TRANSFORMING DATA MANAGEMENT WITH DATABRICKS LAKEHOUSE FEDERATION

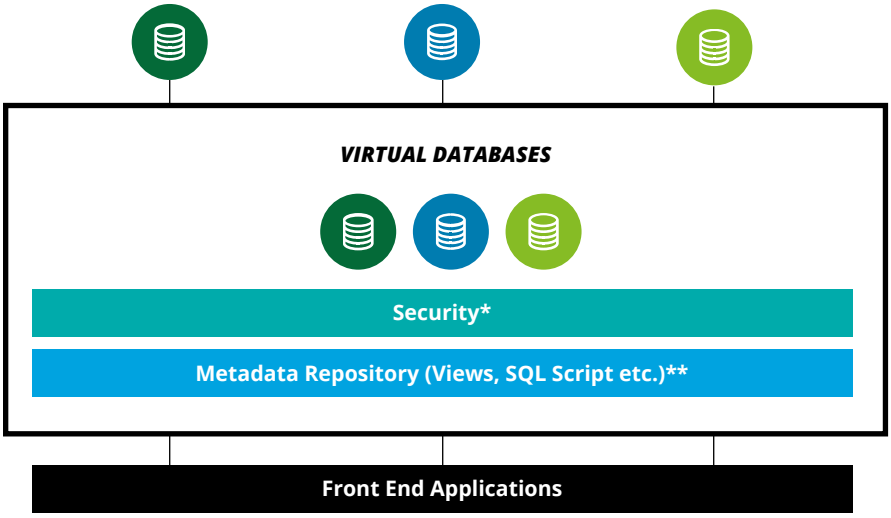
As data management challenges grow and evolve, organizations that embrace innovative yet trustworthy solutions can leverage real-time, accurate data to drive performance and competitive advantages.

The Databricks Lakehouse Federation is an advanced solution built on established data federation concepts. It plays a critical role in the Databricks Data Intelligence Platform, integrating into a Data and Analytics Platform (DAP) and offering extensive benefits to the Databricks ecosystem.

This paper explores the Databricks Lakehouse Federation's transformative power and how it leverages a data federation approach to optimize performance and deliver real-time data access.

WHAT IS A DATA FEDERATION?

Data federation, a form of data virtualization, integrates multiple sources into a unified entity. This virtual database standardizes diverse data into a single repository, ensuring centralized and consistent access for front-end applications.



