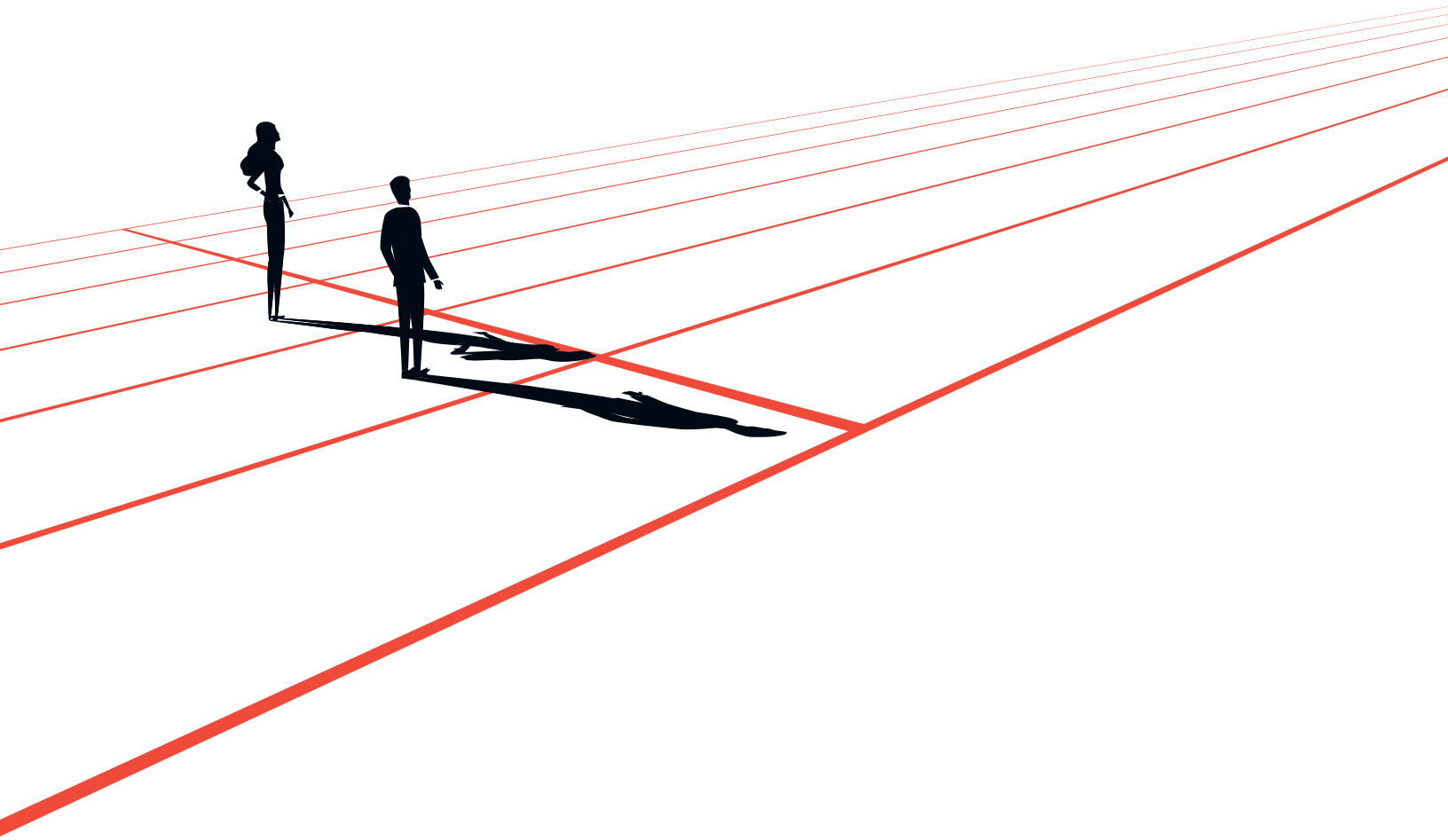
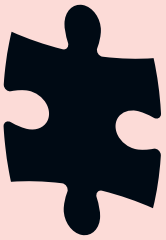
denodo × databricks

Whitepaper

Denodo and Databricks: A Modern Data Strategy

Table of Contents





Introduction

In today's dynamic business landscape, data-driven decision-making has become a critical factor for success. Organizations that empower their employees with timely access to accurate data and insights gain a significant competitive advantage. However, many companies still face challenges in implementing effective data-driven strategies across all levels of their organization.

