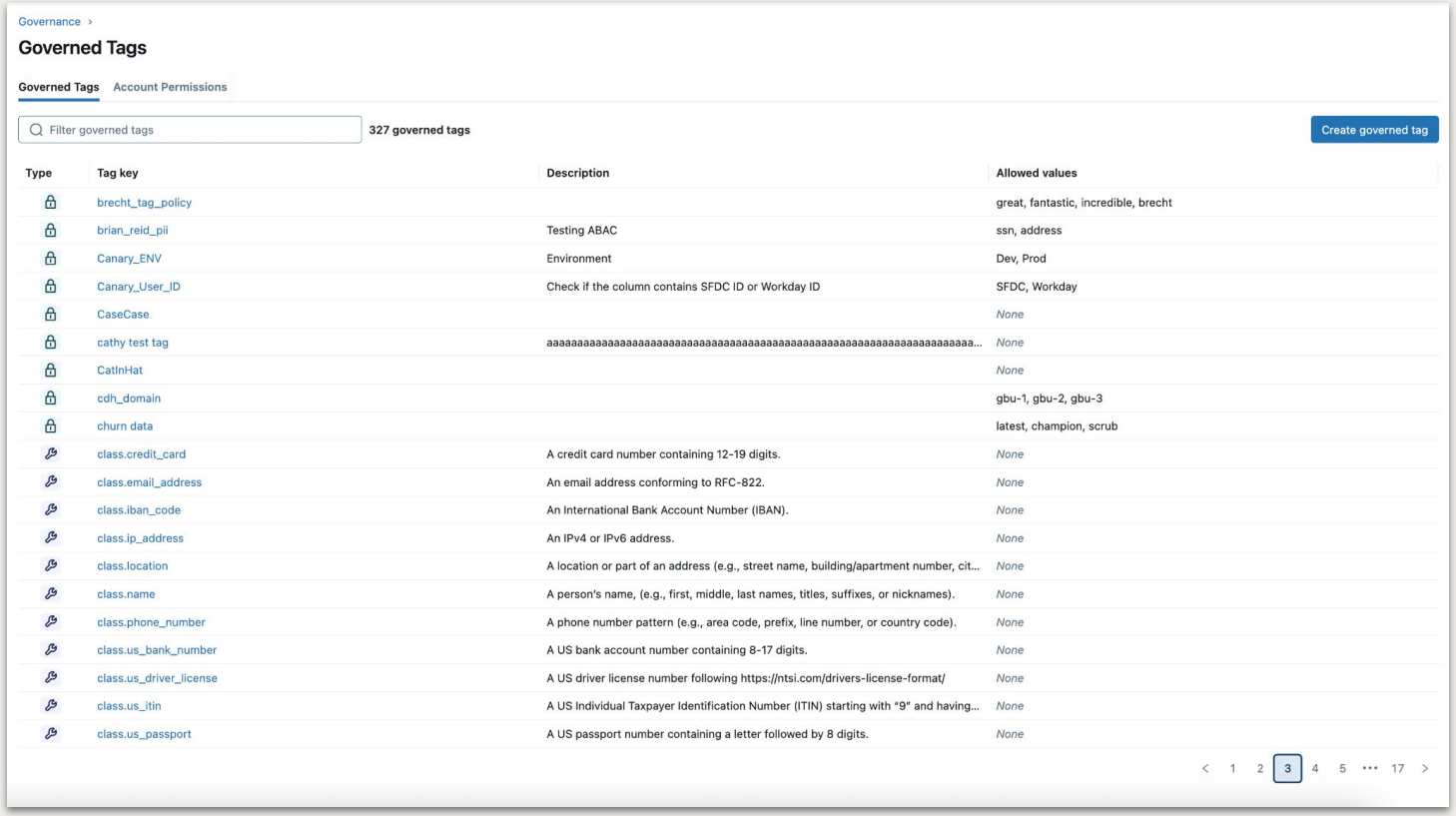
Governed Tags

Enforce consistent, secure tagging across data assets

Govern tags across your data estate
Define and enforce account wide user permissions on tags on UC objects and workspace assets (coming soon).

Standardize tagging for key use cases
Enable consistent tagging for cost attribution, compliance, and discovery

Support access control with trusted attributes
Use governed tags as inputs for ABAC policies and secure access



Core Use Cases

Governed Tags

Discovery

- Governed Tags make data easy to find with consistent labeling across teams
- No more missed assets due to inconsistent naming

Governance and Compliance

- Used by Data Classification to automatically identify/mark PII data
- Can be leveraged in ABAC policies to restrict access access to sensitive data
- Simplifies classification and auditing

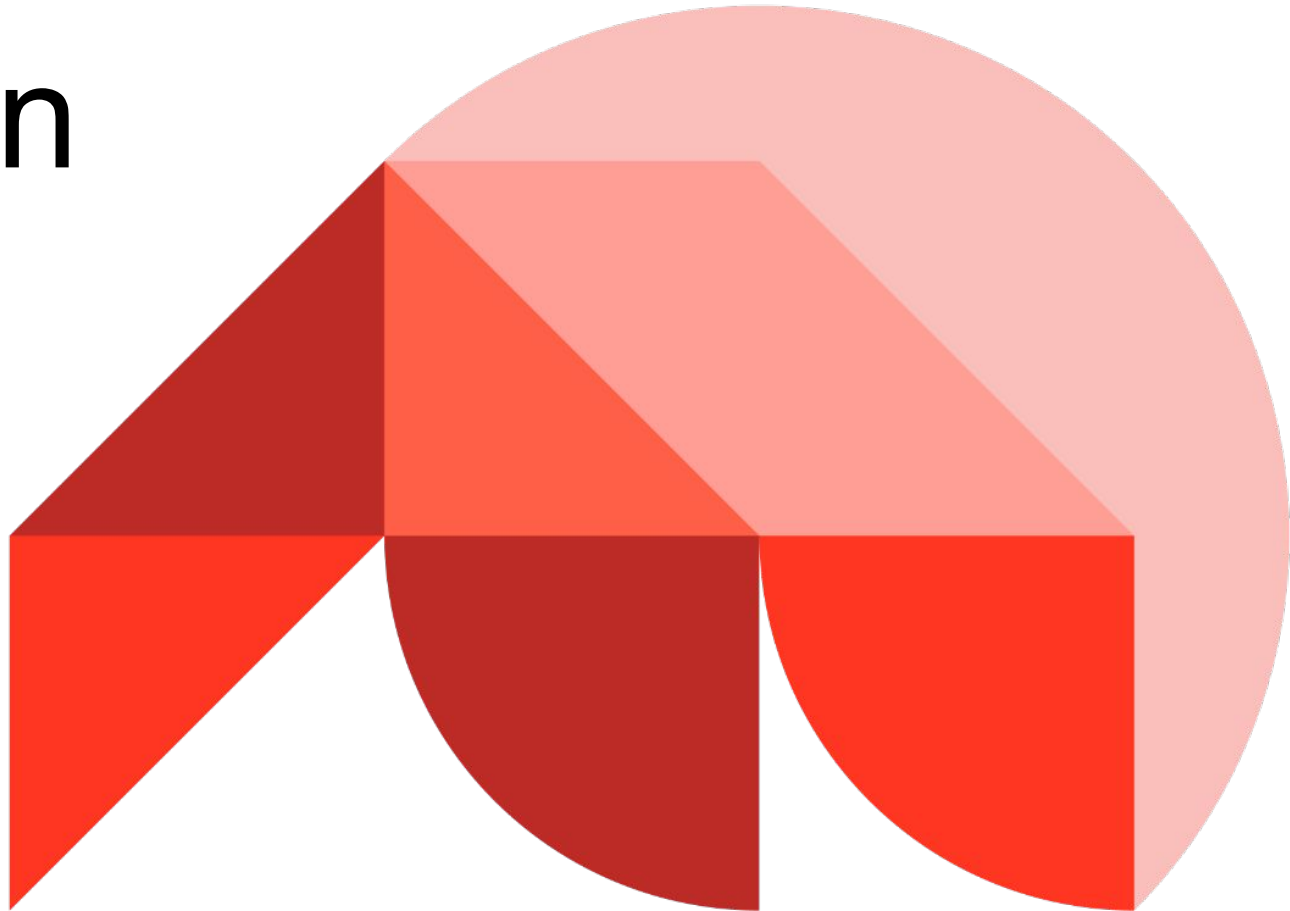
Cost Attribution

- Tags can be used to track usage and costs
- Governed tags will soon be integrated with Serverless Budget Policies
- Standardization provides consistency in labeling for better cost tracking

databricks

Resources:
Public Docs

Data Classification



Agentic Data Classification

From Governance Blind Spots to Complete PII Visibility at Scale

Agent-powered Detection

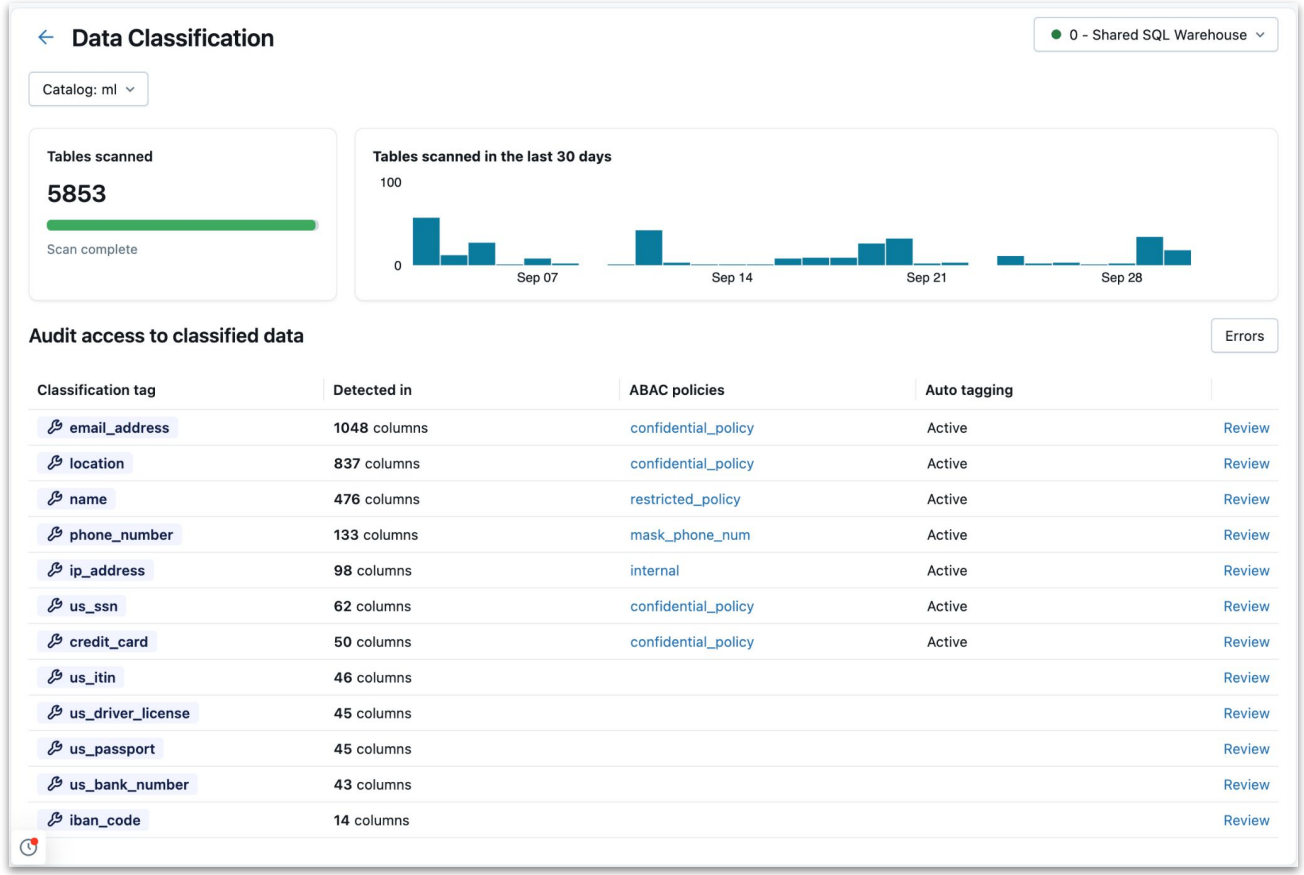
Automatically detects PII using an agentic AI system. Broader classifier support coming end of 2025

