LumenData

DATA SHEET

Modern Data Management with the Databricks Data Intelligence Platform - I

databricks

Current challenges faced across Industry is correct way of defining and implementing data management with modern data stack.

Companies are rapidly adopting the Data Lakehouse architecture to enable their organizations to better utilize data for analytics and AI. The move to the Lakehouse requires a different mindset regarding the lifecycle of data.

Data management has been a common practice in the industry for many years, although not all organizations have used the term in the same way.

At Databricks, data management is defined as all disciplines related to the lifecycle of data which includes gathering data, processing data, governing data, sharing data, analyzing it and optimizing it.

LumenData being a Platinum Enterprise Partner of Informatica and expertise in Data Management using IICS has been at forefront of implementing industry best practices for data management.

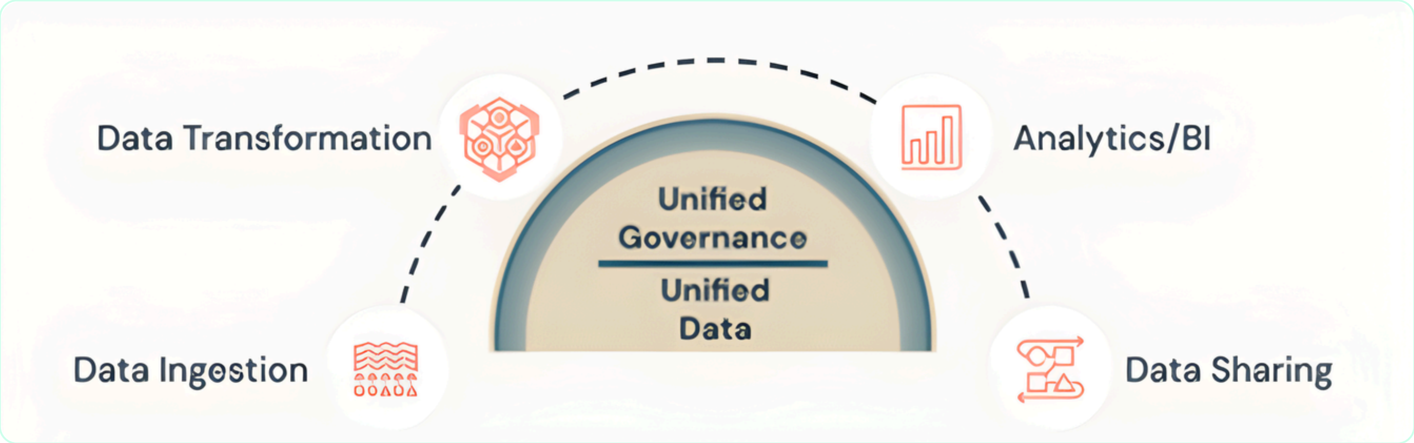

Image Credit: **Databricks**

Challenges of Modern Data Management

1 Fragmented Data Flow Across the Organization

- Having a steady and trustworthy stream of data across teams and business processes is essential for innovation and success.
- As organizations begin to leverage advanced analytics and generative AI, data needs to be trusted, timely, and properly governed to enable faster decision.

2 Growing Demand for Analytics and AI at Scale

- Data is being leveraged more for product innovation, inter-team collaboration, and market development.
- Industry studies indicate that all the organizations that have adopted a lakehouse architecture have been able to meet their data and AI objectives, thus underlining the importance of scalable analytics and AI infrastructure that does not compromise on reliability and governance.

3 Split Architectures Create Data Movement Overhead

- Most of the data in the enterprise ends up in data lakes for preparation and machine learning, while the curated data is constantly being loaded into data warehouses for business intelligence and reporting.
- This is causing latency, duplication, and complexity.

4 Inherent Limitations of Lakes and Warehouses

- Data lakes are best suited for machine learning because of their open nature and flexibility in the ecosystem, but they are not good at BI performance and data quality.
- Data warehouses, on the other hand, are good at BI but are proprietary, SQL-only databases with poor support for machine learning.

The Core Problem: Architectural Incompatibility

The root problem in data management lies in the attempt to unify systems that are designed to manage different workloads.

If this is not achieved, then there will be fragmented governance, performance, and agility in the delivery of analytics and AI. Data and usage patterns evolve with time.

As new data is ingested into the data lake and processed for the data warehouse, the schema must evolve to accommodate new data types and sources.