This white paper introduces a comprehensive approach to overcoming the barriers to data-driven decision-making, consisting of a Databricks data lakehouse supported by the Denodo Platform, a logical data management solution. By leveraging these modern data architectures and embracing data democratization, businesses can unlock the full potential of their data assets and drive better outcomes.

Key challenges addressed in this paper include:

- Limited access to data and insights for line-of-business staff
- Over reliance on IT departments for basic data needs
- Underutilization of valuable data in decision-making processes
- Lack of self-service analytics tools for employees

By implementing the strategies outlined in this paper, organizations can:

- Empower employees at all levels to make informed, data-driven decisions
- Improve agility and responsiveness to market changes
- Enhance operational efficiency and reduce reliance on IT for data access
- Maximize the value of existing data assets
- Leverage the full power of AI to streamline and accelerate data management

This white paper will guide you through the essential steps to transform your organization into a data-driven powerhouse, enabling faster, more accurate decision-making across all business functions. It explores how the Denodo Platform and Databricks can work together to create a modern data strategy that can help organizations overcome data-access issues and bring agile decision-making across the organization.

We will lay the foundation for what a modern data strategy should offer, a strategy that can overcome today's data-access issues. We will explore the role of modern cloud data platforms like Databricks, and we will explain how the Denodo Platform can elevate an enterprise data warehouse (EDW) modernization effort into a successful enterprise-wide data strategy. By leveraging the Databricks platform, organizations can build a robust, scalable, and flexible data infrastructure that supports their evolving analytical needs and drives business growth.

The Need for a Modern Data Strategy

Many businesses still struggle to make quick, data-driven decisions, despite advancements in technology. Employees often lack direct data access or struggle with data silos, hindering their ability to leverage data for insights. Modern data strategies aim to democratize data access and foster a data-driven culture.

Putting such a strategy in action requires a number of important technical capabilities, including:

- **Self-Service Data Access:** Different personas need to be able to discover and access data themselves, without requiring support from IT. Organizations need to be able to leverage AI to support self-service data access with natural language queries.
- **Support for Multiple Personas:** Beyond self-service data access, multiple personas should also be granted selective access, based on their profiles.
- **Agile Data Provisioning:** Data provisioning should be fast and straightforward, without requiring extensive data preparation, and this should be true not only of all relevant internal data sources but also a variety of external ones.
- **Performance and Cost Optimization:** Organizations should have control over which queries to run, when, and at what cost, as well as the ability to optimize strategies for the best performance and the lowest cost. Organizations should also be able to leverage AI to boost performance.
- **Trust and Security:** Organizations need to be able to trust that their data is up-to-date, authorized, and free from viruses or malicious code.
- **Support for Enhanced Data Analytics Stacks:** Organizations need to be able to support virtually any data source or reporting tool, so they can address the widest possible variety of business use cases.

Databricks: A Modern Data Lakehouse

Databricks, a cloud data platform built on the data lakehouse architecture, is a central component for any modern data strategy. It addresses the limitations of traditional data warehouses, enabling organizations to store and process all of their data in a single, centralized location.

Unified Platform: Databricks provides a unified platform for data engineering, data science, machine learning, and business analytics. This eliminates the need for multiple different systems, simplifying data management.

Scalability and Performance: Databricks offers scalable storage and on-demand elastic compute, so organizations can process massive data volumes with high performance.

Open Format and Flexibility: Databricks leverages open data formats, improving compatibility with a wide variety of tools while reducing vendor lock-in. It supports diverse data types, including structured, semi-structured, and unstructured data, offering the flexibility to address many different use cases.

Cost-Effectiveness: Databricks can help reduce infrastructure costs through features like auto-scaling and spot instances.

