Unity Catalog Upgrade

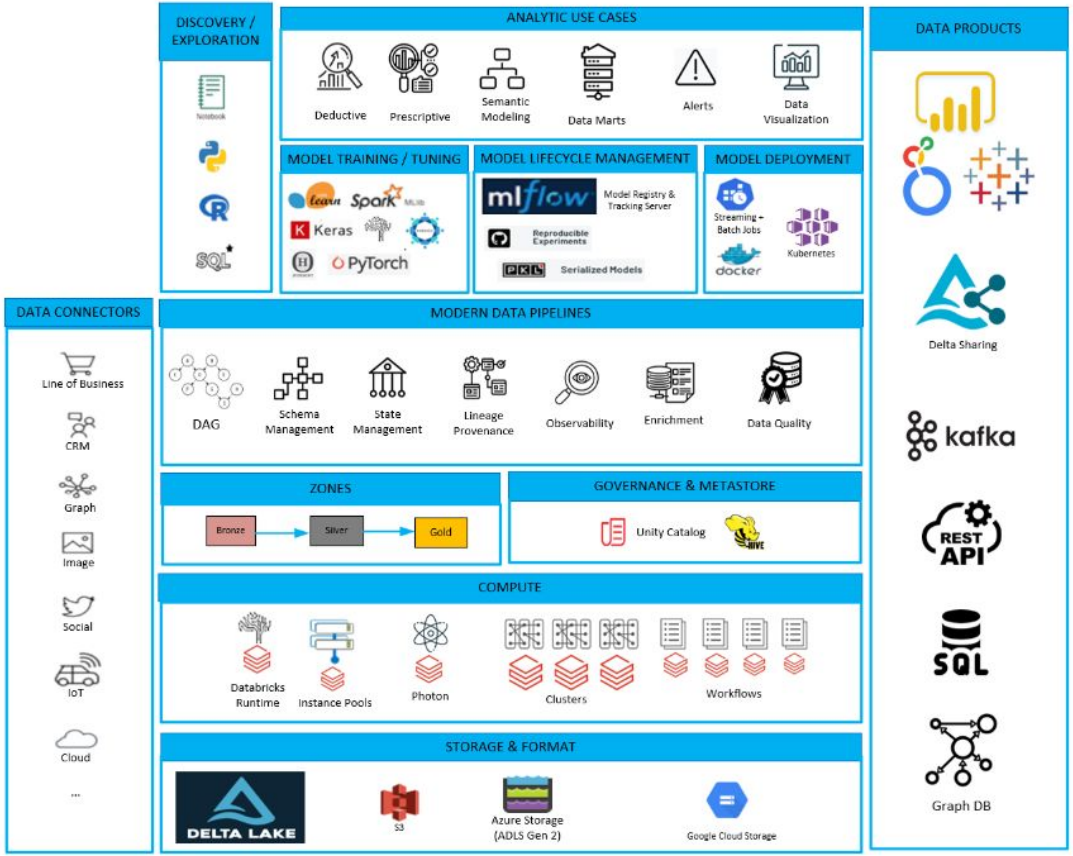
Goals and Objectives

PROJECT OBJECTIVES

The purpose of this discussion is to understand the following:

1. What are you hoping to get out of this engagement?
2. Why do you want to upgrade to Unity Catalog (what is your major motivation behind this upgrade)? Are there specific features you care about more than others (that we should prioritize)?
3. How do you want to implement Unity Catalog?
    a. Via Source Control and full CI/CD
    b. Or are you OK with some Manual Configuration
4. What deliverables do you most care about?
    a. Assessment
    b. Roadmap (Design Document)
    c. Project Plan
    d. Implementation
        i. Account Level Configuration
        ii. Workspace to Account Security migration
        iii. Jobs and Tables Upgraded

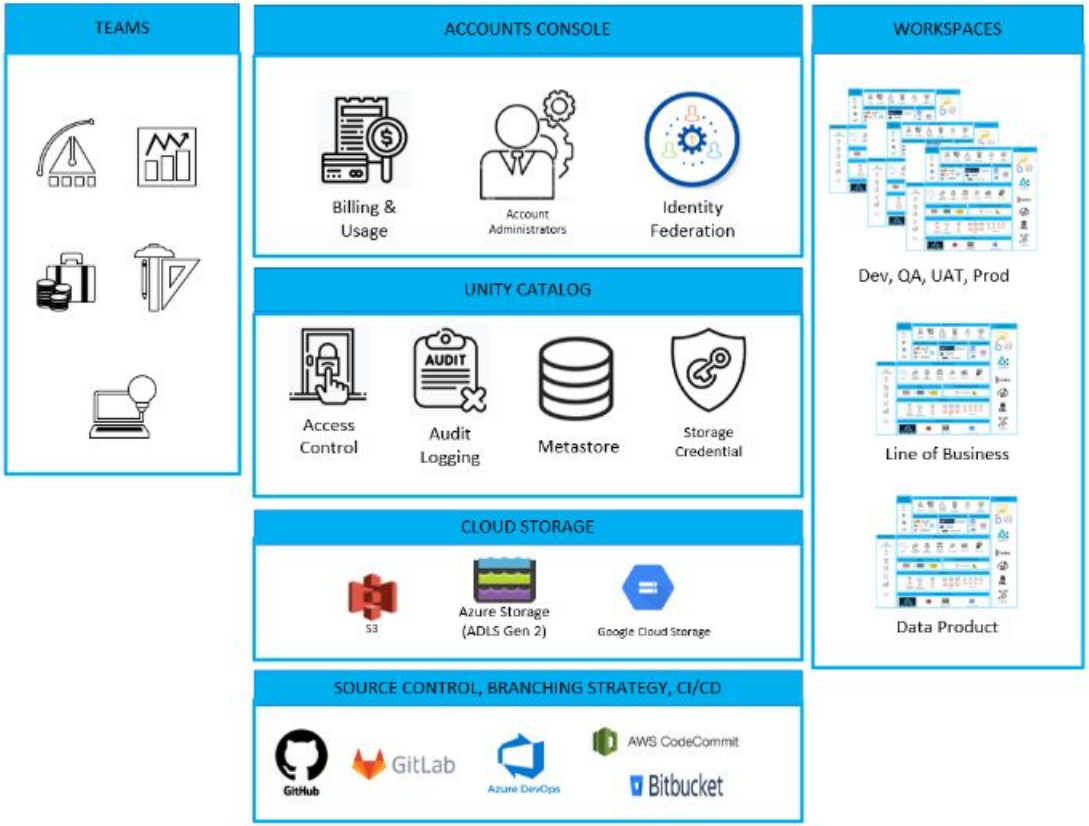
The Databricks Lakehouse



- Unified Data & AI Platform combining the strengths of Data Lakes and Warehouses.
- Support any Data and Processing workloads
- Build any Data Product
  - AI Models & Algorithms
  - Data Visualization Tools
  - Data Sharing Platforms
  - Data APIs
  - Data Management Tools
  - Data-driven Products and Services
  - Data Monetization Platforms
- Scalable, Flexible, Agile
- Make more informed decisions and achieve better business outcomes

The Databricks Lakehouse allows you to INNOVATE.

Databricks in Production



- Support multiple Teams and Environments
- Deliver Data Products
- Governance
- Code Quality
- Data Quality
- Security
- Administration & Monitoring
- Scalability
- Cost Management
- Continuous Integration & Deployment (CI/CD)
- Regulatory Compliance

The Production Databricks Lakehouse allows you to DELIVER VALUE in a reliable, secure and repeatable way.

What are your goals for this engagement?

- A Unity Catalog Roadmap
- Unity Catalog Configuration (Implementation done by PS/Partner)
- Jobs Upgraded (or a subset) (Implementation done by PS/Partner)
- PATTERN

What is your motivation for upgrading to UC?