Efficient coverage at scale

Only scans new or changes tables and columns, covering the possibility of new PII.

Audit readiness

Logs all results to system tables: When PII was found, where it was detected, and sample data that triggered each classification.



Core Use Cases

Data Classification

🔎 Automated Discovery

- Automatically detect and tag sensitive data with Governed tags (“class.ssn”, “class.email_address”, etc.)
- ⭐ *Ensure any new table landing into UC is automatically scanned for PII*

📝 Actionable audits

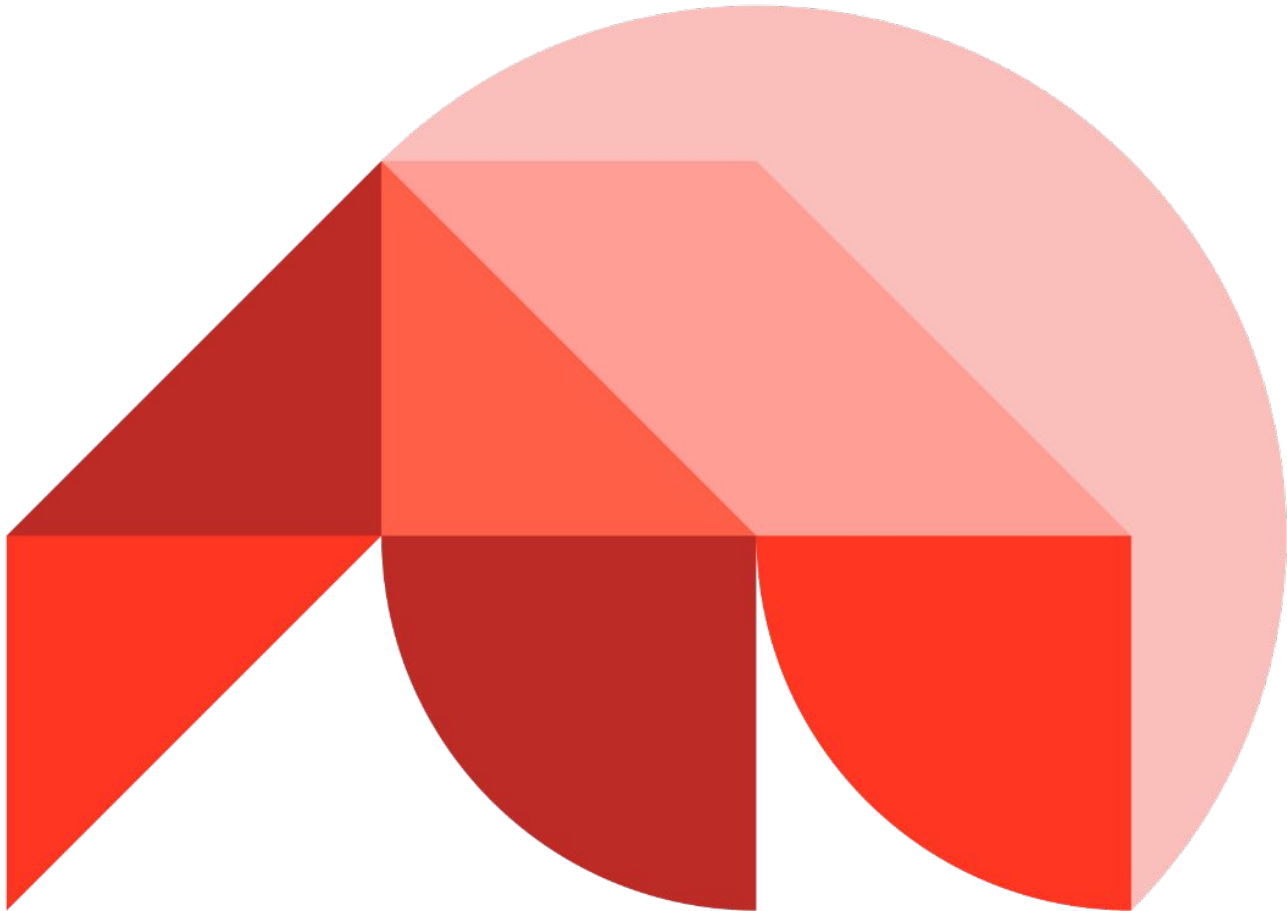
- **Logs all results to system tables:** what PII was found, where it was detected, and sample data that triggered each classification
- ⭐ *Understand who has access to sensitive data based on classification results*

🔒 Governance

- **ABAC policies:** Use Data Classification tags with ABAC to automatically restrict access to sensitive data
- **Inform sensitivity tiers:** Use Data Classification tags to automatically categorize tables for access control. (E.g. Tag a table as “confidential” if “ssn” is found)

databricks

Databricks Data Quality Monitoring



Data Quality Monitoring at Databricks

Simple, scalable, and consistent quality for the data platform

Simple setup

✅ Easy to set up, intelligent, built on your data

❌Highly manual threshold configuration

Coverage that scales

✅ Scales with platform growth. Reliable datasets for all!

❌Quality blind spots that undermine trust of the data platform

Platform-native

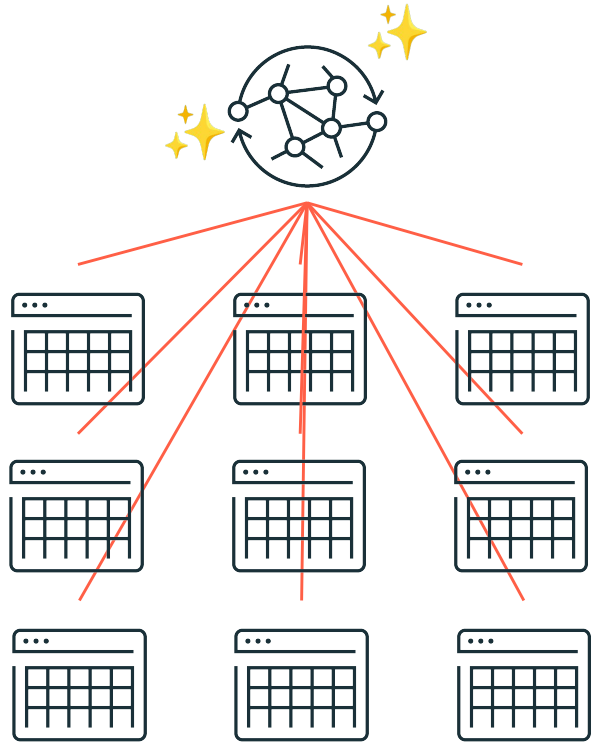
✅ Consistent quality signals across upstream pipelines and downstream consumers

❌Fragmented quality visibility across different vendors

Data Quality Monitoring Features

Anomaly Detection

- Enabled at the **schema level**
- ✨ Intelligently detects data quality anomalies for all tables
- Meant for: Scalable quality monitoring and actionable insights

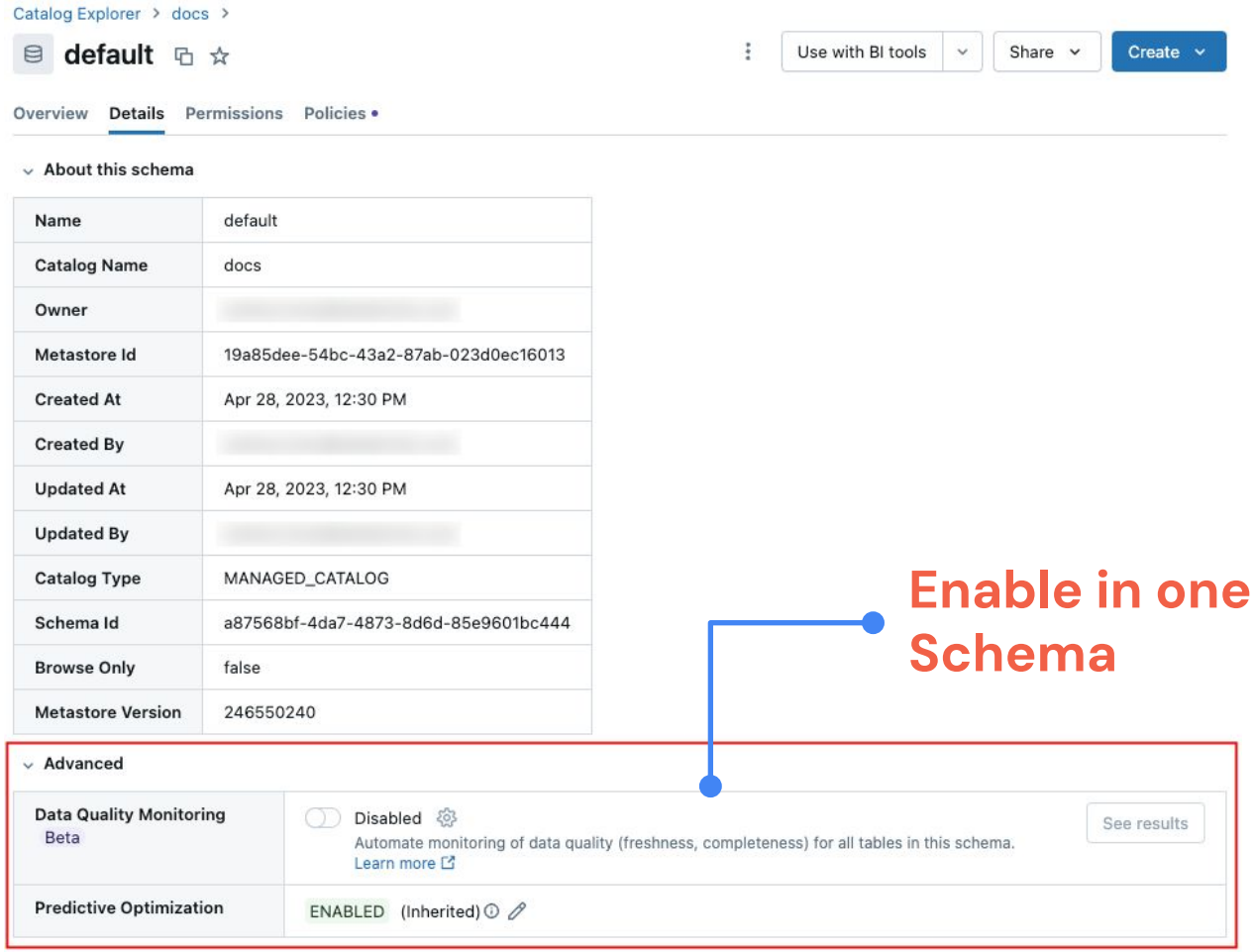


Data profiling

- Enabled at the **table level, f.k.a Lakehouse Monitoring**
- 📊 Provides a table profile for your most important tables
- Meant for: DIY table monitoring and quick summary statistics