New analytics and AI applications drive new queries that join data in more complex ways. Consequently, tables that were optimized for previous applications may become suboptimal over time.

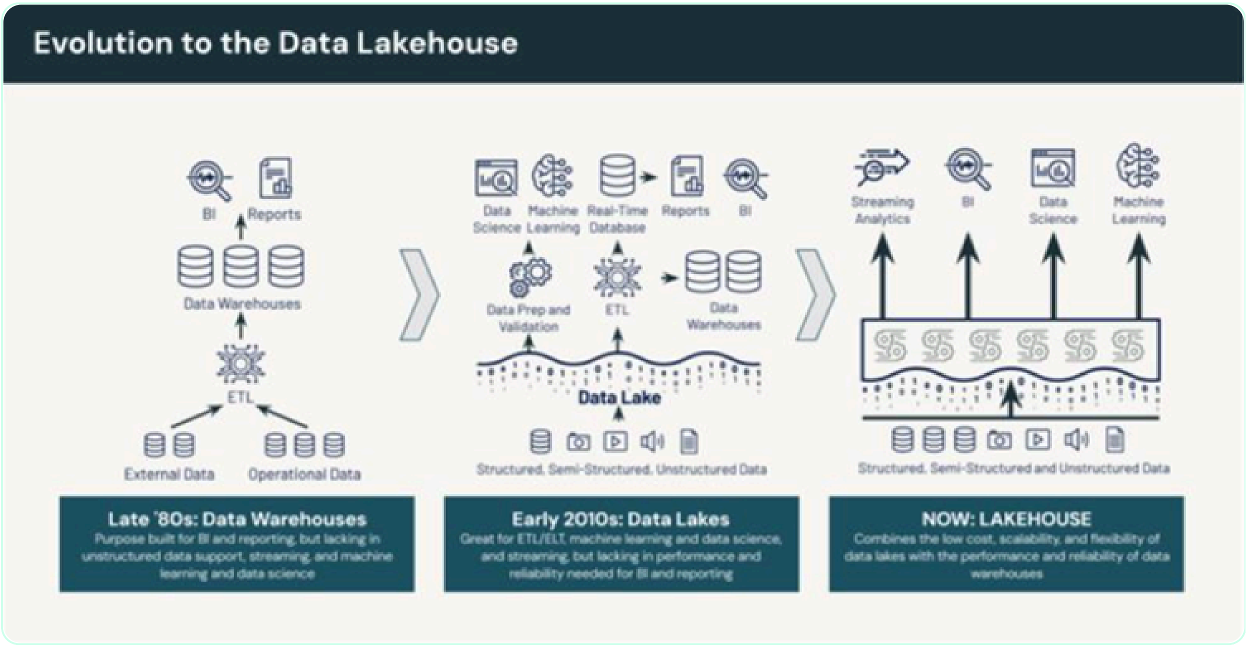
Evolution to the Data Lakehouse
BI
Reports
Data Warehouses
ETL
External Data
Operational Data
Late '80s: Data Warehouses
Purpose built for BI and reporting, but lacking in unstructured data support, streaming, and machine learning and data science
Data Science
Machine Learning
Real-Time Database
Reports
BI
Data Prep and Validation
ETL
Data Warehouses
Data Lake
Structured, Semi-Structured, Unstructured Data
Early 2010s: Data Lakes
Great for ETL/ELT, machine learning and data science, and streaming, but lacking in performance and reliability needed for BI and reporting
Streaming Analytics
BI
Data Science
Machine Learning
Structured, Semi-Structured and Unstructured Data
NOW: LAKEHOUSE
Combines the low cost, scalability, and flexibility of data lakes with the performance and reliability of data warehouses

Image Credit: **Databricks**

Solution: Data Management on Databricks

Unified Platform for the Entire Data Lifecycle

- The Databricks Data Intelligence Platform integrates data engineering, analytics, business intelligence, and AI on a single platform.
- It removes fragmented systems and helps organizations manage data end-to-end, from ingestion and governance to analytics and AI, with the same levels of performance, security, and scalability.

Few Features and Services of Databricks

Open Data Lakehouse as the Core Foundation

- At the core of the platform is an open data lakehouse architecture, which enables organizations to store data in their chosen cloud storage solution using open formats such as Delta Lake, Apache Iceberg™, Parquet, JSON, and AVRO.
- This openness enables data portability, flexibility, and freedom from vendor lock-in.