- Lineage
- Auditing
- Ability to apply granular security
- Improve Data Quality
- Improve Discoverability
- Readiness for Lakehouse IQ
- Regulatory and Compliance Mandates
- Simplified Administration (moving to Account-level Administration)
- Keeping the Lakehouse viable and entrenching it in the client's ecosystem/solution
- Federation (Debezium, SQL Server)

Improving the Databricks experience with UC

Beyond blockers, Databricks is simply better with UC

- Enterprise reach
  - Governance across workspaces
  - Lakehouse Federation
- Governance for AI
  - UC for AI—models, features, volumes
- AI for governance
  - Lakehouse Monitoring
  - System tables & dashboards
- Delta sharing

- LakehouseIQ/Databricks Assistant
- Collaboration
  - Databricks Marketplace
  - Lakehouse Apps
  - Databricks Cleanrooms
- Real time lineage
- Search and insights
- Governed tags, ABAC and RBAC

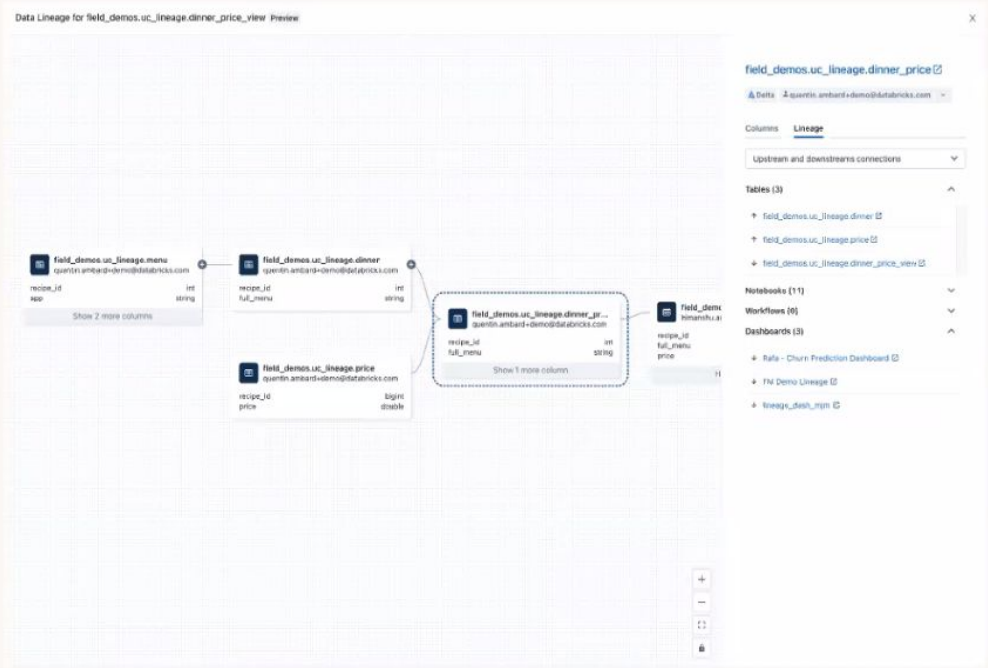
And the future looks bright

Simplified UX w/global data discovery	Multi-region, multi-cloud	Unified resource hierarchy	Fine-grained admin groups & roles w/RBAC	Global policy engine with federated pushdown
Platform HA / DR	Catalog of catalogs	CDO / CIO-relevant governance reporting & analytics	Biz user experience: Domains, semantics, glossaries	Governed tags & ABAC

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



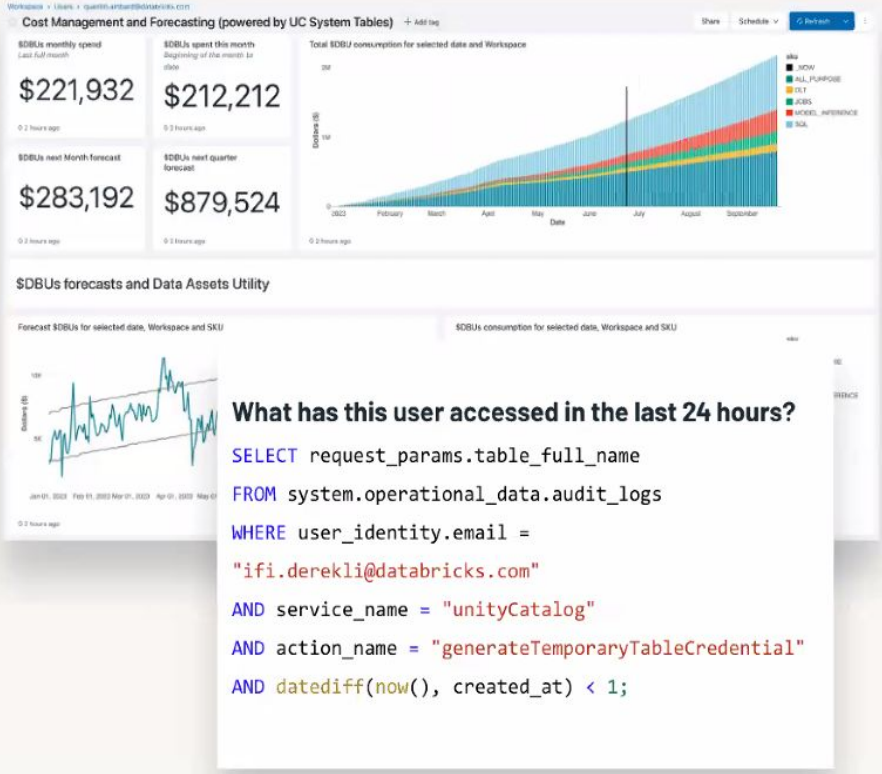
In Preview

Lakehouse Observability

Out of the box System Tables powers insight

Databricks-hosted analytical store for **all** of Databricks operational data - warm path used for historical customer observability, including

- Cost/usage analytics
- Efficiency analytics
- Audit analytics
- Data Quality analytics



Public Preview

In Preview

Lakehouse Federation—Databases/Warehouses

Unify your data estate with the Lakehouse

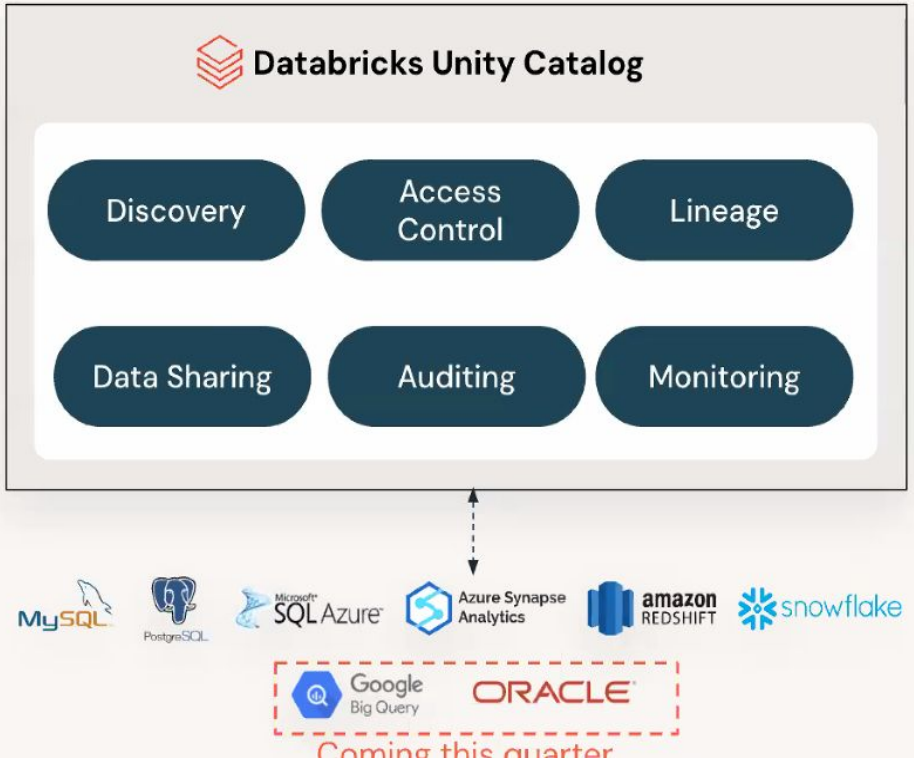
Discover, query, and govern all your data—no matter where it lives

Connect and query **external databases and warehouses**

Single point of access for all your data with performance optimizations

Unified governance across all data sources

Public Preview



Discovery Tags

Semantic layer for your lakehouse

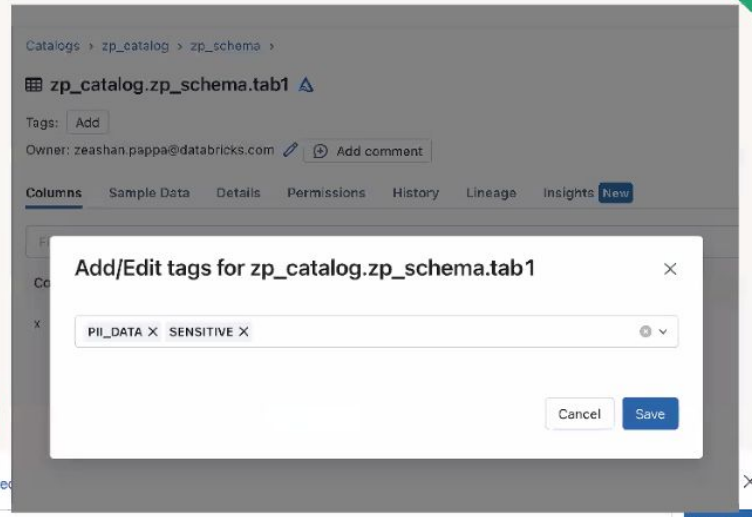
In Preview

Problem

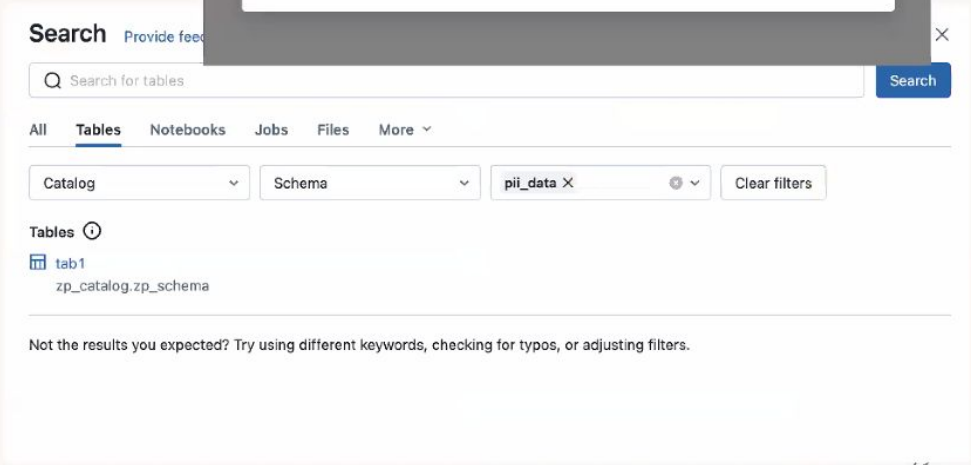
Searching for data assets in business terms or generally agreed upon taxonomies usually requires additional catalog tools.