Simple setup

One-click, quality scales with your data platform



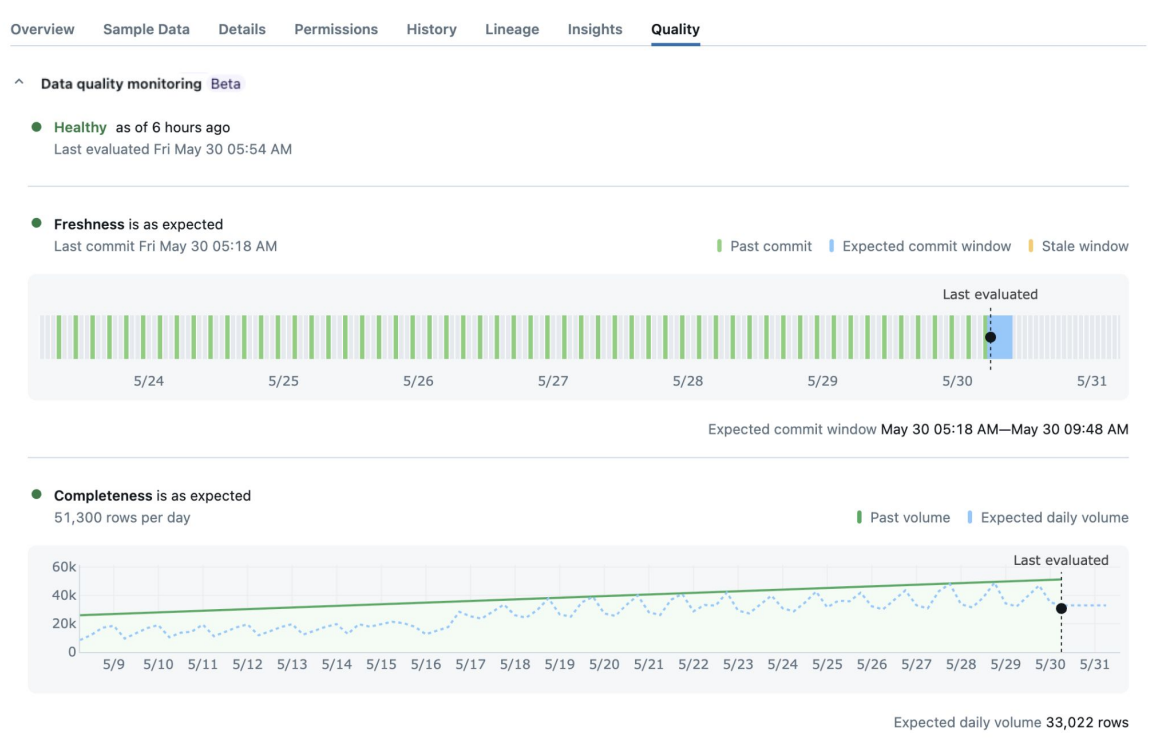
Enable in one-click for your entire Schema

- **Freshness**
- **Completeness**
- **Segmentation**
- **Custom rules**

AI-Powered

Intelligent and built on your data

- **Freshness:** tracks how recently data was updated, flagging tables as stale if updates are delayed.
- **Completeness:** checks if row volume in the last 24 hours falls below expected levels.
- **Segmentation:** View quality metrics by slice (e.g., freshness of vendor_id="walmart")



~~*Bespoke rules per table*~~

Simple and scalable to all tables

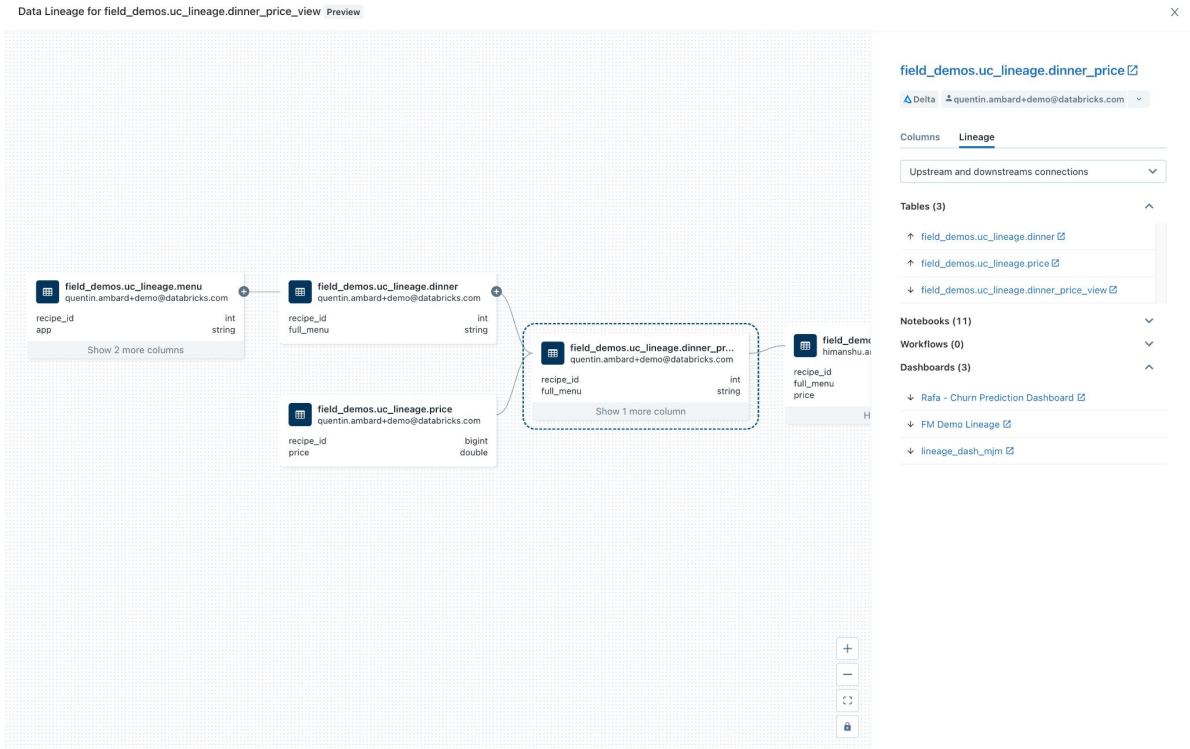
databricks

Lineage

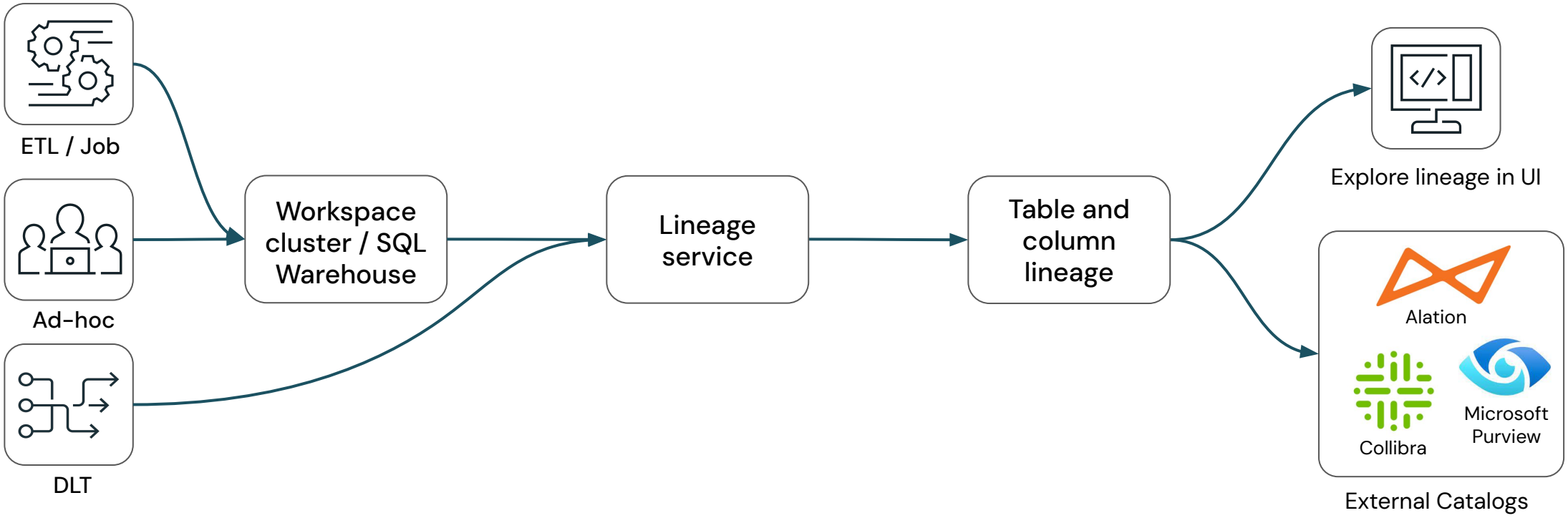
Automated lineage for all workloads

End-to-end visibility into how data flows and consumed in your organization

- Auto-capture runtime data lineage on a Databricks cluster or SQL warehouse
- Track lineage down to the table and column level
- Leverage common permission model from Unity Catalog
- Lineage across tables, dashboards, workflows, notebooks, feature tables, files, and pipelines



Lineage flow - How it works



- Code (any language) is submitted to a cluster or SQL warehouse or DLT* executes data flow