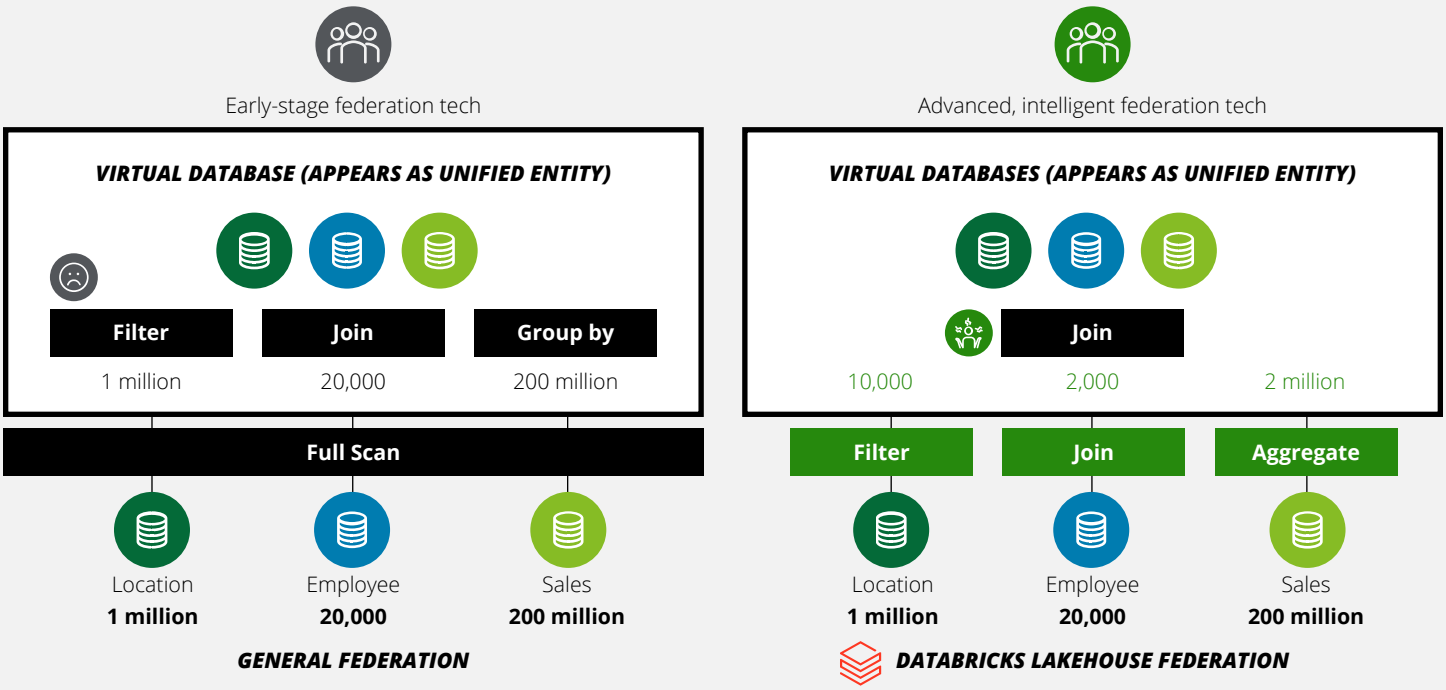
Data Federation - Illustrative

*Technology such as Databricks Lakehouse Federation offers security, access management, and governance controls through its Unity Catalog and provides the capability to define and manage the datasets.

3 physical sources appear as "One" database (virtual) for the users under the Unity Catalog

When leveraging data federation, the data remains within its source system, yet end users seamlessly gain access to a unified dataset using "Federated Views." Thanks to this seamless integration, end users may not even be aware of the precise location of the original data.

GENERATE A REPORT HIGHLIGHTING THE HIGHEST-PERFORMING SALES REPRESENTATIVE IN THE STATE OF ILLINOIS



In the early stages of federation technology, the approach involved consolidating large volumes of data (e.g., 1 million, 20,000, and 200 million records) into a centralized federation server layer. This method necessitated potent federation servers, which extended processing times, diminished overall performance, and significantly increased costs due to the need for enhanced computing and memory resources. This design was implemented to prevent overloading source systems, which are often engaged in critical business operations, as any additional overhead could negatively impact their performance.

However, advancements in intelligent federation technology have revolutionized this process. Modern federations now push relevant portions of processing back to the more powerful source systems, retrieving only the necessary data (e.g., 10,000, 2,000, and 2 million records). This approach ensures optimal performance, reduces costs, and efficiently uses computing and memory resources.

Concerns about overburdening source systems are managed by allowing the data to be brought into the federated servers if needed, keeping the process flexible for different operational demands.

DATABRICKS DATA FEDERATION

EXTEND DATA INTELLIGENCE ACROSS ALL SYSTEMS

Databricks Lakehouse Federation seamlessly integrates various data sources. It maintains metadata for user-friendly interfaces, stores federated queries for efficient retrieval, and emphasizes robust security and governance. Additionally, it dynamically scales compute resources to help ensure high quality performance.

Databricks Lakehouse Federation

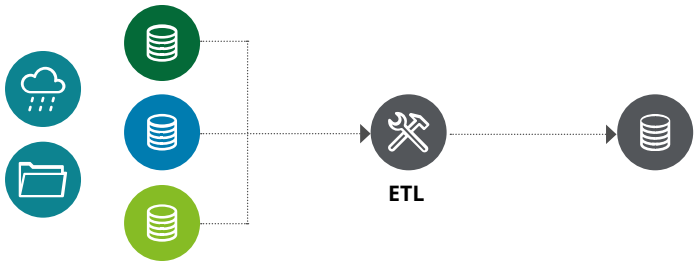
One significant challenge in data federation is managing the complexities associated with various SQL dialects. Databricks Lakehouse Federation effectively mitigates this issue by translating Databricks SQL into the corresponding SQL dialects of the source databases. This can help ensures compatibility and smooth integration.

Another major challenge in data federation is generating SQL queries optimized for source system push-down and execution within federation servers.

Databricks' data federation virtualization addresses this concern by balancing query optimization across both environments, enhancing performance and outcomes.