Solution

- Discovery Tags allow you to tag Column, Table, Schema, Catalog objects in UC
- **Integrated search mechanism in UC allows you to search for objects by tag.**



Tag your data



Search based on tags

In Preview

Lakehouse Monitoring

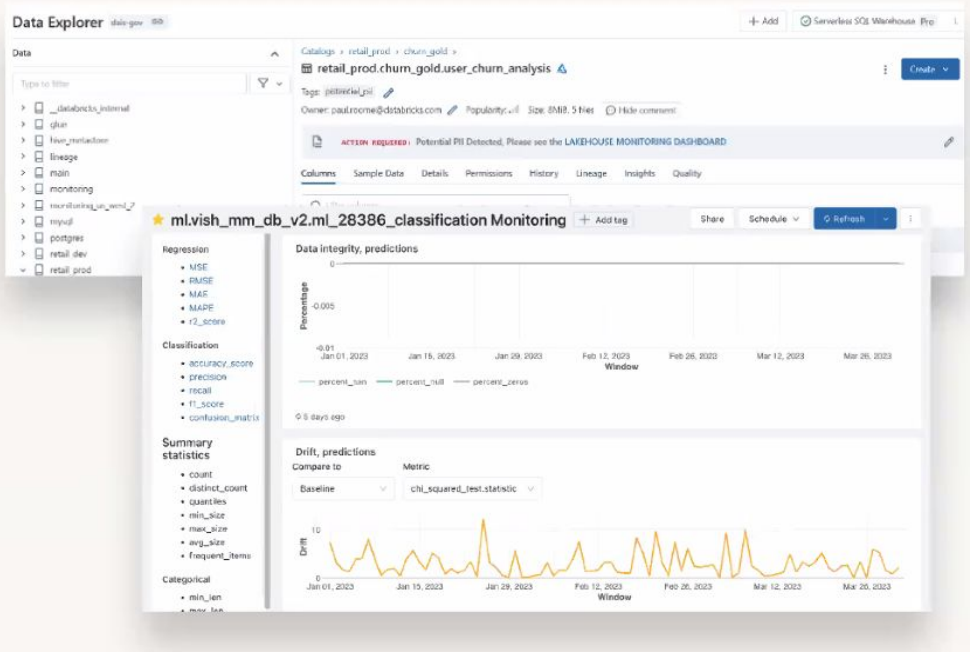
AI powered quality monitoring for data and AI

Unified solution to monitor entire data pipelines, from data and features to ML models

Proactive alerts on quality issues

Auto-generated dashboards for data and ML quality

Automated root cause analysis and impact assessment



Public Preview on AWS and Azure

Coming Soon!

LakehouseIQ

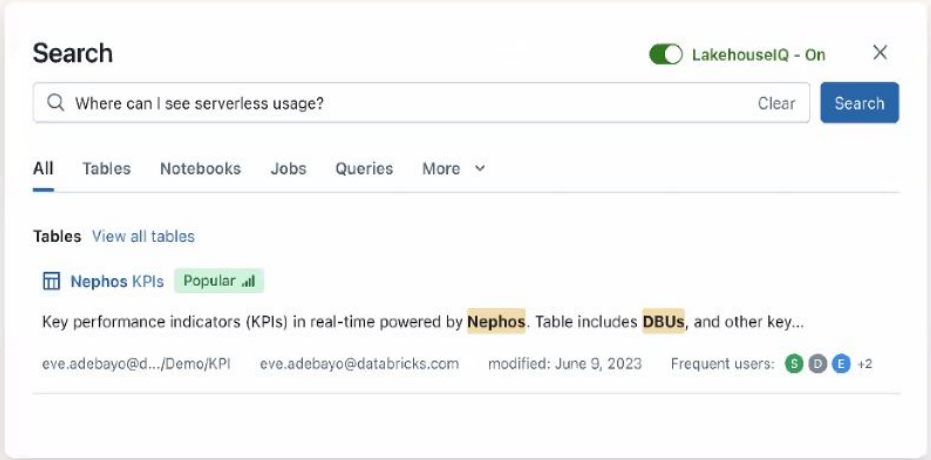
AI-powered knowledge engine that uniquely understands your business

Seamlessly access, analyze, and generate precise answers using natural language

Understands context using natural language (jargon, organizational structure, etc.)

Powers Search, Help, Assistant, Text-to-SQL

Secure and privacy-preserving



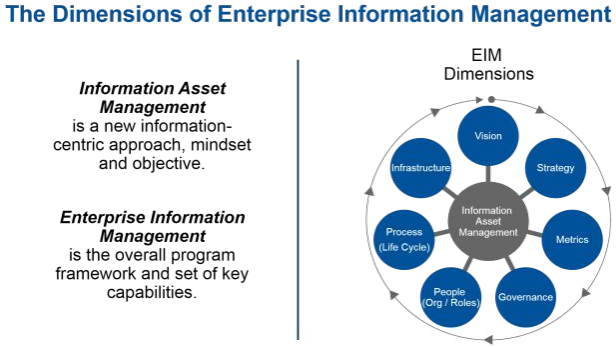
Public Preview in Q4

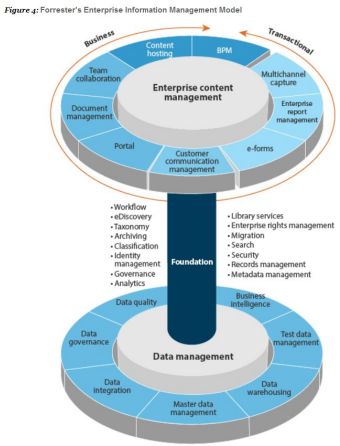
Why Data Governance?

- Adopt an Organization-wide Data Governance Strategy
  - Data Quality assurance needs to exist around all pipeline steps that produce a data product.
  - Data Catalog
    - Semantic Model for business concepts
    - Data Discovery: users should be able to discover relevant data easily
    - High quality metadata around the proper use of the data
  - Access Control
    - Implement fine-grade permission schemas from the beginning (column/row level access controls, role-based or attribute-based control).
    - Continuous Audit logging should be in place
- Use a centralized governance model to allow for distributed production and reuse of data products across your enterprise.
- Enable central platform operations teams to enforce compliance in operation and provide tooling while allowing data teams the flexibility to produce and work on data products.

Enterprise Information Management (EIM)

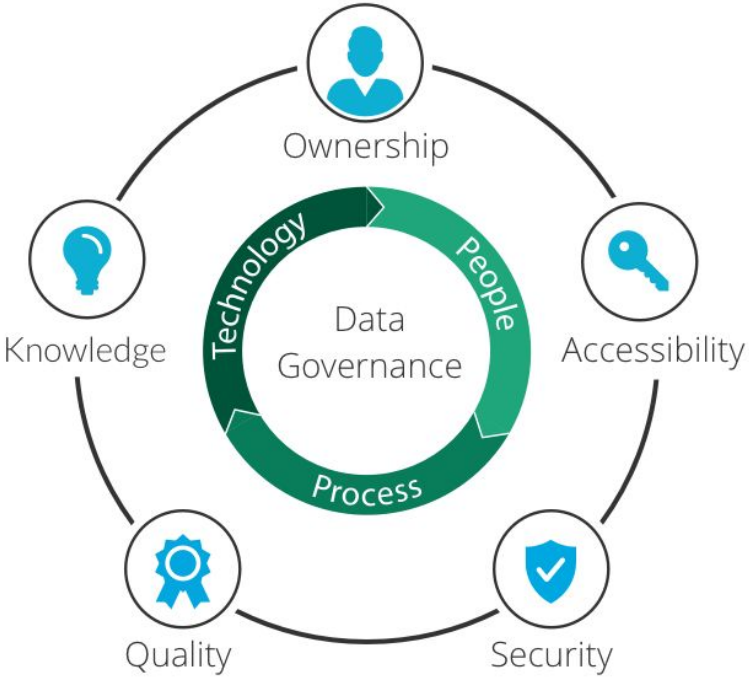
- Improve decision-making: Provide timely and accurate information to decision-makers, help business make better decisions and respond more quickly to changing market conditions.
- Increase efficiency and productivity: EIM helps businesses streamline their processes and eliminate redundant activities, which can result in cost savings and improved productivity.
- Enhance customer experience: EIM can help businesses gain a deeper understanding of their customers and their needs, which can result in more personalized and effective customer experiences.
- Mitigate risk: EIM helps businesses ensure that their information is accurate, complete, and secure, which can help mitigate the risks associated with data breaches, compliance violations, and other forms of data misuse or mishandling.
- Enable innovation: EIM provides a foundation for innovation by providing businesses with access to the data and insights they need to identify new opportunities and develop new products and services.