- Lineage service analyzes logs emitted from the cluster, and pulls metadata from DLT
- Assembles column and table level lineage

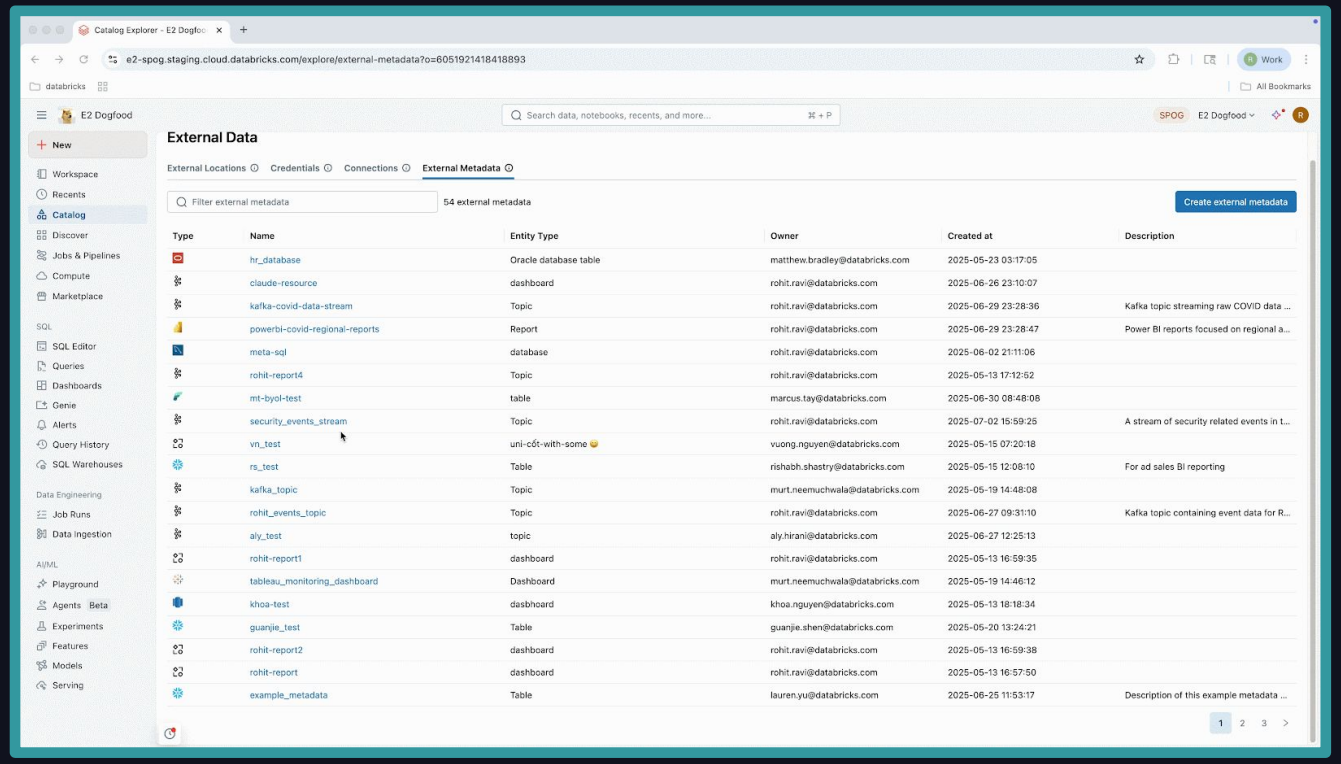
- Presented to the end user graphically in Databricks
- Lineage can be exported via API and imported into other tool

Custom Lineage

A BYO party? Count us in.

Augments auto-captured lineage with manually configurable relationships

Introduces a new external metadata object that can be tied to tables, views, models, apps, etc.



databricks

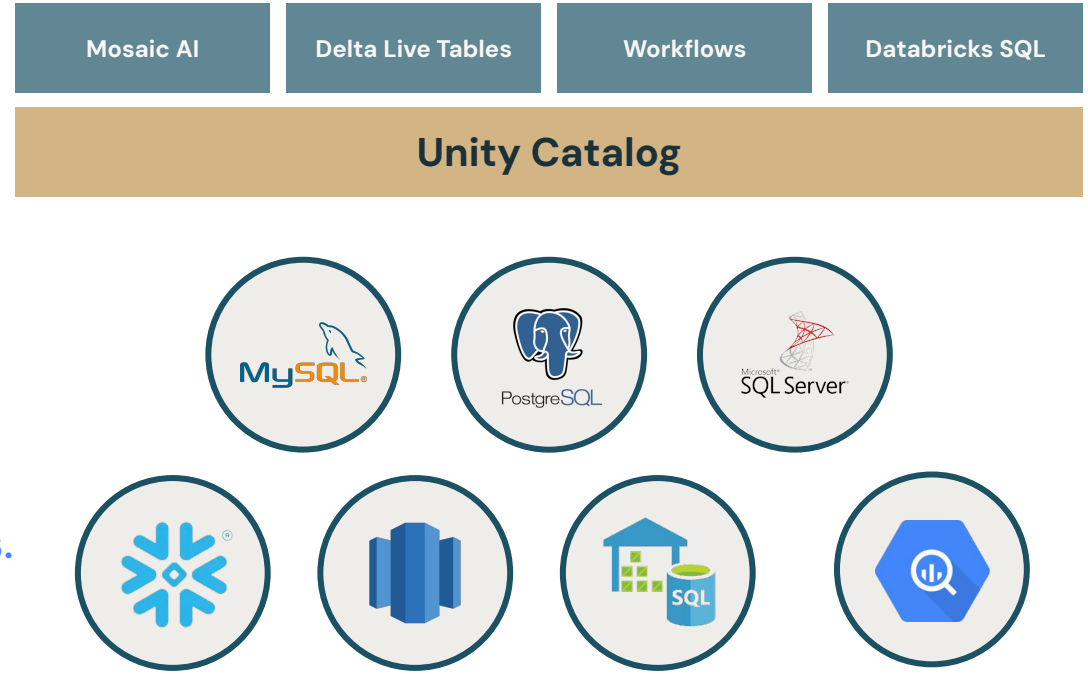
Lakehouse Federation

Discover, query, and govern all your data - no matter where it lives

Lakehouse Federation: Databases & Warehouses

Unify your data estate with the Lakehouse

Discover, query, and govern all your data – in any system



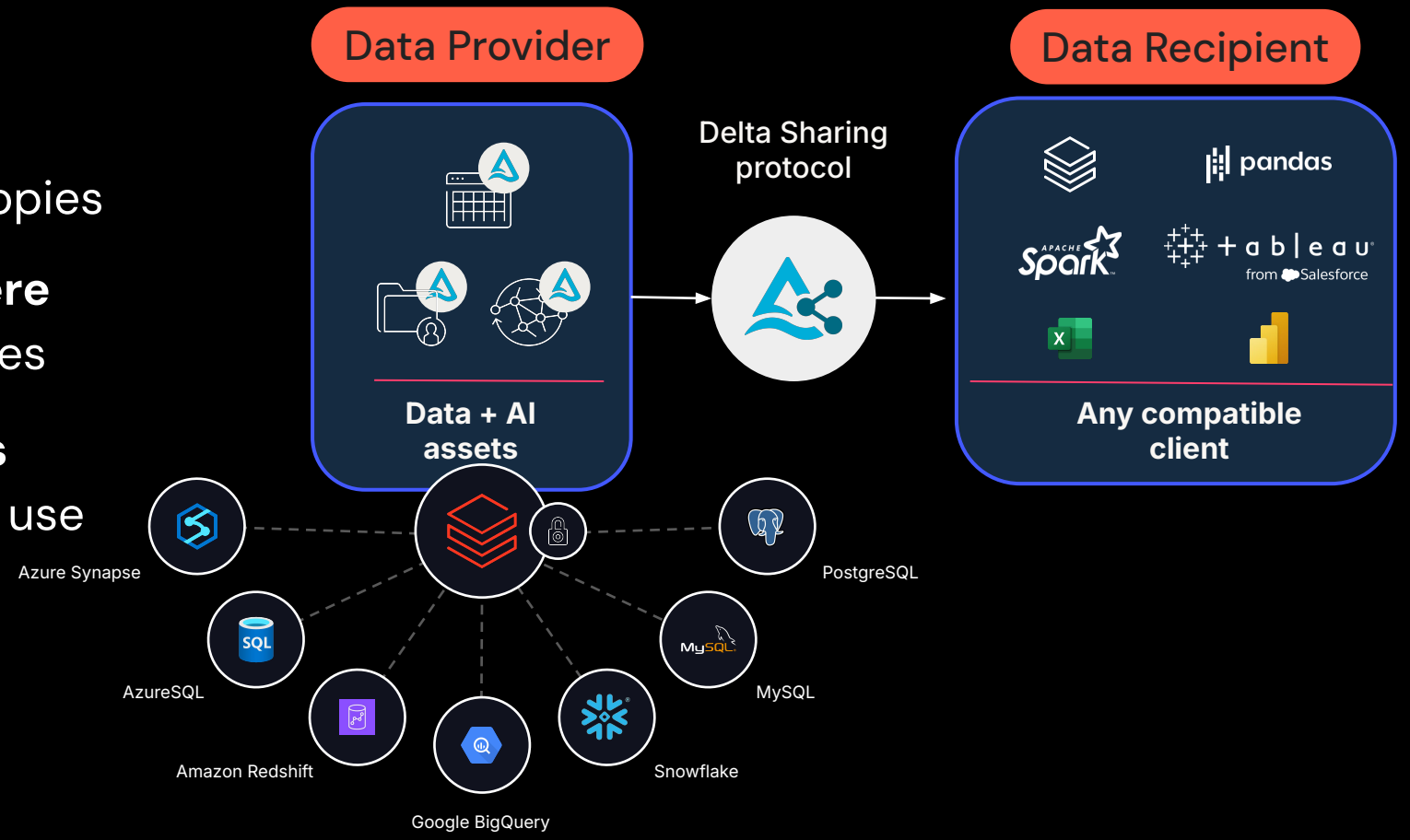
Database & DW - General Availability

- Improved pushdown coverage & performance for Snowflake, BigQuery, SQL Server, Postgres, Redshift & Synapse.
- OAuth support for Snowflake connections.
- Azure AD support for Azure ecosystem connections.
- Case sensitive namespace support
- Salesforce Data Cloud Connector (Preview)

Delta Sharing

The only open protocol for cross-cloud, cross-platform data sharing

- Share **all asset types:** data sets, notebooks and AI models
- Share data **across clouds and regions without replication** or copies
- Share live data **directly from where it's stored** – incl external databases
- Share **openly to all your partners** regardless of which platform they use



Share Data & AI assets

Unlock use cases, from data exploration to model deployment

- Share tables, dynamic views, materialized views, streaming tables
- Share Delta or Iceberg tables
- Share AI models, unstructured data (audio/video/files) and notebooks
- Enable data engineers, data scientists, and business analysts to work on one platform