ETL/ELT VS FEDERATION

No data duplication and real-time access



In contrast to Batch and near real-time data transfer methods—where data is typically duplicated from the source to the target system—the Databricks federation approach keeps the data securely housed within the source system and optimizes responses with efficient data caching for subsequent queries, enhancing overall performance.

While useful, data federation is not a replacement for Extract, Transform, Load (ETL) and Extract, Load, Transform (ELT) processes in both streaming and batch data processing. Performing ETL or ELT in an enterprise data and analytics platform offers numerous advantages. This explanation distinguishes how ETL and ELT function compared to data federation. Most organizations operate their Data Analytics Platforms (DAP) primarily with ETL/ELT, complementing them with data federation as needed.

WHY DATABRICKS DATA FEDERATION?

Databricks Data Federation Services offer valuable advantages, including:

Unified data appearance
The Unity Catalog centralizes data objects, making it appear that they originate from a single source, even though they come from various sources. This integration ensures a unified data environment. Additionally, Databricks dynamically scales compute and memory resources based on demand, allowing clients to pay only for what they use and obtain insights without hardware scalability limitations.

Elimination of complex coding
Users are relieved from the burden of coding complex joins and cross-data source operations in the client application. This helps streamline development efforts and reduces the potential for errors.

Effective use of network bandwidth
Databricks Lakehouse Federation enhances network bandwidth by delegating queries to the source for execution. This can help reduces data transfer overhead and enhances performance by minimizing unnecessary data movement.

Standardization
Clients no longer need to handle diverse SQL dialects and data types, as the service abstracts away these complexities, providing a standardized interface.

Real-time data freshness
Queries are executed against backend data sources in near real-time, ensuring users can access the most current data available. This is particularly useful for applications requiring up-to-date information.

Security and access controls
Unity Catalog catalogs all source system objects, facilitating intelligent business definition generation, data lineage, and governance. It supports fine-grained security measures through Role-Based Access Control (RBAC), Attribute-Based Access Control (ABAC), and Tag-Based Access Controls, ensuring robust security and precise access management.

High-performance data processing and lower total cost of ownership (TCO)
Databricks federation technology enhances query optimization by harnessing parallel in-memory processing within high-performance photon-enabled spark clusters. It also implements efficient data and aggregate-aware caching, preserving aggregate results in caches. Additionally, Databricks Lakehouse Federation enables query rewriting, allowing queries to tap into pre-computed, broader cached data, expediting results retrieval without querying the source data directly.

Delta share and cleanrooms
Open cross-cloud, cross-platform data sharing without ETL enables secure data sharing across organizations without duplicating data. This facilitates collaboration across multiple business units and is essential for marketplace operations and data monetization. Unity Catalog supports secure data clean rooms, allowing two parties to collaborate without revealing each other's sensitive information, ensuring data privacy while enabling valuable insights and joint analysis.

AI/BI dashboards and Genie
As technology evolves, tech-savvy business users increasingly seek to create dashboards by leveraging Natural Language Processing (NLP) capabilities or fine-tuning NLP-generated queries, including data related to federated objects. Databricks artificial intelligence and business intelligence (AI/BI) tools are particularly useful in these scenarios. Additionally, other business users seek to understand business outcomes through NLP without technical expertise, using tools such as Genie. Cataloging all data objects, including metadata from federated sources, enables users to access previously unreachable data, ensuring that insights are accurate and pertinent.

LAKEHOUSE FEDERATION USE CASES

- Situations where a swift solution is imperative due to unavailable conventional data integration methods or to reduce risks when deploying a new solution without prior prototyping.
- When there is a pressing need for real-time or near real-time access to rapidly changing data, data federation allows direct access to the original data while carefully considering factors like performance, security, availability, and privacy.
- Whenever the business seeks to conduct exploratory data analysis across various sources of differing types—including files, services, applications, databases, and more—Federated Views emerge as a valuable means to address these challenges.
- When the expense of duplicating data surpasses the cost of remotely accessing it. It is essential to evaluate the impact on source system performance and the network expenses incurred while querying remote data sets in comparison to the costs related to storing numerous copies of the data, including network, storage, and maintenance costs.

In specific scenarios, opting for a federation-based approach becomes the preferred choice, such as when:

- The data volumes from the sources exceed the rationale for replication.
- The infrequent utilization of the data does not warrant its duplication.
- Only a minuscule or unpredictably small portion of the data sees any use.