The Denodo Platform: Unlocking the Power of Data with a Logical Approach

The Denodo Platform leverages a logical approach to data management. Unlike traditional data management platforms, logical platforms enable companies to manage data, which includes accessing it, governing it, and delivering it, all without having to first replicate all of the data into a single repository. While Databricks provides the foundation, the Denodo Platform acts as a crucial enabler, addressing several key requirements of a modern data strategy:

Self-Service Data Access:

- The Denodo Platform offers a unified data access layer that can expose all data through a single interface, regardless of the source's location or format. This creates a centralized marketplace for users to access data, whether it resides in Databricks, legacy systems, SaaS applications, or Excel spreadsheets. Data is presented as simple tables that can be queried using standard SQL, while the platform transparently manages execution across diverse systems. Additionally, it enables centralized management of data security and governance protocols, facilitating impact analysis and version control across various data sources.

- The Denodo Platform features an integrated data catalog that enhances data discovery, documentation, and querying for users of all technical levels. This comprehensive tool expands controlled data access, reducing shadow IT and security risks. The catalog offers advanced functionalities like usage statistics, collaboration features, data lineage, and direct integration with reporting tools. Uniquely, the Denodo catalog is part of the data delivery infrastructure, allowing direct data querying through a user-friendly query wizard, SQL shell, and personalized query saving. As data sources are imported into the platform, they become immediately available in the catalog, automatically synchronizing reporting tools with catalog content.

- The Denodo Platform offers a powerful feature for creating logical semantic models that are tailored to business needs, making data more accessible and understandable for non-technical users. Traditionally, organizations would create multiple data replicas using tools like dbt or by establishing separate data marts, which can be both time-consuming and difficult to maintain. This approach also introduces governance challenges due to the proliferation of data copies. In contrast, the Denodo Platform enables the creation of semantic layers that expose data in the required formats without replication. These logical models are developed using graphical wizards, preserving data lineage and offering impact analysis. The extended benefits of using the Denodo Platform's semantic layer include faster prototyping, reduced time-to-market, and lower operational and maintenance costs. By minimizing data replication, organizations can achieve greater efficiency and flexibility in their data management practices.

- The Denodo Platform leverages AI to support natural language queries, as well as to enhance data discovery with contextually relevant suggestions, based on individual user profiles.

Support for Multiple Personas: Data analytics today is characterized by diverse users due to self-service and data democratization trends, which brings data-driven decision-making across organizations but also raises security and privacy concerns. This diversity also introduces new data management requirements. The Denodo Platform caters to various user personas. This multi-persona support extends Databricks Lakehouse capabilities, making data more accessible and useful across the organization. Databricks provides the processing power for large data volumes, while Denodo offers a comprehensive logical layer for diverse data management needs.

- Business users: Easy-to-use data catalog for exploring and documenting information.
- Data engineers/BI Analysts: Design Studio with SQL shell, modeling tools, and performance tuning capabilities.
- Operations: Solution Manager for software provisioning, monitoring, and dataset lifecycle management.
- Data scientists: Unified data view across sources, with Python and notebook integrations.
- Data stewards: Data catalog for basic tasks, with integration options for enterprise governance tools

Agile Data Provisioning:

- The Denodo Platform simplifies data onboarding from both internal and external sources with just a few clicks. Users can access data on-demand from Databricks and other sources without needing replication, enabling features like advanced scheduling and incremental loads. This flexibility enhances agility and governance, particularly since some datasets cannot be replicated due to compliance issues.

- The Denodo Platform also enables the quick prototyping of data resources, facilitating faster validation by stakeholders who can choose between direct access or data persistence, for better performance. The Denodo Platform's logical data models integrate multiple sources and technologies and feature a multi-source execution engine. The Denodo Platform also supports full lifecycle management, including integrated version control with Git.

Performance and Cost Optimization:

- The Denodo Platform complements Databricks' performance with its Smart Query Acceleration feature. It intelligently optimizes queries while reducing the cost of Databricks processing. The system leverages AI to further improve query optimization. Additionally, the Denodo Platform can offload less critical workloads to less expensive locations, without impacting business operations.