Sharing for Lakehouse Federation

Share data directly where it is stored across platforms, without replication

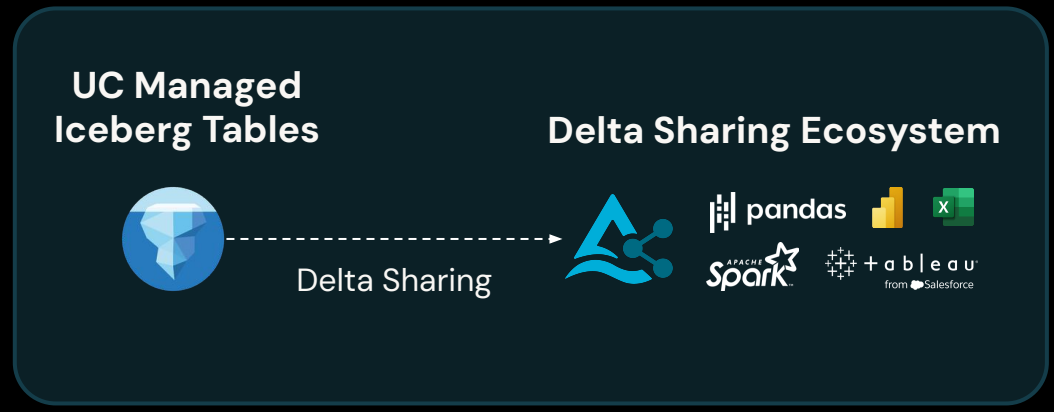
Support for many data sources: Snowflake, Amazon Redshift, Azure Synapse, Google BigQuery, MySQL, PostgreSQL, Microsoft SQL Server

Use the same easy, secure sharing experience from Databricks and Delta Sharing. Share external cross-platform data sources without ETL

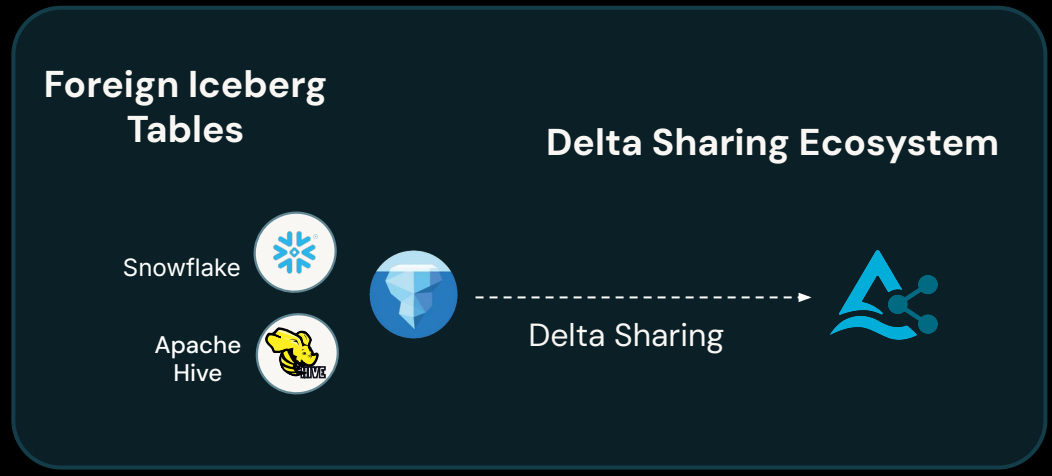
Iceberg Interoperability - Share Iceberg Tables

Unify the formats in sharing and collaboration for an open ecosystem

Share managed Iceberg tables written with external Iceberg clients, all managed by Unity Catalog. Access Iceberg tables using the Iceberg REST Catalog API



Share foreign Iceberg tables via Lakehouse Federation Sharing through direct access with D2D cloud tokens

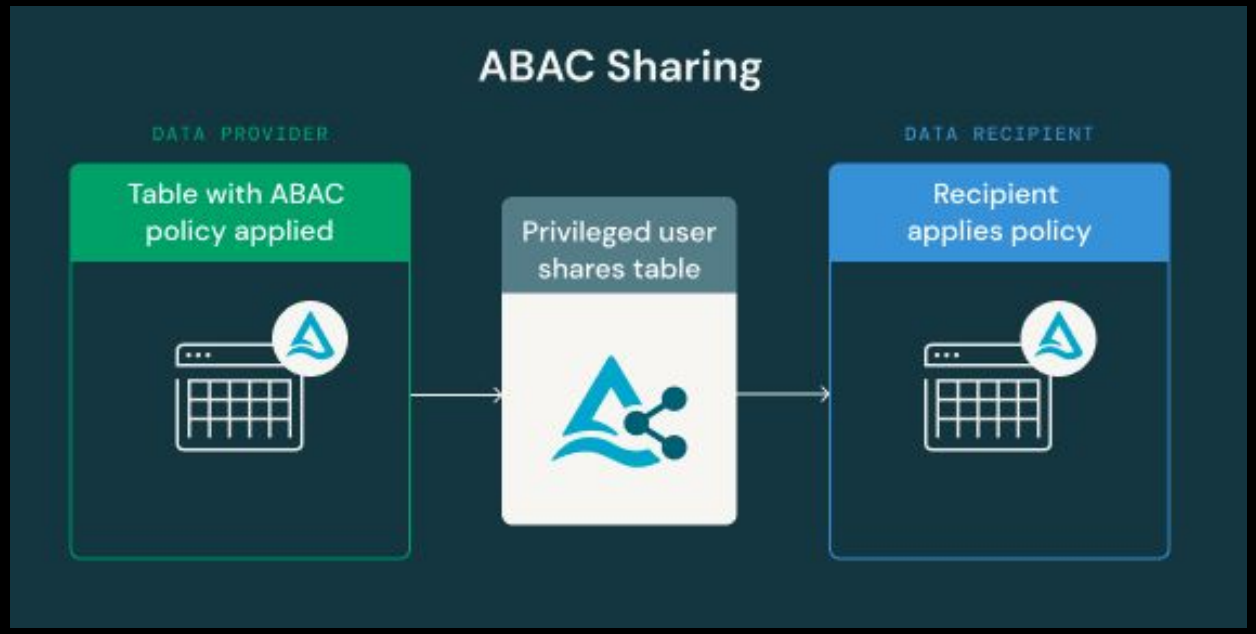


Sharing with Attribute Based Access Control

Govern shared data with fine-grained access control

Privileged users can share tables, views, MVs and STs with ABAC policies without needing to make copies

Recipients can apply the same policy or different policies on the shared assets, as well as RLS/CM



Data & AI Governance Operational Model

Governance Operational Model

There is no one-size-fits-all for data and AI governance

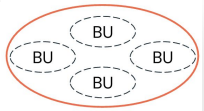
The organizational structure for implementing governance. There are four primary variants:

- **Centralized**
  - Single governing body
  - Highly regulated industries
- **Decentralized**
  - Fully distributed with minimal central oversight
  - Agile organizations prioritizing innovation and rapid iteration
- **Federated**
  - Distributed ownership with strong central oversight
  - Multinational corporations
- **Hybrid**
  - Mix of centralized and decentralized
  - Post-merger organizations

Operational Governance Model (1/2)

Governance models can adopt different structures

Centralized Governance

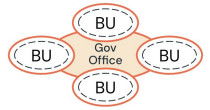


A central governing body or team is responsible for setting and enforcing data and AI policies, standards, and best practices, ensuring uniformity and consistency organization-wide.

Suits organizations operating in highly regulated environments or those prioritizing consistency.

The centralization of decision-making may limit flexibility, adaptability, and responsiveness to local needs.

Federated Governance

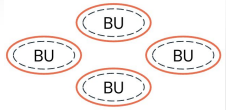


Combines centralized oversight with decentralized execution. Business units govern their data, while a central body ensures alignment and consistency.

Suited for multinational corporations or conglomerates with diverse business units.

Strengths: Balances autonomy with collaboration, leveraging localized expertise

Decentralized Governance

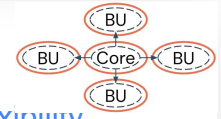


Allows individual departments to independently manage and adapt data and AI practices according to their operational needs while ideally aligning with broader organizational goals.

Provides flexibility and enables departments to apply specialized knowledge, tailoring data and AI solutions to unique contexts.

The decentralized structure can lead to duplication of skills, efforts and technology, and variability in risk management, ethical standards, and regulatory compliance across units.

Hybrid Governance



Combines centralized and decentralized governance elements and balances uniform standards with local flexibility.

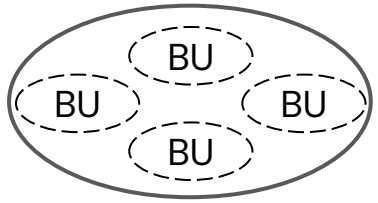
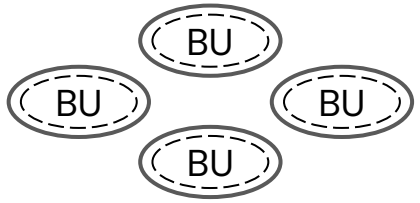
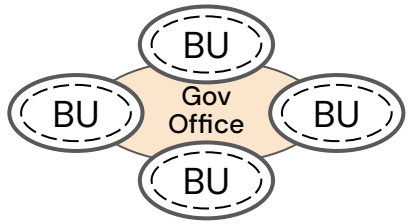
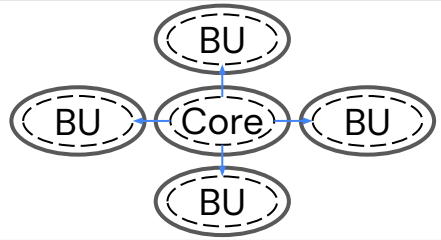
Typically managed by a Center of Excellence, hybrid governance provides centralized oversight while allowing departments to leverage localized knowledge and innovation.

While effective for organizations with complex data and AI programs or those undergoing transformations, hybrid governance requires careful coordination to avoid overlapping roles or inefficient resource use.