Data Governance

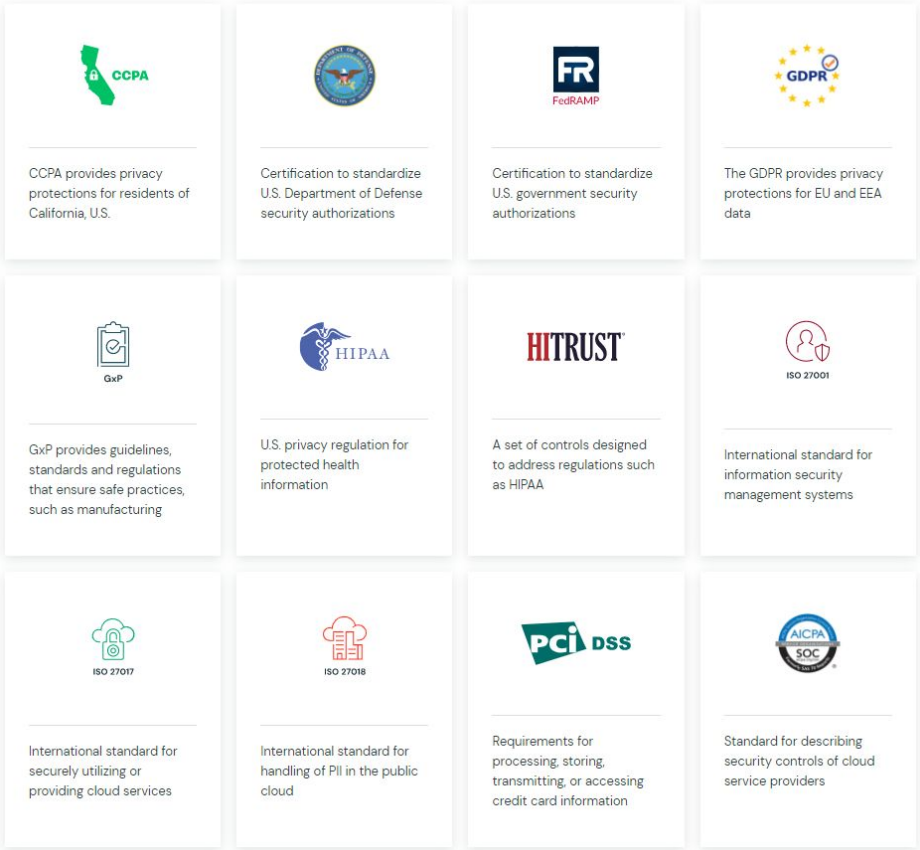
- **Data quality** – ensuring data is correct, consistent and free of “noise” that might impeded usage and analysis.
- **Data availability** – ensuring that data is available and easy to consume by the business functions that require it.
- **Data usability** – ensuring data is clearly structured, documented and labeled, enables easy search and retrieval, and is compatible with tools used by business users.
- **Data integrity** – ensuring data retains its essential qualities even as it is stored, converted, transferred and viewed across different platforms.
- **Data security** – ensuring data is classified according to its sensitivity, and defining processes for safeguarding information and preventing data loss and leakage.



Highly Regulated Industries

Industry	Regulated by	Protects
Healthcare and Pharmaceuticals	FDA, CMS, HIPAA	Patient safety and confidentiality
Financial Services	SEC, Federal Reserve System, FINRA	Investors and Financial Markets
Insurance	NAIC, FIO	Policyholder interests, Fair competition
Energy and Utilities	FERC, NRC	Environment and Public Safety
Aviation and Transportation	FAA, DOT	Public Safety and Transportation Infrastructure
Telecommunications	FCC	Fair competition and consumers
Environmental and Waste Management	EPA, OSHA	Environment and Public Health
Defense and Aerospace	DOD, NASA	National Security and Technological Innovation
Food and Agriculture	USDA, FDA	Food safety and Public Health
Education	Dept. of Education and accrediting agencies	Student welfare and educational quality
Gaming and Gambling	State and Federal Agencies	Fair competition and public trust

Regulatory and Compliance Requirements



Data and AI governance drives business value

"Organizations are finally realizing the value of data as an asset that needs to be protected, managed and maintained to increase asset value"

—

IDC

"Organizations seeing the highest returns from AI have a framework for AI governance to cover every step of the model development process"

—

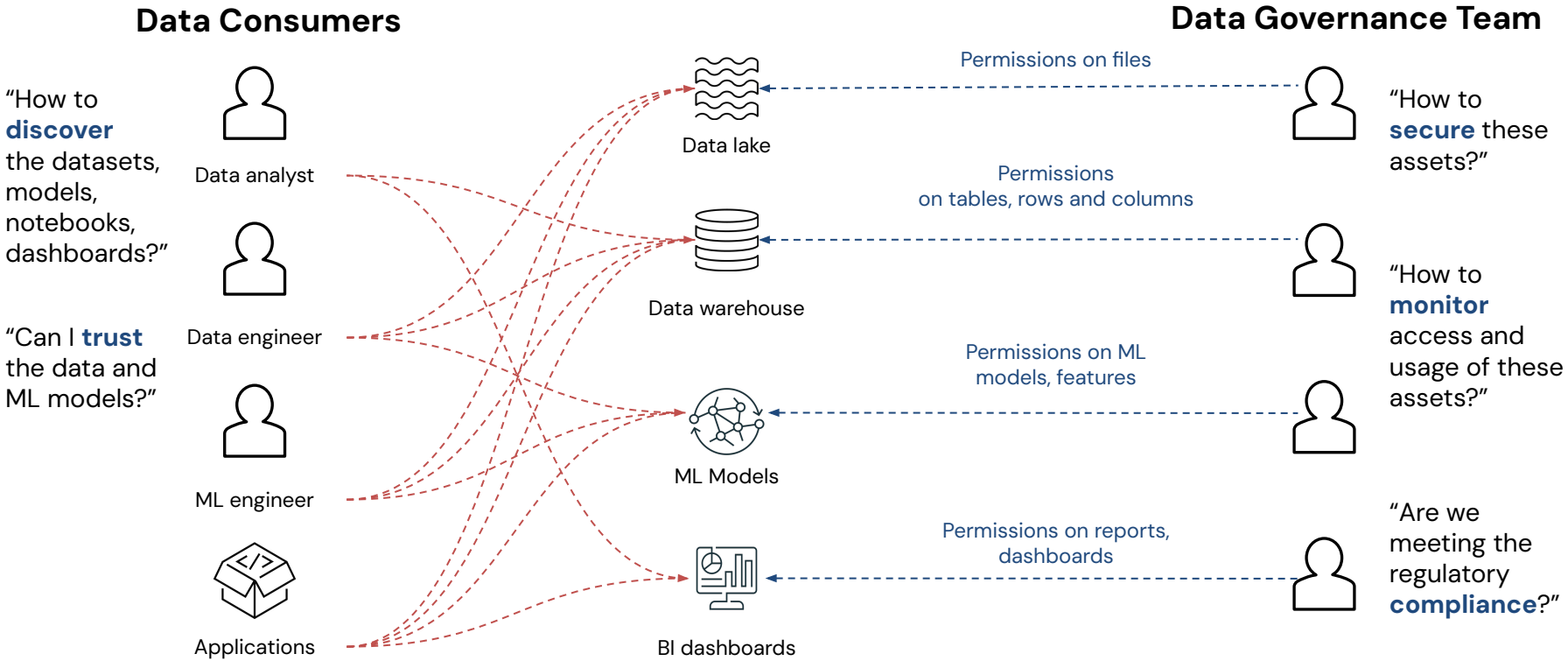
The State of AI in 2022, McKinsey & Co

"AI is now an enterprise essential, and as such, AI governance will join cybersecurity and compliance as a board-level topic"

