

Modern Data Management with the **Databricks Data Intelligence Platform - II**



Data Transformation, Quality and Processing Using Databricks

Building on the [unified Lakehouse foundation introduced in Part I](#), let us dive deeper into how organizations can turn this architecture into sustained business value.

This section explores advanced governance capabilities, performance optimization techniques, and practical implementation patterns within the Databricks Data Intelligence Platform, highlighting how teams can operationalize trusted data, scale efficiently, and apply these principles in real-world scenarios.

✦ From Raw Data to Trusted Data Products

- First step is to ingest data into lakehouse. To provide value to analysts, data scientists, and AI systems, raw data must be processed into clean, reliable, and consumable data products.
- Poor data quality is the root cause of inaccurate insights, operational failures, and a loss of trust.
- Data engineers' responsibility is to clean up various high-volume datasets and transform them for analytics, business intelligence, machine learning, and GenAI applications.
- This involves a strong understanding of various data stacks and the management of complex transformation logic, which is prone to risks and delays.

Solution: Databricks Delta Live Tables: Declarative Data Pipelines

- Databricks Delta Live Tables (DLT) makes transformation and processing easier with a declarative ETL framework in SQL or Python.
- Engineers specify what the data should look like, and Delta Live Tables takes care of orchestration, infrastructure, monitoring, and error handling.



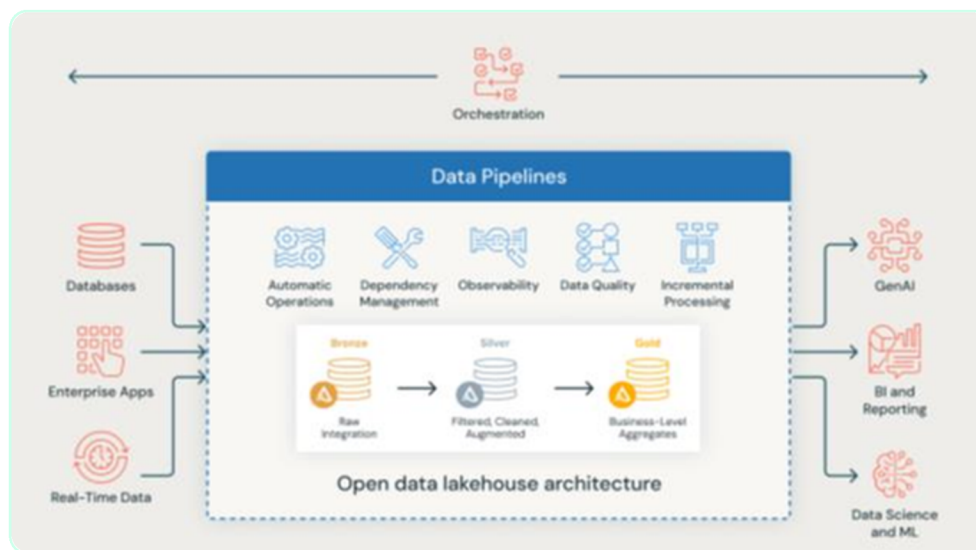


Image Credit: **Databricks**

Few Features & Services of DLT

1 Built-In Data Quality & Validation

- DLT allows teams to set up data quality expectations within the pipeline itself, making it possible to validate data in real-time as it comes in and is processed.
- Data quality rules are always on the lookout for the integrity of the data, ensuring that no erroneous data is passed along.

2 Incremental Processing at Scale

- Based on minimal code, DLT automatically determines the optimal execution plan to optimize performance and cost while enabling end-to-end incremental processing.
- With serverless compute, DLT enables incremental refresh of complex transformations for both batch and streaming workloads.

3 Workflow Orchestration with Databricks Workflows

- In addition to ETL, data platforms have to manage notebooks, queries, model training, dashboard updates, and scripts.
- Databricks Workflows offers a single solution for defining, executing, and monitoring multi-task workflows that encompass data engineering, analytics, and machine learning.

4 Adaptive & Self-Maintaining Pipelines

- Declarative pipelines developed using DLT are capable of adapting to changes in data, code, or execution environments.
- This makes it easier to maintain and allows data engineers to concentrate on developing, validating, and improving transformation logic.





✦ Traditional ETL vs Delta Live Tables (DLT)

Dimension	Traditional ETL	Delta Live Tables (DLT)
Pipeline Design	Imperative, step-by-step logic across multiple tools	Declarative pipelines in SQL or Python
Orchestration	External schedulers and custom dependency management	Built-in, automatic orchestration
Infrastructure Management	Manual cluster sizing and tuning	Fully managed, serverless execution
Data Quality Handling	Custom validation logic, often post-processing	Native data quality rules (EXPECT) within pipelines
Incremental Processing	Manually implemented and error-prone	End-to-end incremental processing by default
Schema Evolution	Manual handling and frequent breakages	Automatic schema inference and evolution
Error Handling & Recovery	Custom retry and alerting logic	Built-in monitoring, retries, and failure isolation
Operational Overhead	High maintenance and DevOps effort	Low operational effort, platform-managed
Scalability	Limited by pipeline design and infrastructure	Massively scalable with adaptive execution
Time to Value	Slower due to setup and maintenance	Faster pipeline delivery and iteration



Data Sharing & Collaboration using Databricks

- With the adoption of lakehouse architectures by organizations, the importance of trusted data goes beyond the boundaries of analytics and machine learning within organizations.
- Contemporary approaches to data must be aligned with business goals and include secure data sharing and collaboration.