Governance Boundary Business Unit (BU) Boundary

Governance Operational Models (2/2)

Different types of Operational Models and some of their characteristics

	Centralized Governance	Decentralized Governance	Federated Governance	Hybrid Governance
Gov Boundary Business Unit (BU) Boundary				
Definition	Single authority controls data & AI policies	Each BU manages its own data & AI policies (follows Conway's Laws)	Central coordination with local autonomy	Explicit split (core vs. BU data & AI)
Decision-Making	Centralized, top-down	Distributed, independent	Federated - central guidelines with local execution	Tiered (central dictates core rules, BUs their own rules)
Data Ownership	Owned by a central team (central governance)	Owned by individual teams or BU (shared nothing)	Distributed ownership across entities (shared governance)	Core data owned centrally; BU data owned decentralized (shared core data)
Scalability	Limited by central team's capacity	Highly scalable, but may lack coordination	Scalable with structured coordination	Moderate (core data centralized; BUs scale independently)
Compliance & Security	Strong enforcement, high consistency	Varies by team, potential security inconsistency	Central standards with local enforcement	Strict compliance for core data; flexible for domains
Operational Efficiency	Efficient but bureaucratic	Agile but may duplicate efforts and introduce silos	Balanced - efficient yet flexible, but high federation might introduce silos	Potential bottlenecks for core data; efficient elsewhere
Use Case Suitability	Highly regulated industries (finance, healthcare)	Agile businesses, startups, innovation-driven sectors	Large enterprises, multinational corporations	Mixed compliance/innovation needs (e.g., fintech)

Centralised Governance

(Multi-Tenant Lakehouse)

Each BU/domain operates in a different business boundary, however all BUs/domains fall under the same operational boundary that governs all data & AI assets (multi-tenancy)

Pros:

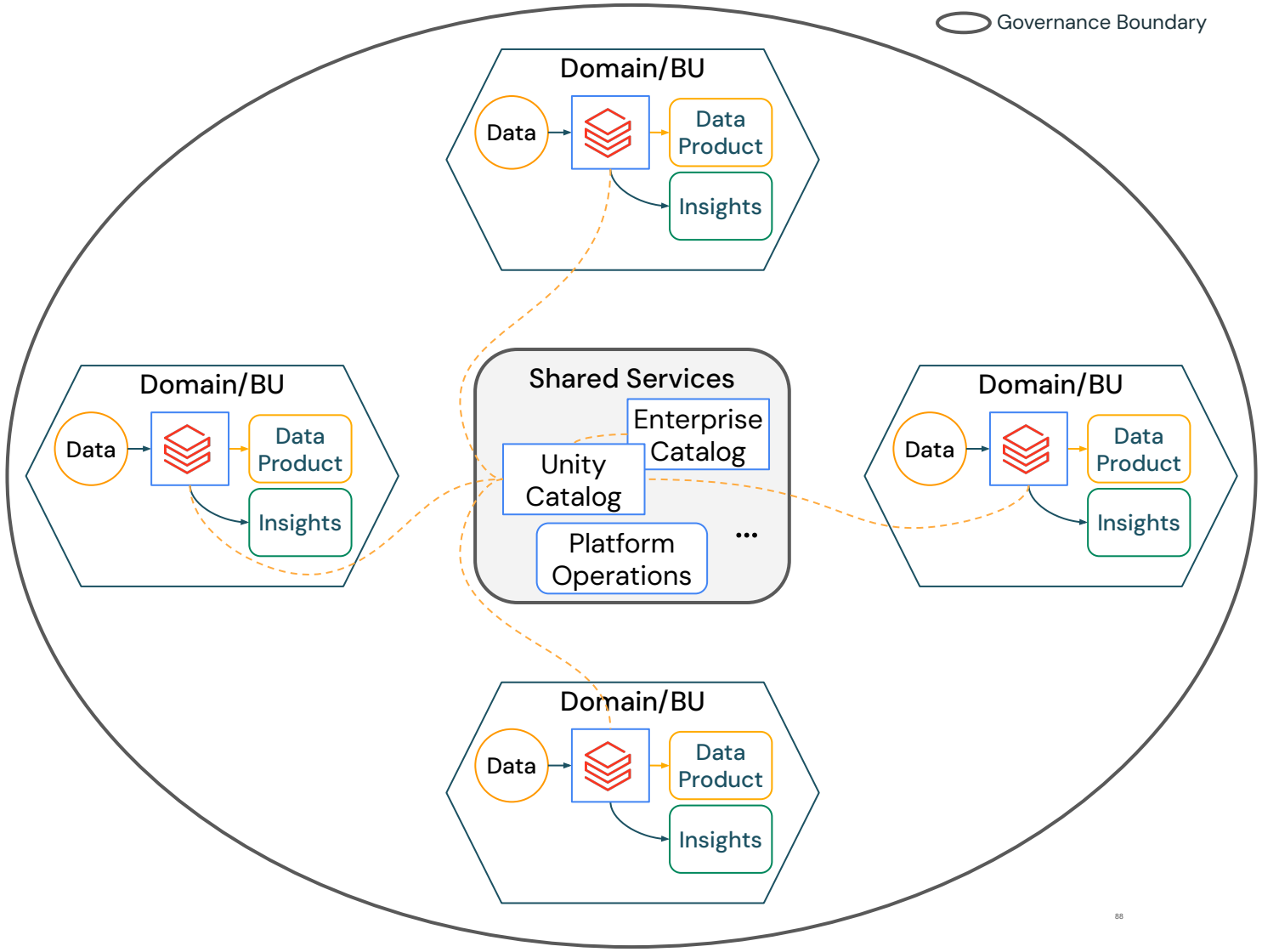
- Strong governance across all BUs/domains
- Unified view of all data and AI assets across the group

Cons

- Might be slow for innovations and decisions making
- Changes are not easy as they have to be agreed across all BUs/domains

Use When:

- Corporate business model is pretty much the same for all BUs/domains
- Operating model is leaning towards centralization (with a centralised IT)
- Strong inter-BU/domain data sharing requirements



Decentralised Governance

(Single Tenant Lakehouses)

Each BU/domain has an independent business and operational boundary (i.e own tenant). Governance is delegated to each BU/domain to define and enforce their own policies

Pros:

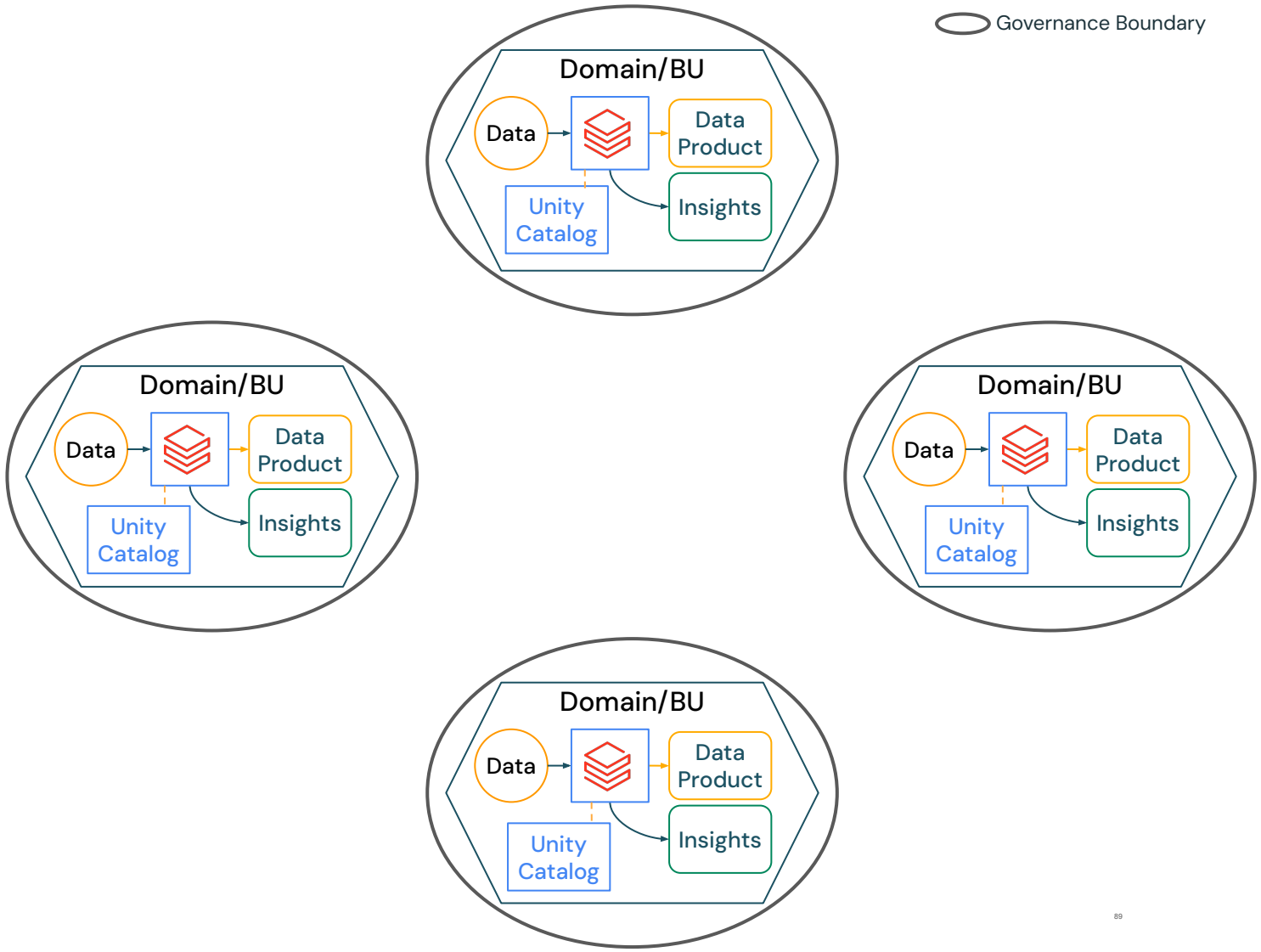
- Greater autonomy for each BU/domain to decide their own data & AI management and strategic goals
- Speed up of innovations and decision making

Cons

- Inconsistent governance perspective
- No strict rules on how each BU/domain defines or maintains their data & AI assets

Use When:

- BUs/domains have different business models independent from one another
- Minimal to none sharing or exchanging information between BUs/domains expected



Federated Governance

(Harmonised Lakehouse Mesh)

Each BU/domain has an independent business and operational boundary. However, governance policies are defined by a centralised Governance-Office, enforced by each BUs - critical assets are governed by the governance office (e.g. shared data products)

Pros:

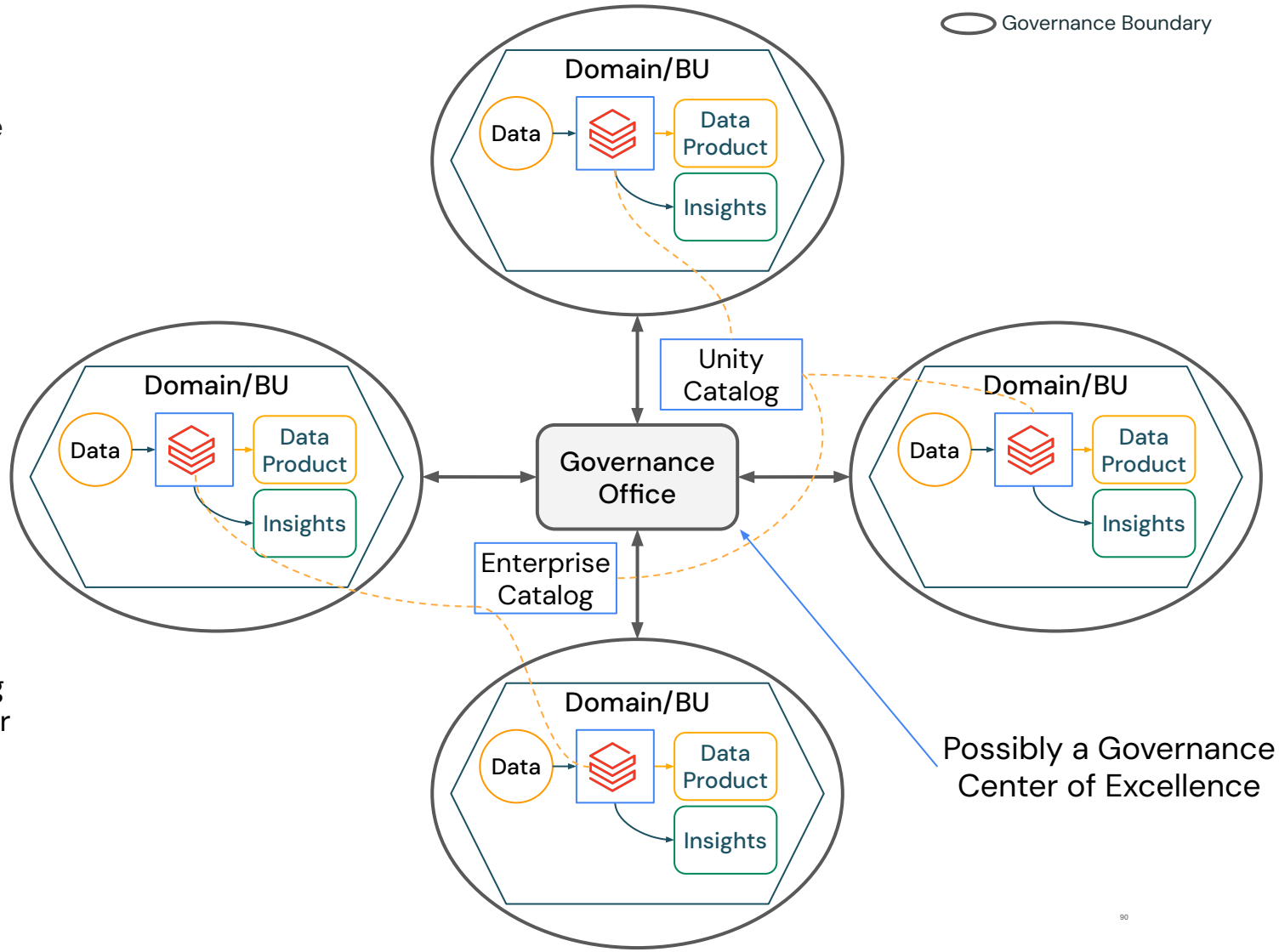
- Speed up of innovation and decision making, yet still maintain centralised governance guidelines
- Possible use of Center of Excellence

Cons

- Hard to enforce governance guidelines across all BUs/domains, without strong governance operations/teams (Governance office)

Use When:

- The company operates each BU/domain independently, however, still requires data sharing between BUs/domains to deliver business goals or comply with regulatory requirements



Hybrid Governance

(Hub & Spoke Lakehouse Mesh)

Midway between centralised and decentralised, allows each BU/domain to govern their own data (decentralised). However, critical business data is collected and governed in one place for all to use (centralised)

Pros:

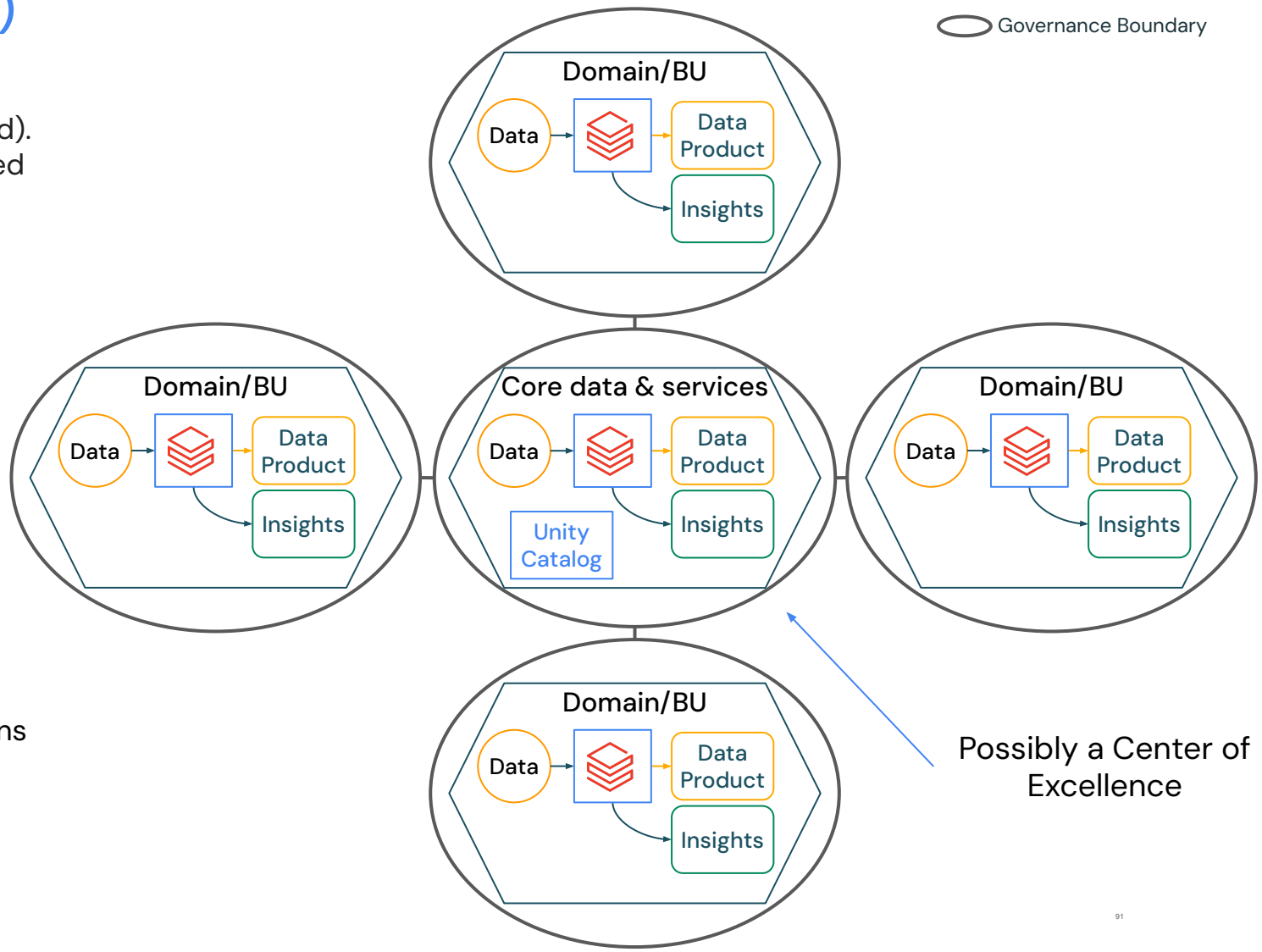
- Speed up of innovations and decision making at the BU/domain level
- Strong governance for critical data that are centrally managed (eg Group Financial details, customer PII data, etc)
- Possible use of Center of Excellence

Cons

- Core data & services will almost act as an independent BU/domain, requires own management, operations and funding

Use When:

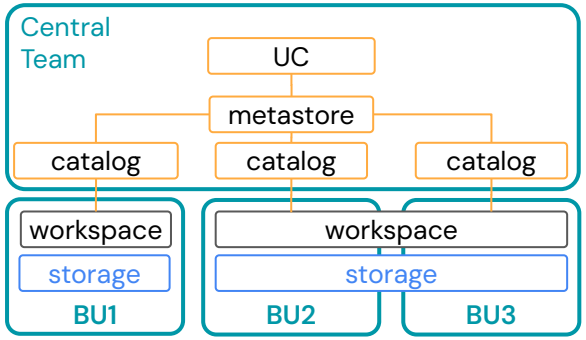
- The company doesn't want to restrict BUs/domains from pursuing own innovation, however, critical business data such as PII/PCI must be centrally managed to comply with regulations



Operational Model impacts UC and Databricks setup

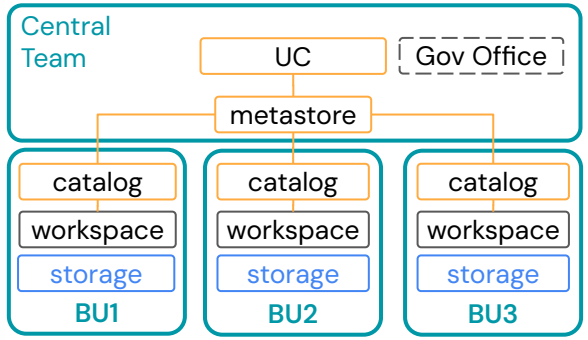
Single Cloud and Single Region

Centralized Governance



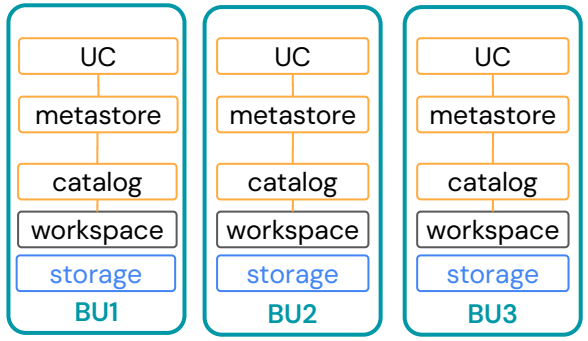
- Central team owns UC down to the catalog level (and sometimes schema)
- It dictates structure and governs data and AI
- BUs own their workspaces and storage (can be shared between BUs)
- BUs can share workspaces and storage accounts

Federated Governance



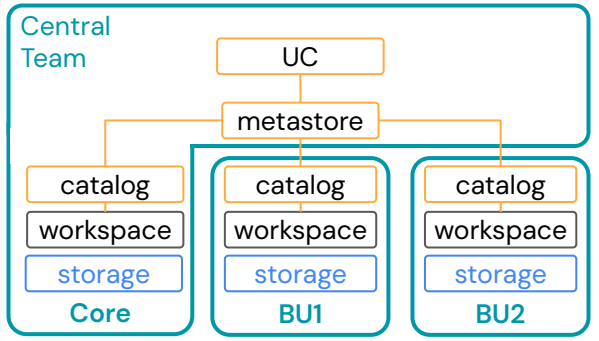
- Central team owns UC down to metastore level (Governance oversight via governance team and metastore admin)
- Central teams has no data
- Independent BUs owning their catalogs