Trust and Security:

- Trust in data democratization relies on security, privacy, and proper access control. The Denodo Platform offers comprehensive security features, including pass-through credentials, OAuth support, and fine-grained access control with RBAC and ABAC.

- The Denodo Platform provides centralized security management across a variety of data sources, supporting external sources and data sovereignty. The platform also integrates with external data governance tools, simplifying policy enforcement and enabling strategies like data mesh. This approach provides consistent, secure access to the right data for authorized users across the entire data ecosystem.

Enhances data analytics stacks: This versatility enables seamless integration with diverse applications and simplifies data interactions across platforms without additional coding.

The Denodo Platform enables:

- Out-of-the-box polyglot output

- Automatic exposure of logical layer data models in a variety of formats:
  - SQL: JDBC, ODBC, Python drivers
  - MDX: For cube format querying (e.g., Excel pivot tables)
  - Web services and APIs: RESTful, OData 4, GeoJSON, GraphQL

- Compatibility with:
  - Major reporting tools (Tableau, PowerBI, etc.)
  - Data science notebooks (Jupyter, Zeppelin)

- Support for read-only and IDU operations

- Self-documented APIs using OpenAPI and SwaggerUI

Futureproof Your Data Architecture

A logical data layer, such as the one provided by the Denodo Platform, can significantly future-proof data infrastructure by creating an abstraction layer between data sources and consuming applications. This separation enables seamless backend changes, including migrations between different systems (e.g., Teradata to Databricks), without disrupting end-user experiences or requiring modifications to existing reports and dashboards. The platform's multi-source federation capabilities also facilitate hybrid environments during transition periods, supporting both cloud and on-premises datasets simultaneously. This flexibility empowers organizations to strategically manage data placement, eliminating vendor lock-in, offering unparalleled adaptability in today's rapidly evolving data landscape.

Denodo and Databricks Integration Points

The Denodo Platform seamlessly integrates with Databricks through connectors for SparkSQL, support for lakehouse file formats like Delta Lake, and integration with data science notebooks.

KEY BENEFITS OF THE JOINT SOLUTION

A full-featured data platform like Databricks is often not quite enough to fulfill all the requirements of a modern data strategy. However, combined with the Denodo Platform, Databricks can enable a future-proof data strategy that can meet virtually any modern use case.

Working together, a joint Databricks/Denodo Platform implementation enables:

Accelerated Data Delivery: Denodo and Databricks work together to streamline data access and delivery, empowering users with real-time data for faster decision-making.

Enhanced Data Governance and Security: Denodo's centralized security and governance features extend Databricks' capabilities, ensuring data security and compliance.

Improved Data Discovery and Awareness: The Denodo Platform's data catalog and self-service tools make it easier for users to find, understand, and use all data.

Reduced Costs: Denodo's optimization capabilities and Databricks' cloud-native architecture help organizations optimize infrastructure costs.

Unified Data for BI and ML: This solution unifies data from Databricks and other sources, supporting a wide range of use cases, from BI reporting to advanced analytics and ML.