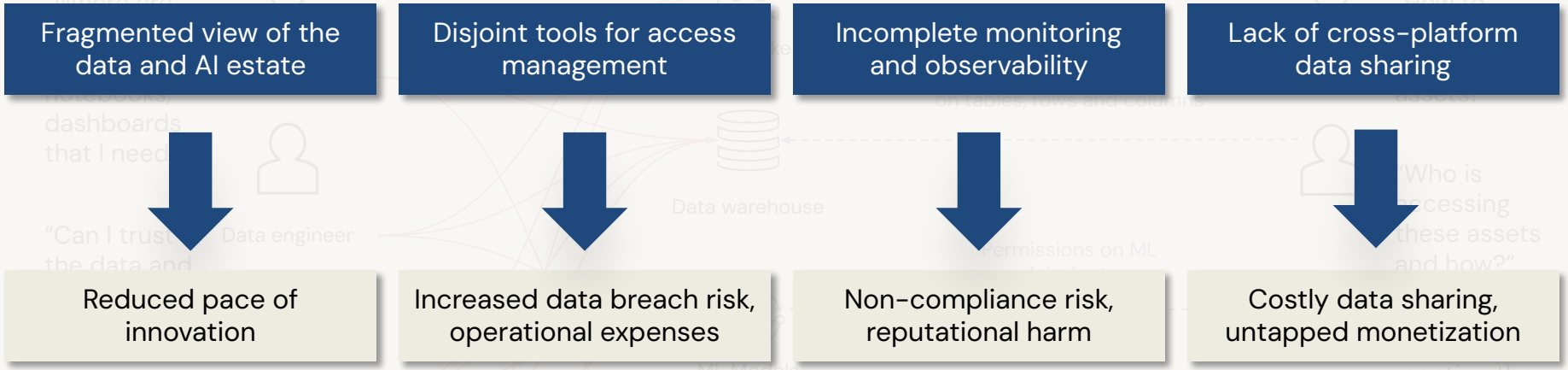
—

Forrester, 2023 AI Predictions report

Today, data and AI governance is complex



Today, data and AI governance is complex

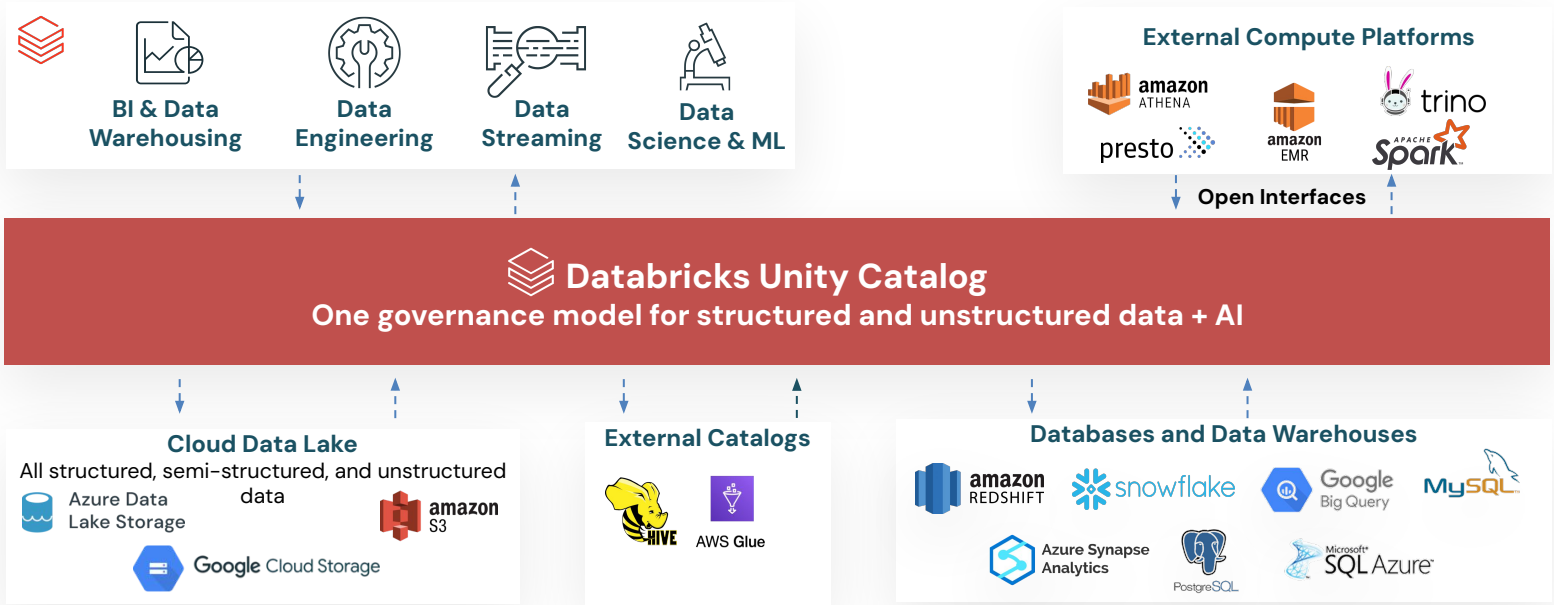


Organizations need a unified approach to governance for data and AI

Gartner®

By 2026, 20% of large enterprises will use a **single** data and analytics governance platform to **unify** and **automate** discrete governance programs

Databricks Lakehouse unifies data and AI governance



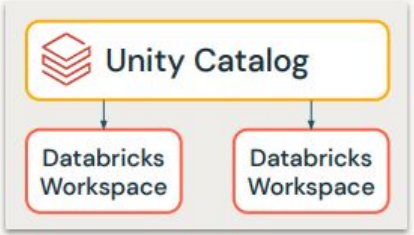
Unity Catalog

Unified governance for data, analytics, and AI

GA on AWS/Azure/GCP

Unity Catalog - Key Capabilities

- Centralized metadata and user management
- Centralized data access controls
- Data lineage
- Data access auditing
- Data search and discovery
- Secure sharing with Delta Sharing
- Marketplace for data, analytics and AI




```
GRANT ... ON ... TO ...
REVOKE ... ON ... FROM ...
```

Catalogs, Databases (schemas), Tables, Views, Storage credentials, External locations