Multi-Format, Multi-Engine Interoperability

- In traditional lakehouse architecture, the implementation was done in a single table format, which meant that organizations had to make decisions on tooling based on storage formats.
- **Databricks** eliminates this limitation by allowing multiple open table formats, which means that teams can select the best compute engines and tools.

Unity Catalog: The Governance Control Plane

- Historically, each lakehouse provider had proprietary catalogs with limited visibility and access across tools and engines.
- This led to fragmented visibility, governance, and collaboration across platforms. There was no single catalog that could govern data and AI resources holistically across the ecosystem.
- The Unity Catalog addresses these issues by offering a single governance layer across formats, engines, and workloads. It manages read and write permissions for both Delta Lake and Apache Iceberg and is fully compatible with the Iceberg REST Catalog API.

We will deep dive into Unity catalog in the coming sections.

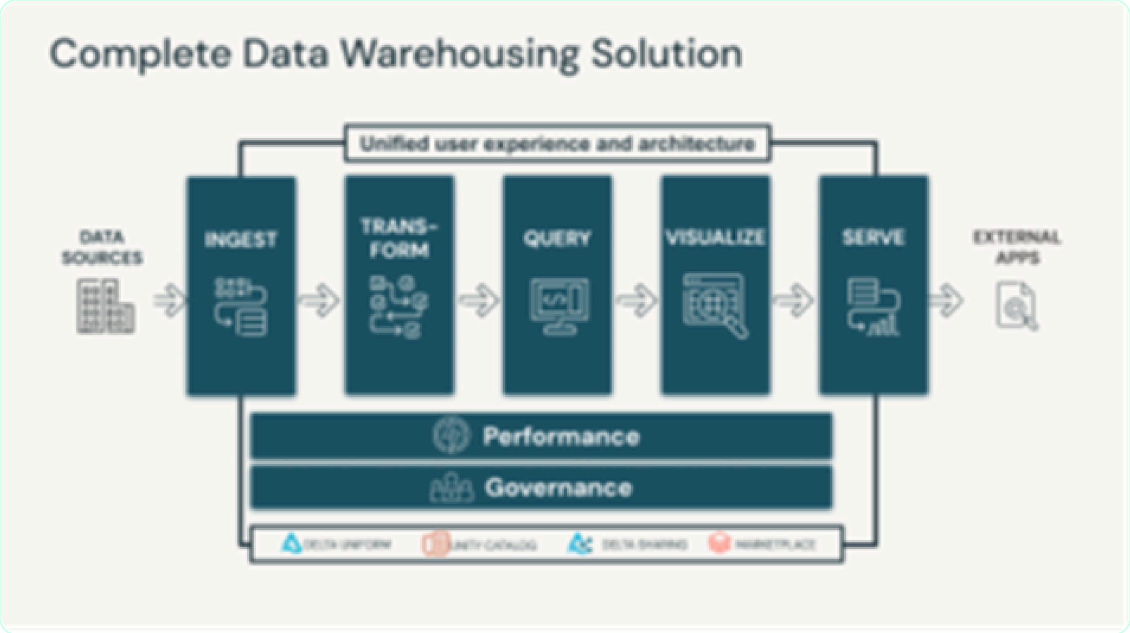

Image Credit: **Databricks**

Leveraging Databricks Unity Catalog for Data & AI Governance

The current data and AI platforms lack the intelligence to make connections between the data and the business concepts.

This has resulted in organizations relying on technical personnel to make sense of the data and provide insights.

This has resulted in a bottleneck effect that limits the use of the data and AI in the organization, especially among non-technical personnel.

To meet these important key governance issues, the **Databricks Data Intelligence Platform offers the open and unified governance solution for all data and AI assets**, called **Unity Catalog.**