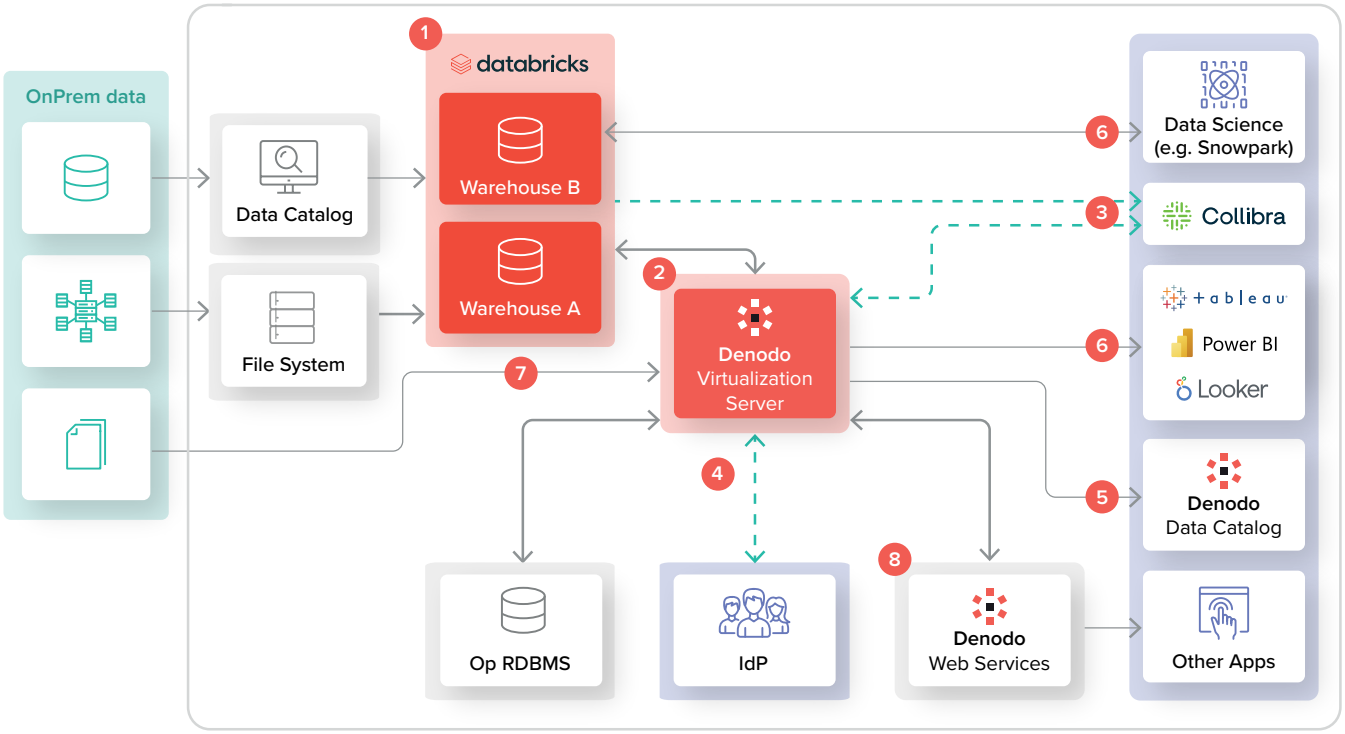
Future-Proof Data Architecture: Denodo's logical approach allows organizations to adapt to future changes in their data infrastructure without impacting data access.

Conclusion

By combining the Denodo Platform's logical data management capabilities with Databricks' lakehouse architecture, organizations can implement a modern data strategy that empowers data-driven decision-making across the enterprise. This joint solution delivers agility, flexibility, performance, and security, unlocking the full potential of your data.