Databricks Unity Catalog

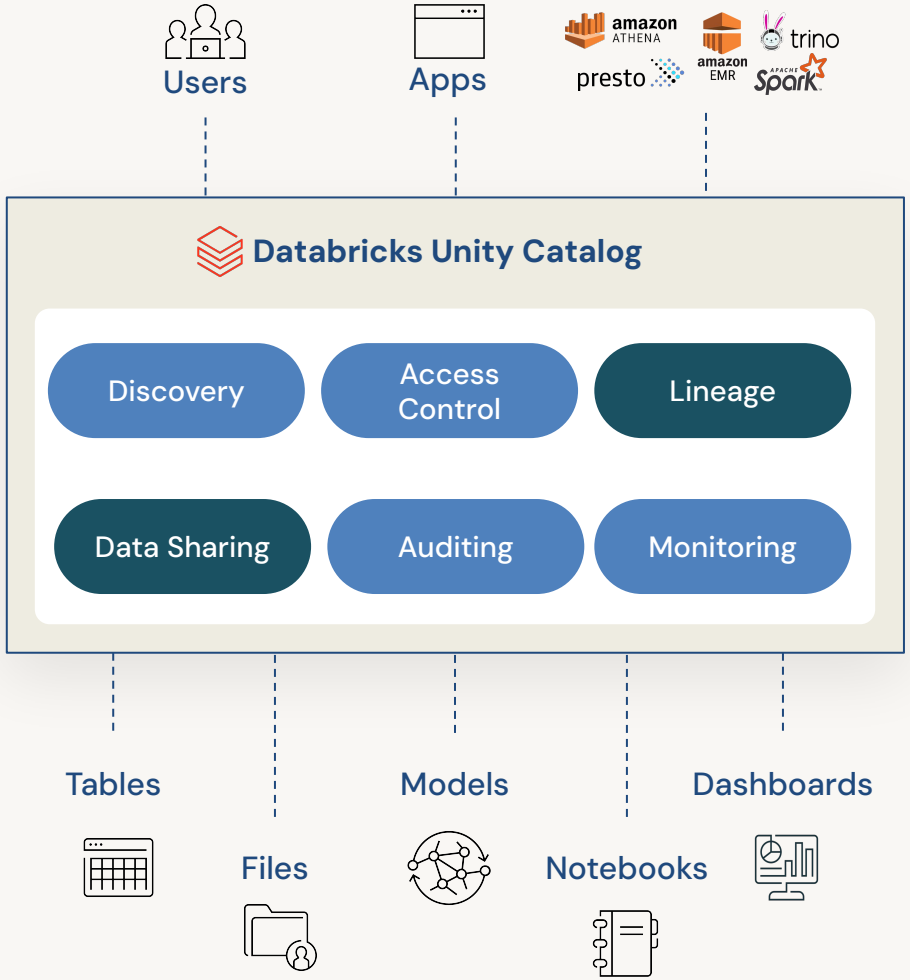
Unified governance for data & AI

Unified visibility into data and AI

Simple permission model for data and AI

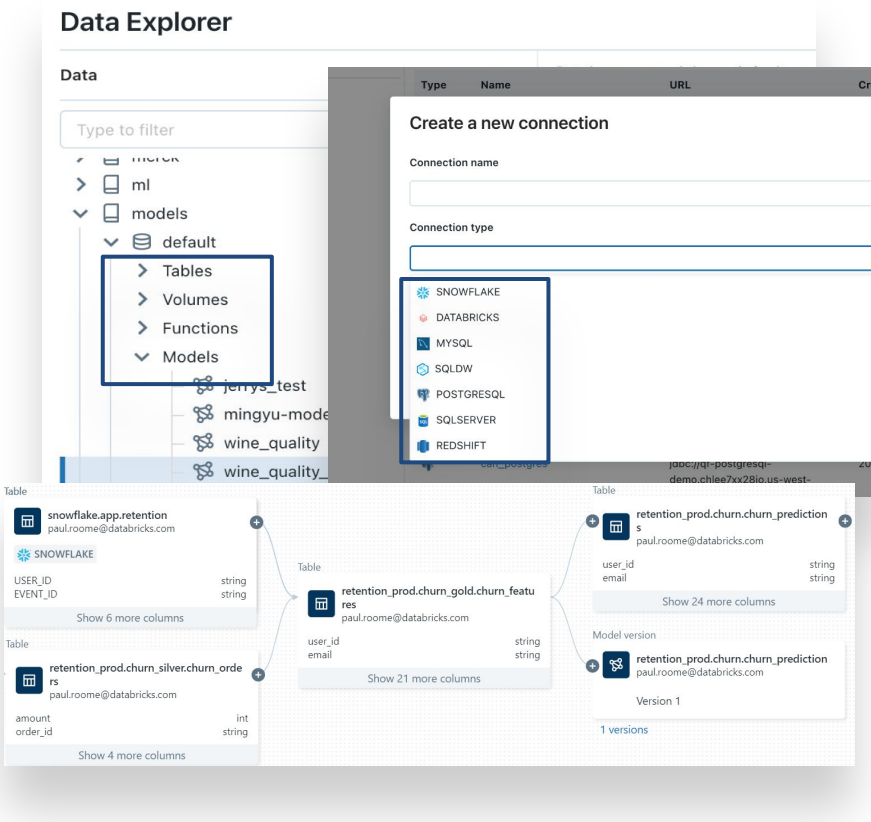
AI-powered monitoring and observability

Open data sharing



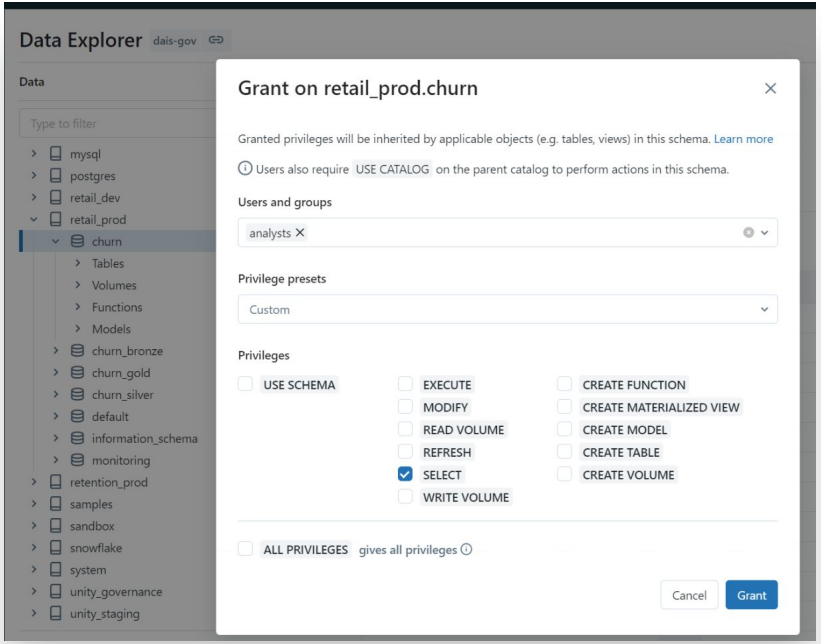
Unified visibility into data and AI

- **Discover and classify** structured and unstructured data, files, notebooks, ML models, and dashboards at one place
- Consolidate and query data from **other data platforms** using a **single point of access,** without moving or copying the data
- Build better **understanding of your data estate** with automated lineage, tags and auto-generated data insights
- Boost productivity by effortlessly exploring and gaining insights from your data and AI assets, using natural language **Coming Soon**



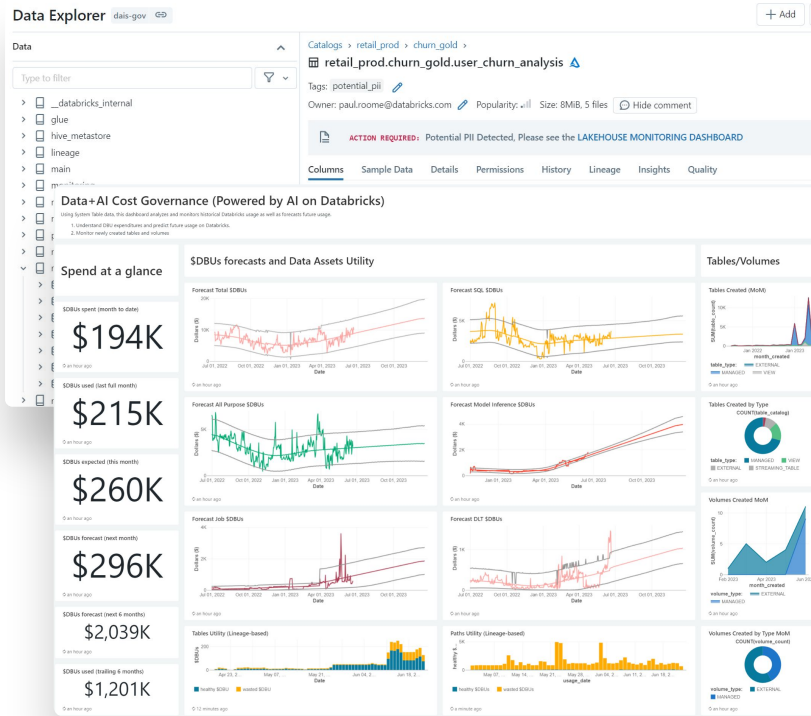
Single permission model for data and AI

- Secure your data estate with a **unified and simple interface** for managing and auditing access policies for all data and AI assets at one place
- Enable **fine-grained access controls** on rows and columns for enhanced security
- Securely access data from other computing platforms using **open interfaces,** with consistent permissions managed in one place



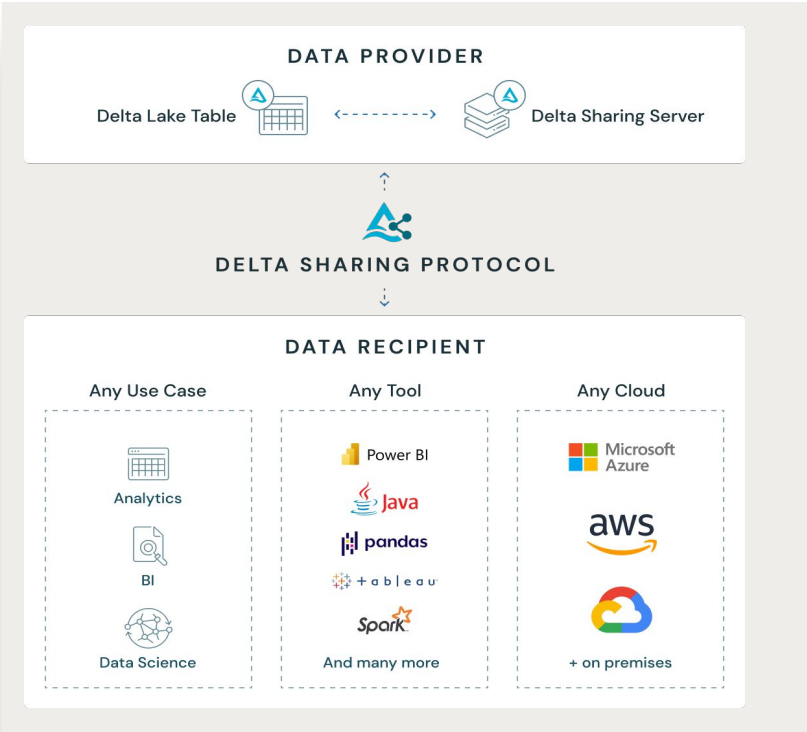
AI-powered monitoring and observability

- Receive **proactive alerts** for quality issues with data and ML models
- Access **real-time data lineage** down to the column level for efficient root cause analysis and impact assessment
- Utilize **auto-generated** dashboards to easily share data and ML quality reports with stakeholders
- Achieve full data and AI **observability** with operational intelligence using system tables for billing, auditing, lineage, and more.