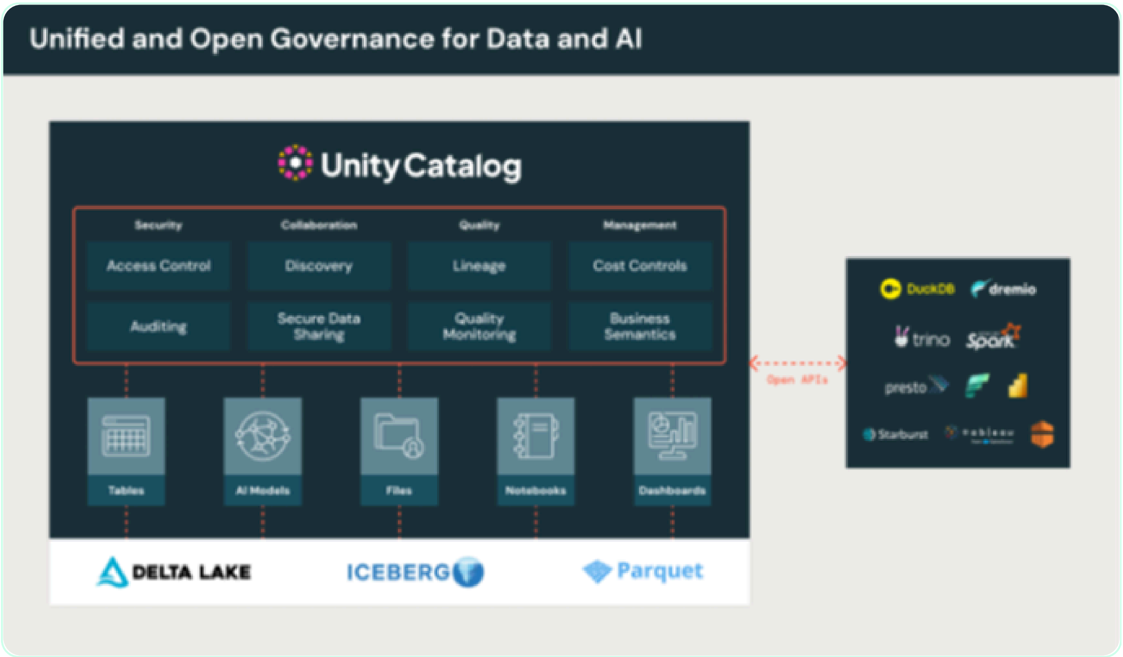

Image Credit: **Databricks**

As an open-source solution, Unity Catalog allows for the discovery and sharing of trusted data and AI assets across any tool, engine, or cloud.

Unity Catalog leverages the strength of the lakehouse and AI to offer domain-specific context and insights that increase productivity for both technical and business users.

Features of Unity Catalog

- Create and build an enterprise catalog for the curation of all structured and unstructured data, ML models, AI tools, notebooks, metrics.
- Use any open data formats of your choice, such as Delta, Iceberg, and Parquet
- Simplify security and compliance with a single interface for access control and auditing
- Scale and simplify governance with tag-based and attribute-based access controls
- Improve security with fine-grained access controls on rows and columns

- Analyze usage and cost with out-of-the-box observability dashboards.
- Tear down data silos between databases, data warehouses, and catalogs with integrated discovery, metadata management, and end-to-end lineage
- Share data and AI assets across data platforms, clouds, and regions without data replication

Data Ingestion Using Databricks

The Core Ingestion Challenge

The biggest problem for data engineers is how to move the data from various systems into a unified open lakehouse architecture.

Fragmented Data Sources and Ingestion Complexity

Contemporary businesses involve on-premises infrastructure, databases, data warehouses, and SaaS apps often proprietary and siloed. This creates a problem for data teams, who have to develop and manage intricate ingestion logic, which is prone to failure, introduces latency, and doesn't support analytics and machine learning.