Denodo and Databricks Reference Architecture

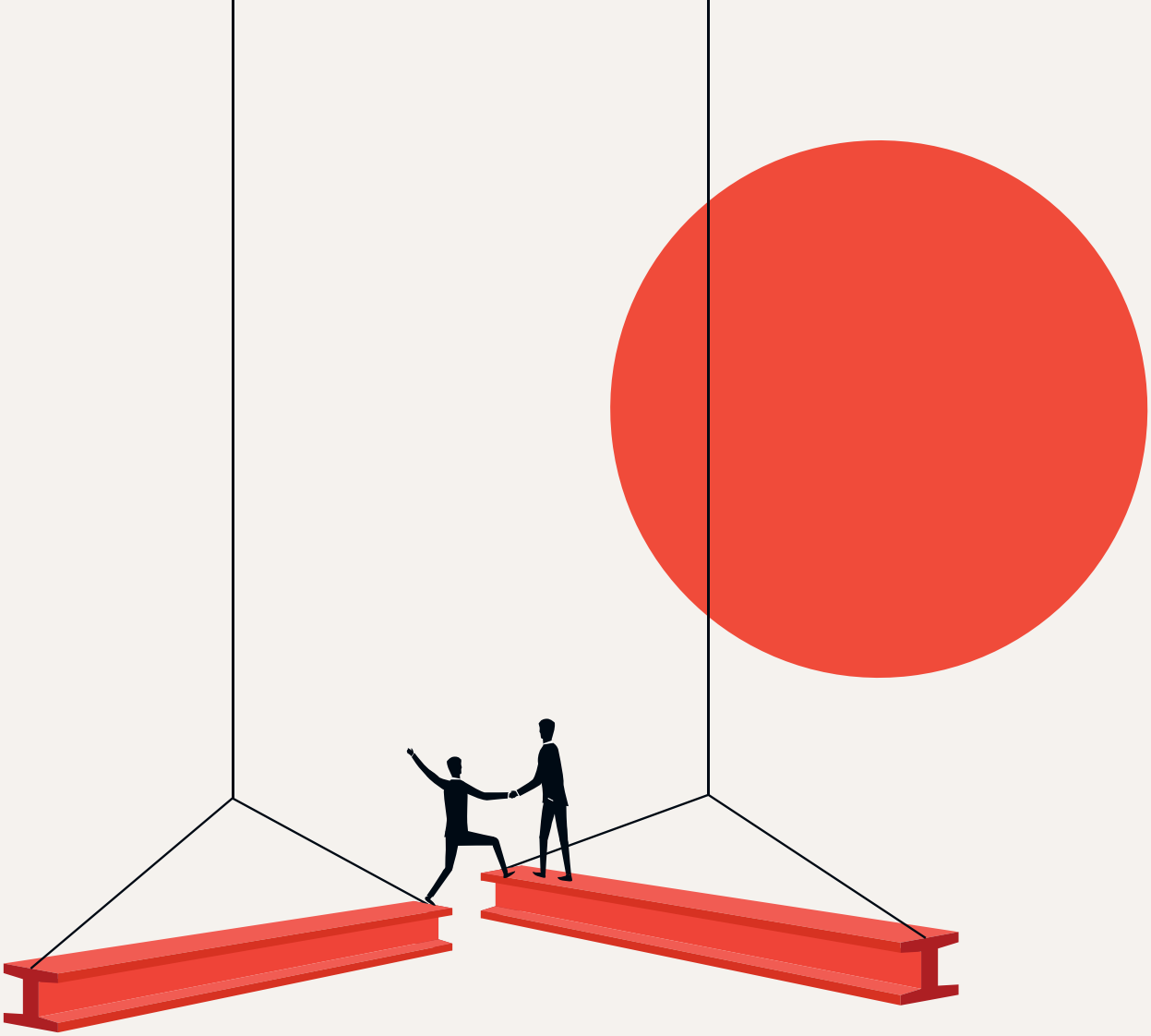
IN AWS, MICROSOFT AZURE, AND GOOGLE CLOUD



CLOUD SERVICES

Data Catalog	File System	Op RDBMS	IdP
Amazon Glue	Amazon S3	Amazon Aurora	AWS SSO
Azure Purview	ADLS	Azure SQL	Microsoft Entra ID
Google Data Catalog	Google Cloud File System	Google Cloud Spanner	Google Identity Platform

1. Databricks is the central component of the analytics stack, providing storage and processing power for all analytical workloads.
2. Denodo provides a semantic layer that centralizes access to all data, including other databases and data still on prem.
3. Data Governance tools like Collibra often play a significant role for Data Stewards to manage and govern approval workflows .
4. Denodo can enforce access control using ABAC and RBAC policies, that can even leverage metadata managed in Collibra like PII or other tags.
5. Denodo's Data Catalog offers an excellent tool for end users to explore, browse and document data.
6. Most data, for example for reporting, is delivered via Denodo, that enforces access control policies on its semantic models. However, some specialized access (e.g. data scientists using Snowpark) may have direct access to Databricks.
7. Denodo can manage data replication and ingestion across sources, including Databricks, on-prem and other databases.
8. Applications that need API-based access can leverage Denodo's native support to expose data as REST, OData and GraphQL resources.



databricks | Visit: databricks.com Follow: @databricks

denodo | Visit: denodo.com Follow: @Denodo