Decentralized Governance



- BUs use different cloud tenants and hence have separate Unity Catalog instances and assigned metastore
- They have their own workspaces and storage accounts

Hybrid Governance



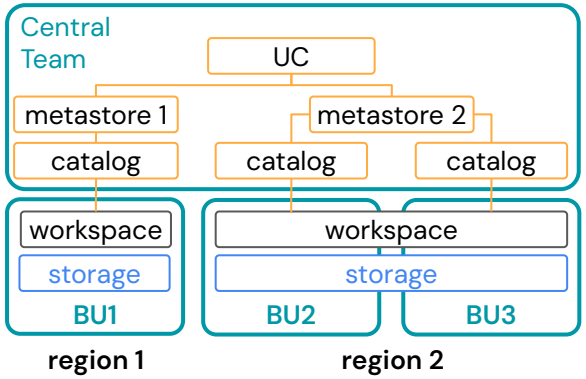
- Central team owns UC, and can own metastore (technical ownership, no metastore admin)
- Central teams owns core (common) data
- BUs have their own catalogs and can be completely independent (own workspaces and storage accounts)
- BUs use core data

BU=Business Unit, UC=Unity Catalog

Operational Model impacts UC and Databricks setup

Single Cloud and Multi-region

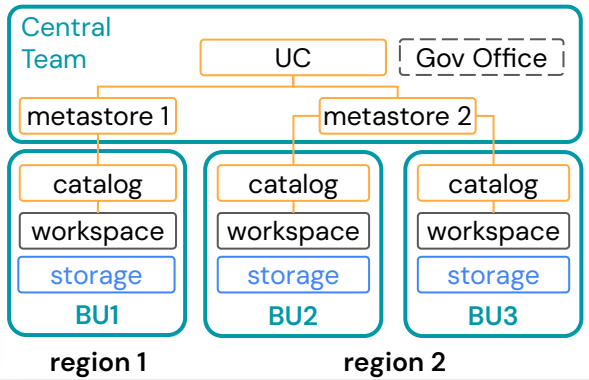
Centralized Governance



Same as single region, plus

- Central team owns UC down to the catalog level across all regions/metastores

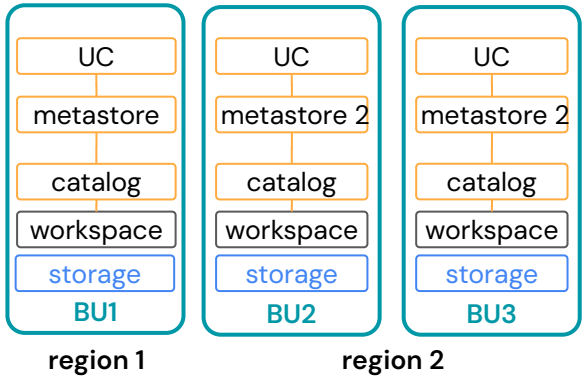
Federated Governance



Same as single region, plus

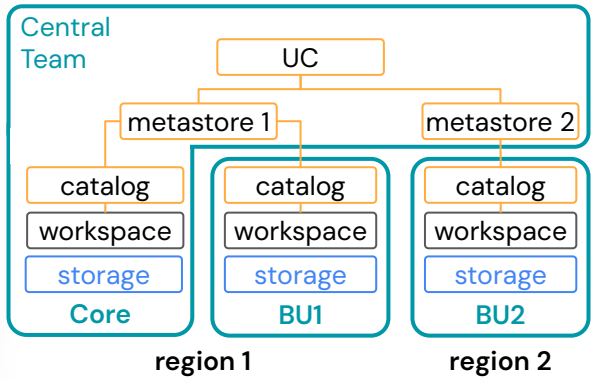
- Central team owns UC down to the level of each metastore (oversight via metastore admin)

Decentralized Governance



Same as single region

Hybrid Governance



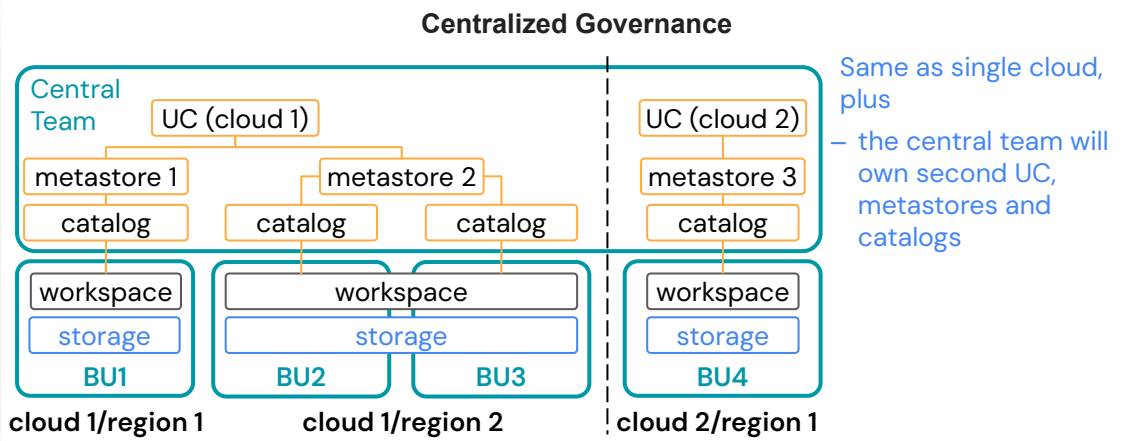
Same as single region, plus

- Central team owns UC, and can own one or more metastores of the regions (technical ownership, no metastore admin)

BU=Business Unit, UC=Unity Catalog

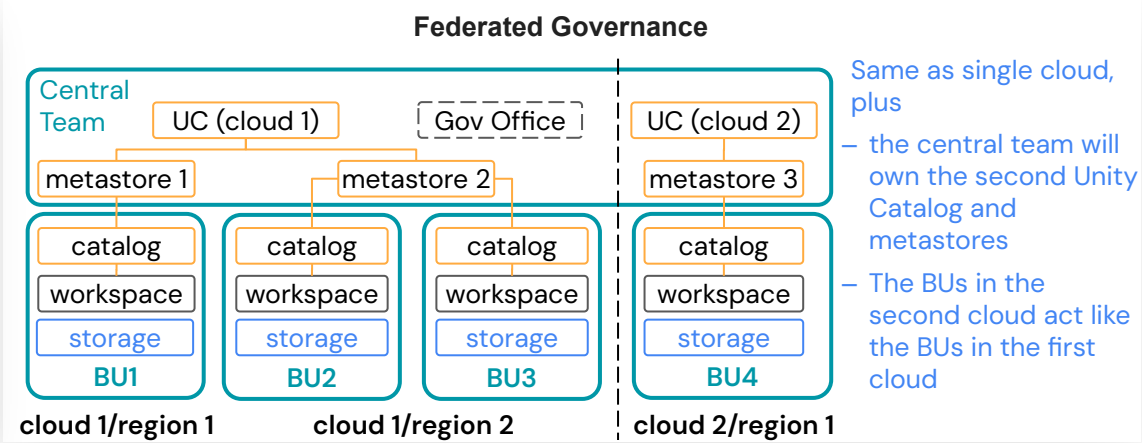
Operational Model impacts UC and Databricks setup

Multi-cloud and Multi-region



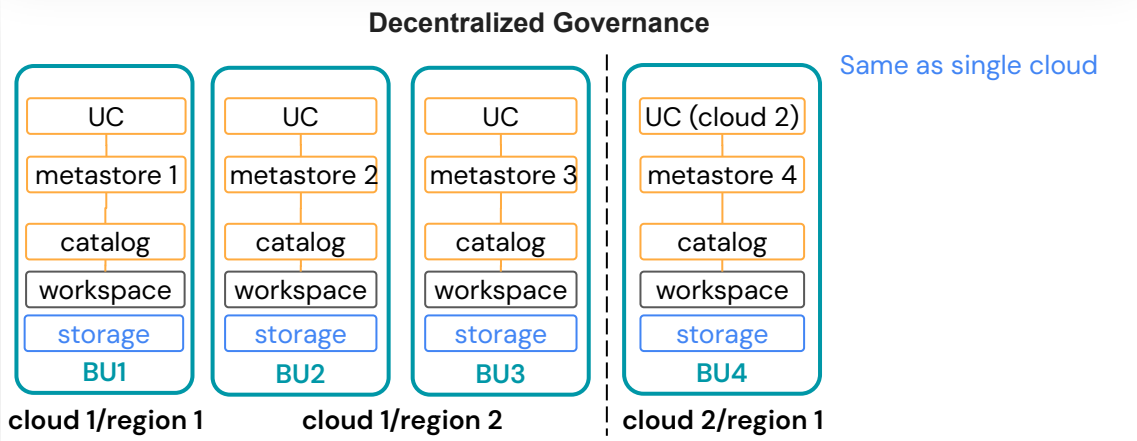
Same as single cloud, plus

- the central team will own second UC, metastores and catalogs

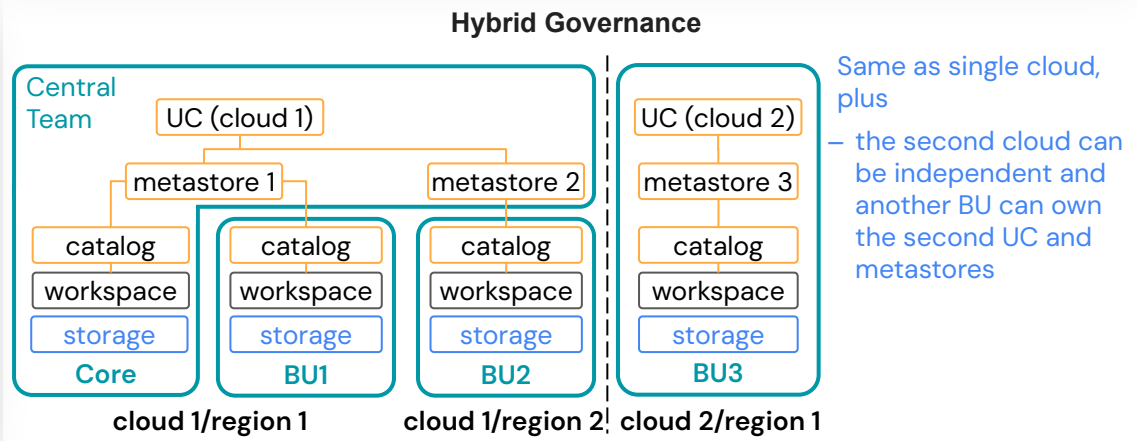


Same as single cloud, plus

- the central team will own the second Unity Catalog and metastores
- The BUs in the second cloud act like the BUs in the first cloud



Same as single cloud



Same as single cloud, plus

- the second cloud can be independent and another BU can own the second UC and metastores

BU=Business Unit, UC=Unity Catalog