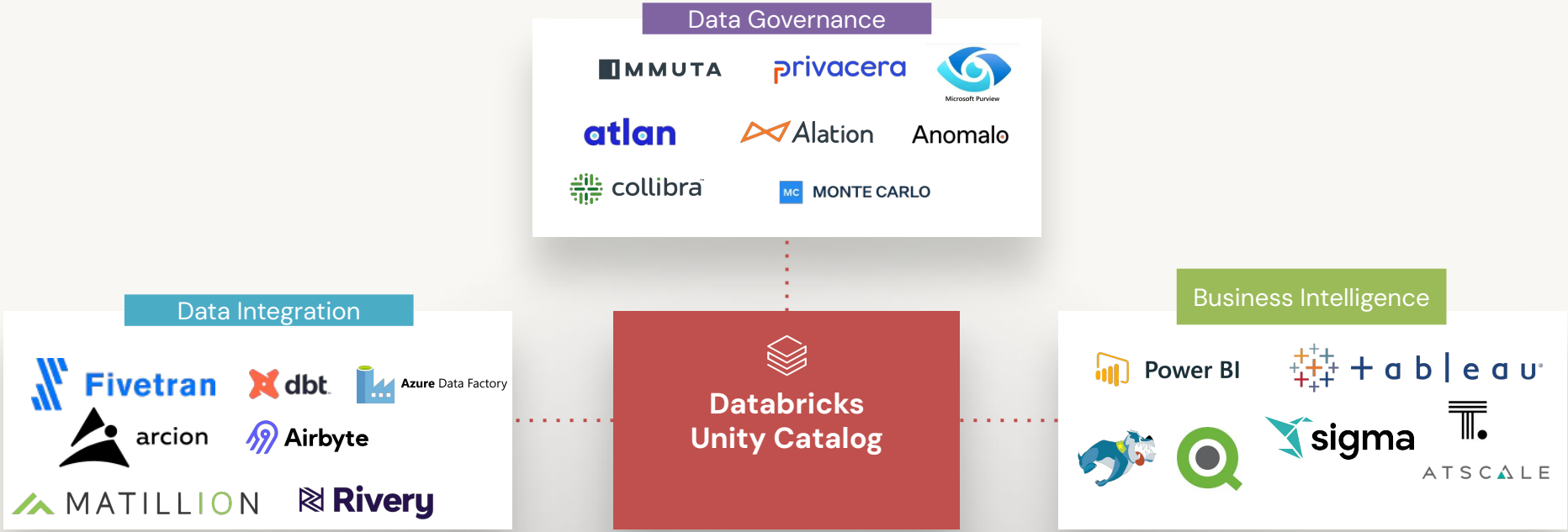


Open data sharing

- **Avoid vendor lock-in** with open source Delta Sharing for seamless data sharing across clouds, regions, and platforms, without replication
- Share **more than just data** - Notebooks, ML models, dashboards, applications
- Explore and monetize data products through an **open marketplace**
- Collaborate securely on sensitive data with **scalable data clean rooms**



Rich partner ecosystem



How do you want to implement Unity Catalog?

- Via Source Control and CI/CD
- Terraform or Alternative (e.g. Databricks SDK Python)
- Via Manual Deployment

Are there specific deliverables you care about more than others?

- Assessment
- YAML for CICD Pipelines
- Terraform or Databricks SDK Python implementation of Account Level Configuration and Security
- Account Level Users, Groups migrated from Workspace
- Job and Table Upgrades or Migrations (either in place or side by side)

DevOps Culture

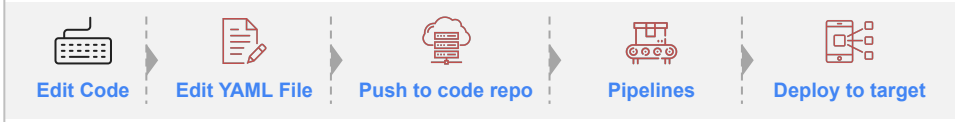
- Use CI/CD Pipelines to make possible the repeatable automation of the entire process from code commit to production.
- Use Infrastructure as Code (IaC) to automate all infrastructure and platform engineering.
- CI/CD Pipelines should automate the deployment of the entire solution stack, including:
  - compiling and building projects
  - deploying infrastructure
  - updating security and IAM Roles
  - applying configuration updates to resources
  - deploying code updates
  - executing automated test suites
  - reporting results
- Use multiple environments (e.g. Development, QA, UAT, and Production).

Why DevOps?

DEVOPS COMPANIES ACHIEVE

- 46x Deployment Frequency
- 2,555x Faster Lead Time For Changes
- 7x Lower Change Failure Rate
- 2,604x Faster Mean Time to Recover
- Faster Time to Market
- Increased Revenue

YAML PIPELINES



- Pipelines defined as code (YAML)
- Commit to Repo, versioned with Code
- Goes through same code review/pull request process
- Automate entire process from code commit to production (if Tests are successful)
- Approval-Gated Deployments

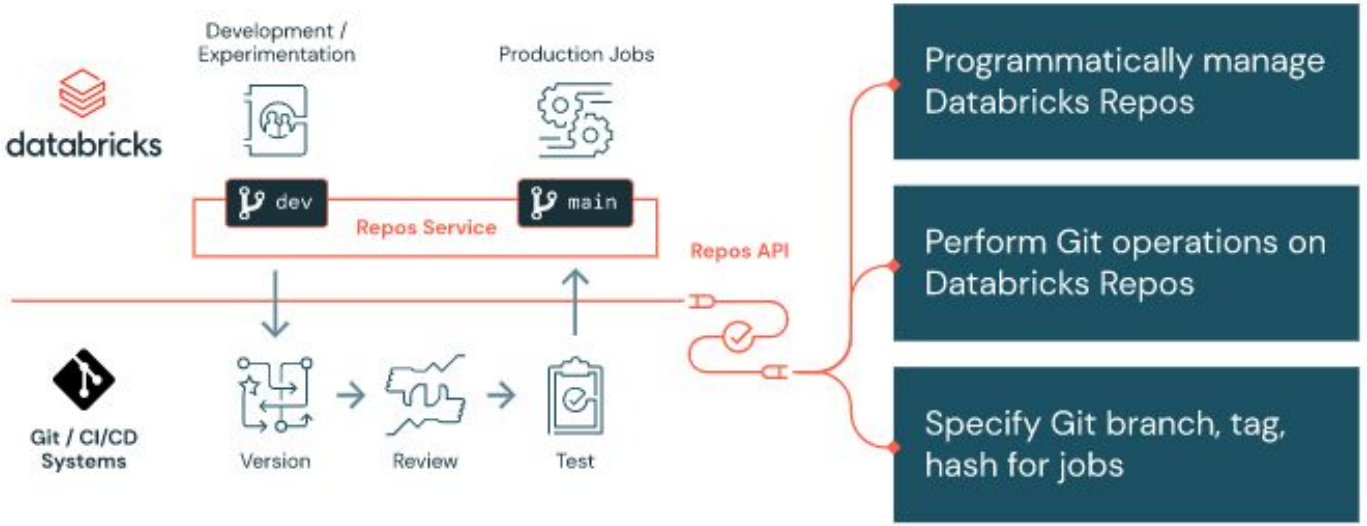
ASCEND CI/CD PIPELINE INCLUDES

Source: 2018 Accelerate: State of DevOps: Strategies for a New Economy. N.Forsgren, j.Humble, G.Kim. DevOps Research and Assessment (DORA)

Source Control

- Use the source control tool of your choice and automate the deployment of all infrastructure and platform engineering (DevOps Pipelines).
- Source Control Repo should contain artifacts to create all Infrastructure (IaC) and all Databricks platform engineering, including accounts, clusters, policies, Databricks SQL, Delta Live Tables, Git, Users/Groups, Instance Pools, Jobs, Security, Libraries, MLFlow, Repos, Secrets, etc.
- Use Databricks Repos Service to link Databricks with your source control tool.
- Create at least 2 shared repo folders: Dev, which references the current integrated feature branch and Prod, which references main. Use the CI/CD process to programmatically pull the latest changes in these branches.
- Allow developers to create their own repos for their individual feature branches in Dev.