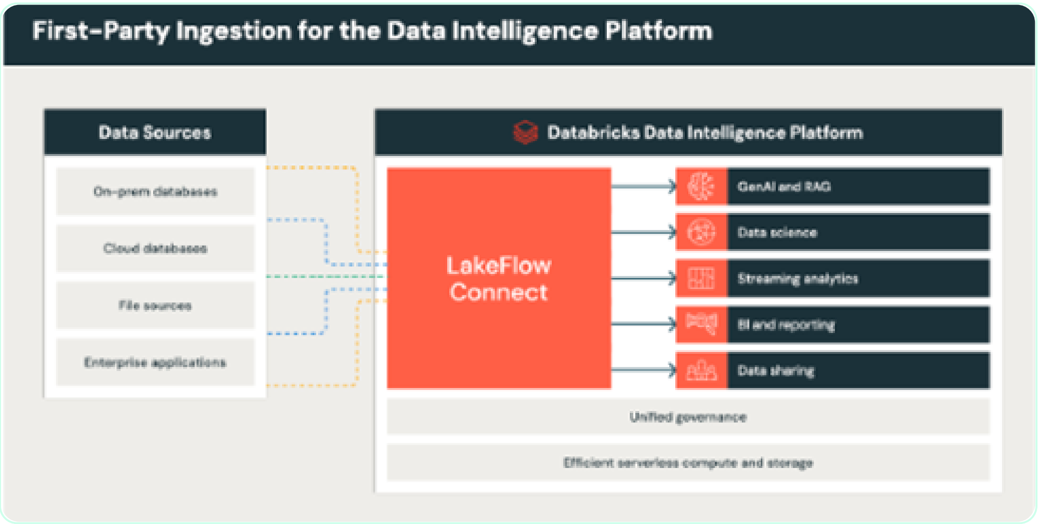

Image Credit: **Databricks**

Solution: Databricks LakeFlow Unified Data Engineering

Databricks LakeFlow is an integrated and intelligent data engineering platform that is native to the Databricks Data Intelligence Platform. It integrates data ingestion (LakeFlow Connect), development of pipelines (LakeFlow Pipelines), and orchestration (LakeFlow Jobs) into a single experience without the need for manual migrations or tool changes.

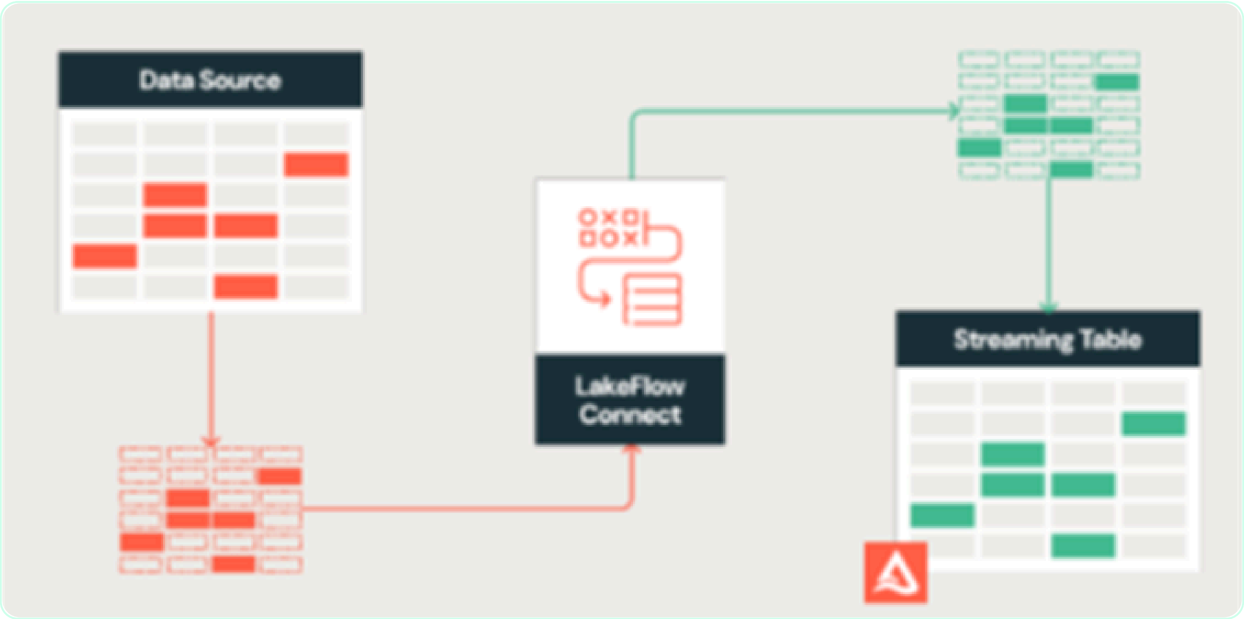

Image Credit: **Databricks**

Few Features and Services of LakeFlow

1 LakeFlow Connect

- LakeFlow Connect offers native connectors for common SaaS apps, relational databases, and file-based sources such as SFTP.
- These connectors allow for end-to-end incremental ingestion, easy configuration through UI or APIs, and native governance through Unity Catalog.

2 Auto-Loader

- For cloud storage-based ingestion, Databricks Auto Loader provides incremental file ingestion support for cloud storage solutions such as Amazon S3, Azure Data Lake Storage, and Google Cloud Storage.
- It is fully integrated with Structured Streaming and Delta Live Tables for schema inference, evolution, and high-throughput ingestion.