✦ Challenges of Cross-Platform Data Sharing

- Fragmentation of data across cloud platforms makes data sharing more complex, resulting in data replication, storage costs, and privacy concerns.
- Conventional data sharing methods restrict AI innovation due to difficulties in sharing models, notebooks, and insights with appropriate governance.

Solution: Databricks Sharing Capabilities

- Databricks facilitates an open way of data sharing and collaboration, maximizing reach without sacrificing control.
- The platform is very interoperable and built to support collaboration across clouds, regions, and analytics tools with minimal operational overhead.



Few Features & Services of Databricks Sharing

1 Zero-Copy Sharing for Lower TCO

- With zero-copy data sharing, organizations can share data in such a way that there is only one copy of the data that can be accessed by multiple consumers without having to replicate the data.
- This approach cuts down the cost of duplication of storage and also enables teams to use their tools of choice while having control over the storage.

2 Built-In Sharing Capabilities of the Platform

- The Databricks Data Intelligence Platform is designed from the ground up for sharing and collaboration.
- Databricks Marketplace offers a platform for an open exchange of data, analytics, and AI assets, helping organizations discover and monetize trusted data products.

3 Databricks Clean Rooms

- Databricks Clean Rooms allow organizations to work with partners and customers in a secure and privacy-safe environment in any cloud.
- Data is protected while enabling joint analytics and insights in a way that does not require direct data access.

4 Delta Sharing: Open Protocol for Secure Data Exchange

- Delta Sharing enables both Marketplace and Clean Rooms with the first open protocol for secure data sharing in the industry.
- Delta Sharing allows data to be shared seamlessly across organizations and platforms without having to adopt particular compute engines.

All data sharing and collaboration workflows are centrally secured and managed by Unity Catalog. This helps ensure consistent access control, auditing, and compliance for shared datasets, AI assets, and collaboration environments.



Dimension	Traditional Data Sharing	Delta Sharing
Data Movement	Requires data replication or exports	Zero-copy sharing from a single source
Storage Cost	High due to duplicate data copies	Low, no additional storage required
Platform Dependency	Often tied to proprietary platforms	Open protocol, platform-agnostic
Cross-Cloud Support	Limited or complex to implement	Native cross-cloud and cross-region sharing
Security Model	File- or system-level access controls	Fine-grained, governed access via Unity Catalog
Data Freshness	Stale data due to batch replication	Near real-time access to live data
AI & Analytics Support	Data sharing only, limited AI assets	Share data, models, and analytics assets
Operational Overhead	High (pipelines, sync jobs, maintenance)	Minimal, policy-driven sharing
Ecosystem Reach	Restricted to same vendor or tools	Share with any tool supporting open formats

Key Takeaways & Conclusion

The fact is that today's modern businesses can't afford to think about data management as a back-office activity. **With analytics, AI, and generative AI increasingly at the forefront of business strategy, businesses need a unified, governed, and intelligent data platform that scales.** The **Databricks Data Intelligence Platform** helps solve the fundamental problems of fragmented architectures by integrating ingestion, transformation, governance, analytics, data sharing, and AI on an open lakehouse platform.

✦ Takeaways

- **Unified data management is essential for analytics and AI.** Fragmented lakes, warehouses, and governance platforms hinder scalability. Open formats and open governance minimize lock-in and complexity. **Compatibility with Delta Lake, Apache Iceberg™, and open APIs** provides flexibility and portability.
- Declarative pipelines and quality enable rapid data delivery. **Delta Live Tables simplify transformation, ensure data quality, and minimize operational complexity.**
- **Data sharing and collaboration are secure and scalable.** Zero-copy sharing with Delta Sharing and privacy-preserving collaboration **facilitate cross-cloud and cross-organization use cases.**





Author



Ritesh Chidrewar

Technical Lead

About LumenData

LumenData is a global enterprise data and analytics solutions provider specializing in master data management, cloud integrations, and AI-ready architectures.

Founded in 2008 and headquartered in Santa Clara, California, LumenData combines strategic consulting with deep technical expertise across platforms such as Salesforce, Informatica, MuleSoft, Snowflake, and Databricks.

With hundreds of certifications and over a decade long partnership with Informatica by Salesforce, LumenData helps organizations transform fragmented data into trusted, governed intelligence that powers personalized experiences, real-time automation, and strategic growth.