Source Control

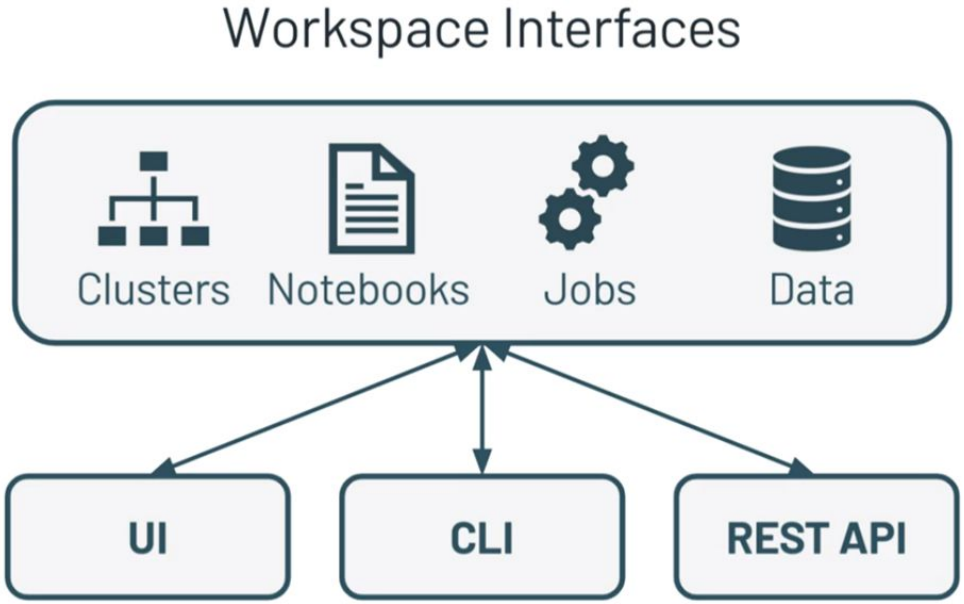


GitHub GitLab Azure DevOps AWS CodeCommit Bitbucket

IaC and Platform Engineering Recommendations

- Use Terraform to create the cloud level infrastructure, networking and security.
- Use your choice of Terraform or REST API in the language of your choice (e.g. Python, Powershell) to perform platform engineering within Databricks.
- Regardless of language, resource deployment should be highly modular and configurable.
- Support the automation of all aspects of platform engineering, including accounts, clusters, policies, Databricks SQL, Delta Live Tables, Git, Users/Groups, Instance Pools, Jobs, Security, Libraries, MLFlow, Repos, Secrets, etc.

Databricks Workspace Interface Options



Terraform

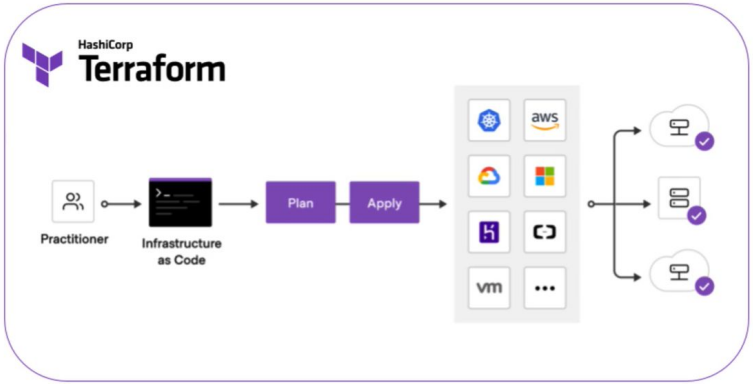
python

PowerShell

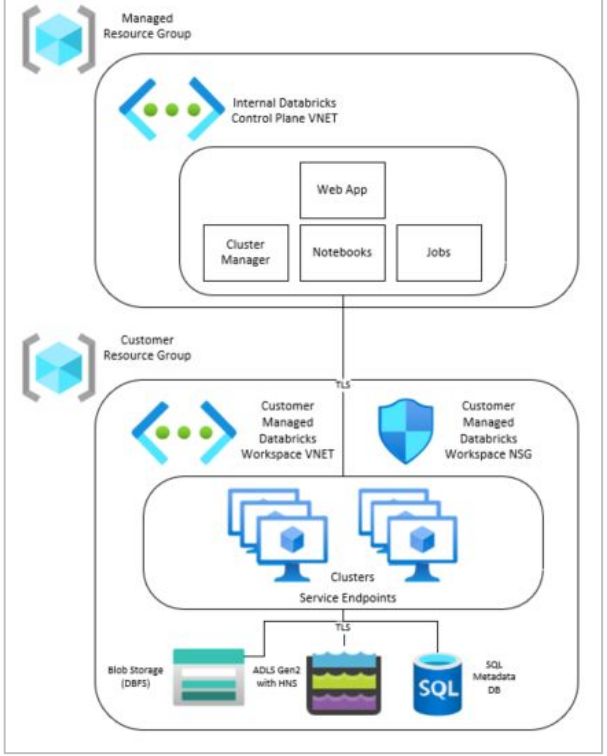
DevOps

Data Engineers

Infrastructure (As Code)

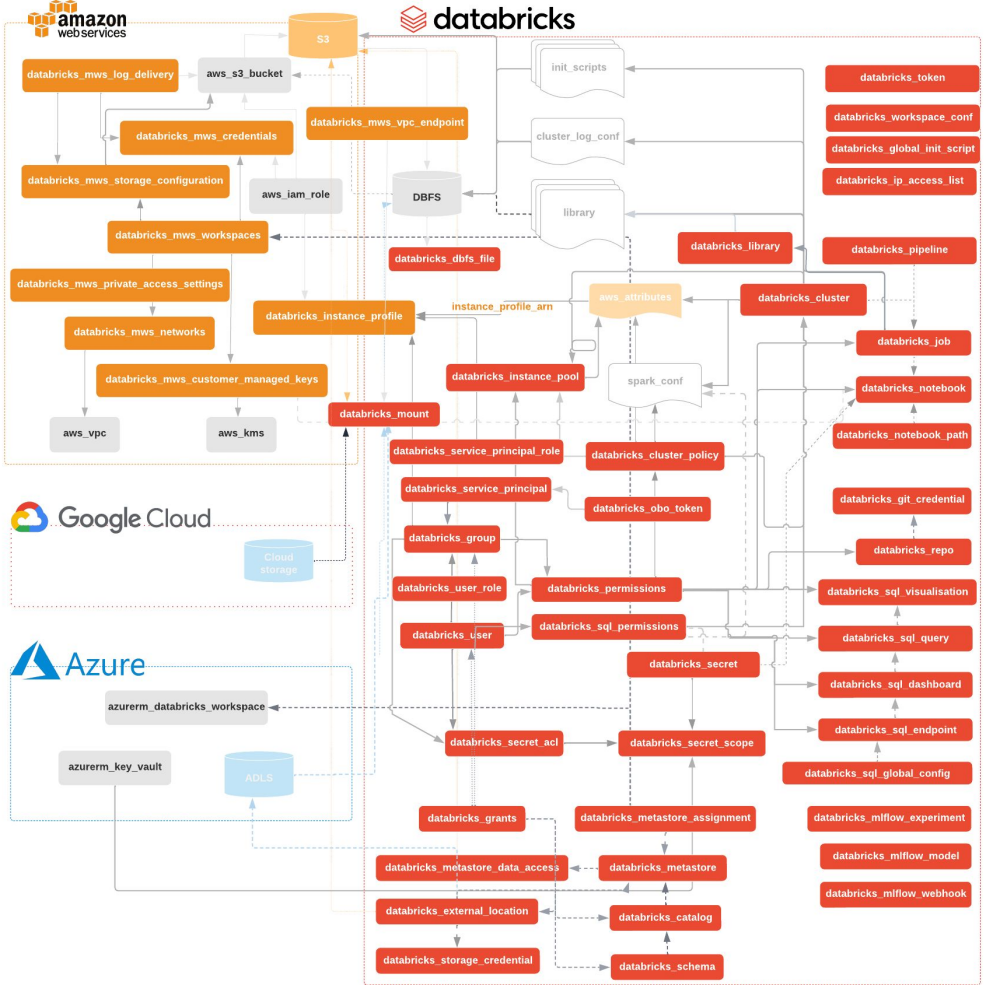


- Automate the creation of production-grade infrastructure
- Composable (expressed as JSON)
- Testable
- Releasable
- Reusable
- Repeatable
- Declarative (Get Projects done faster)
- Open Source (Wide adoption)
- Cloud Agnostic

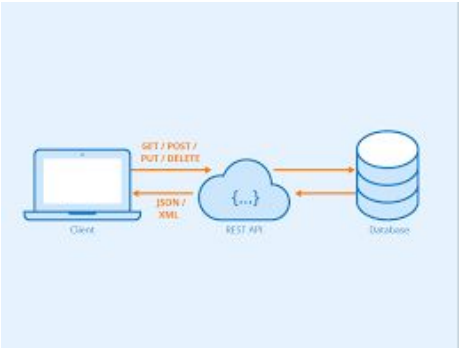


Databricks Terraform Provider

Databricks Labs Terraform



Databricks REST API



- Account API 2.0
- Clusters API 2.0
- Cluster Policies API 2.0
- Databricks SQL Queries and Dashboards API 2.0
- Databricks SQL Query History API 2.0
- Databricks SQL Warehouses API 2.0
- DBFS API 2.0
- Delta Live Tables API 2.0
- Git Credentials API 2.0
- Global Init Scripts API 2.0
- Groups API 2.0 1.2
- Instance Pools API 2.0
- Instance Profiles API 2.0
- IP Access List API 2.0
- Jobs API.0
- Libraries API 2.0
- MLflow API 2.0
- Permissions API 2.0
- Repos API 2.0
- SCIM API 2.0
- Secrets API 2.0
- Token API 2.0
- Token Management API 2.0
- Workspace API 2.0
- API 1.2