3 Optimized for Performance, Cost, and Reliability

- By integrating Auto Loader with Delta Live Tables, Databricks provides low-latency and scalable ingestion pipelines that automatically adjust to changes in the data.
- This not only saves operational costs but also ensures cost efficiency and performance consistency for batch and streaming data.

4 Broad Ecosystem of Ingestion Partners

- In addition to native consumption capabilities, Databricks provides a rich ecosystem of 500+ technology partners who offer pre-built connectors and native integrations.

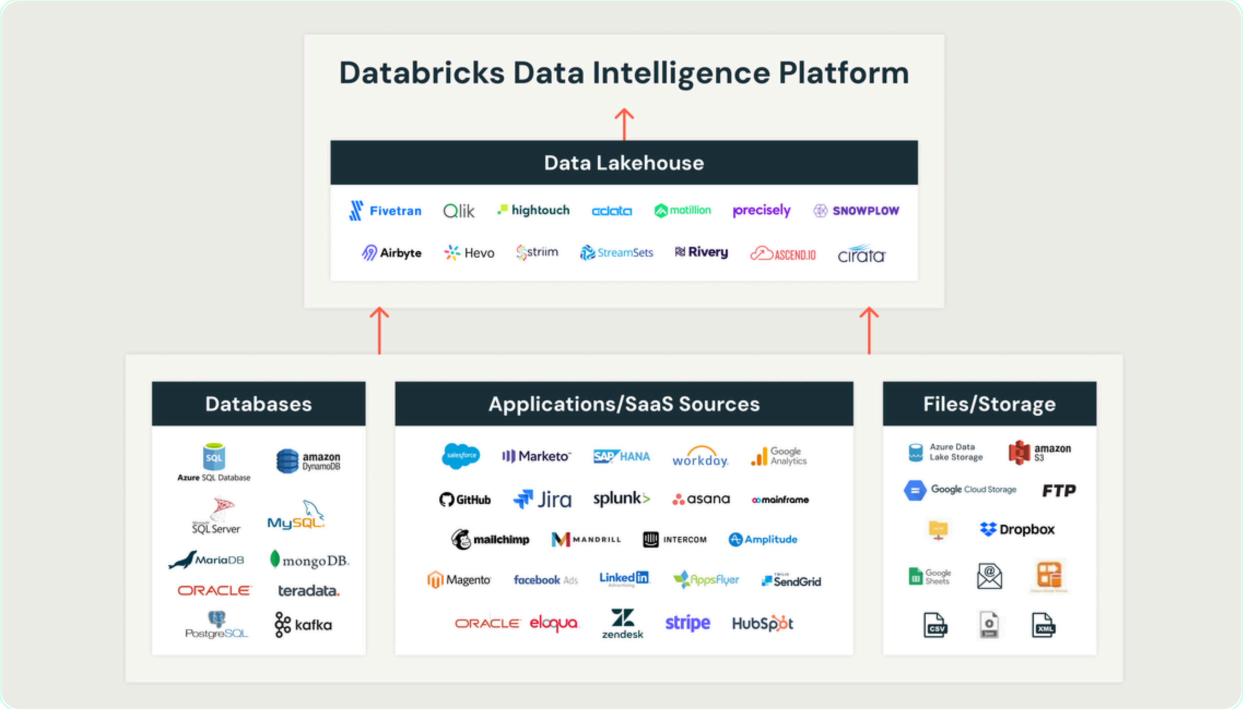

Image Credit: **Databricks**

Conclusion

- Modern data management faces fragmentation, scalability challenges, and limitations of traditional lakes and warehouses.
- The Databricks Data Intelligence Platform overcomes these issues with a unified, open Lakehouse architecture that manages the full data lifecycle, enabling trusted, scalable analytics and AI through centralized governance, multi-format interoperability, and streamlined data ingestion.
- This foundation sets the stage for deeper exploration in Part II, which will further examine advanced governance, optimization, and real-world implementation strategies using Databricks.

LumenData

Author

Ritesh Chidrewar

Technical Lead

About LumenData

LumenData is a leading provider of **Enterprise Data Management, Cloud & Analytics** solutions. We help businesses navigate their data visualization and analytics anxieties and enable them to accelerate their innovation journeys.

Founded in 2008, with locations in multiple countries, LumenData is privileged to serve over 100 leading companies. LumenData is **SOC2 certified** and has instituted extensive controls to protect client data, including adherence to GDPR and CCPA regulations.

Contact us:
info@lumendata.com | (855) 695-8636

Visit our website:
www.lumendata.com