- Establish a temporary virtual data mart or operational data store as a stopgap measure while actively developing a more permanent solution. Choose this approach only if you need to integrate additional data sources beyond the existing data warehouse. Before making this choice, architects should carefully assess the capabilities of the semantic layer to avoid redundancy and minimize any associated overhead.

It is important to note that Data federation is not a universal solution for every data integration challenge. Its suitability varies depending on the specific context and requirements of the data integration problem.

Data virtualization should be a complement to the data and analytics platform architecture domain, not a replacement. It works alongside conventional data integration techniques, providing a supplementary approach.

DATA FEDERATION MYTH VS. REALITY

Myth
The data federation server must be powerful enough to handle complex query capabilities and not be scaled.

Reality
Not anymore. Databricks clusters offer dynamic scalability to manage varying workloads efficiently, alleviating any concerns related to scalability. Additionally, its powerful Photon engine provides an extra layer of performance enhancement.

DATABRICKS LAKEHOUSE FEDERATION CONSIDERATIONS

The following general guidelines can help organizational leaders thoughtfully evaluate data federation options. Deloitte suggests seeking approval from the Architects before deviating from these considerations.

Scalable data and analytics platform

Leveraging the Scalable Data and Analytics Platform becomes essential when seeking a holistic, robust solution to address complex data management, in-depth analysis, and detailed reporting needs. This platform empowers your organization with its versatile capabilities and scalability.

Data governance and security

If you have strict data governance, compliance, and security requirements, a data and analytics platform offers better control and auditing capabilities than pushing all these access controls down to source systems. However, Databricks extends access control down to the source system level.

Demand for historical and trending data

Users often seek historical or trending data not readily available in operational data sources. Getting this requires a data consolidation approach that allows historical data to be collected over time. This is a common necessity in data warehousing.

Balancing performance and availability

Data access, performance, and availability are paramount concerns. Users frequently require swift data retrieval. Despite the effective coordination of well-designed federation services with remote data sources, the volume of data needed may require a locally pre-processed data copy. This scenario mirrors the dynamics often observed in lakehouse environments. Queries may be intricate or demand a multidimensional view of historical or trending data.

Predictable user needs

Predictable user needs, characterized by well-defined and repetitive queries involving access to a known subset of source data, may make generating a local data copy more cost-effective for direct access and utilization. This approach can also safeguard remote operational data sources from the burdens associated with large, intricate, or poorly structured queries.

Complex data transformations and joins

When intricate data transformations or lengthy, complicated joins are necessary, executing them synchronously as part of a user query is unwise. This is due to potential source system performance issues and excessive costs. In such cases, creating a data copy through data consolidation is the more advantageous approach.

Data stability and user experience

The data federation approach offers real-time access to remote data to meet the requirement of read-only access to reliably stable data. However, this real-time access may not always be ideal for end users or applications. They may prefer to accept some degree of data latency in exchange for protection from the constant influx of information from remote operational data sources.

Cost considerations

data federation can be cost-effective when data storage and replication costs are a concern. In contrast, a data and analytics platform may require more upfront investment but provide long-term cost savings.

Data federation offers users flexible access to data from any source instantly using real-time federated queries. However, this approach comes with potential considerations.

LAKEHOUSE FEDERATION VS STANDALONE

Evaluating the use of standalone data virtualization software

The decision to use standalone data virtualization software should be carefully considered in light of the significant benefits previously discussed.

The Databricks Intelligence Platform offers advantages beyond data federation, including the ability to govern all assets through a single pane of glass. However, if your data sources extend beyond those currently supported by Databricks Lakehouse Federation, complementing it with standalone data virtualization software can provide the best of both worlds.

This approach allows you to leverage Databricks Lakehouse Federation for compatible sources while using standalone solutions for unsupported sources as Databricks Lakehouse Federation's compatibility expands aggressively.

CONCLUSION

For organizations managing large-scale data, complex transformations, and advanced analytics, achieving real-time or near-real-time access to distributed data without extensive replication is a significant challenge.

The Databricks Lakehouse Federation effectively addresses these needs by providing a unified approach to data management. This allows access to data distributed across various sources, eliminating the need for extensive data replication and ensuring efficient, real-time data availability with simplified infrastructure, strong security, smart optimization, reduced complexity, and reduced cost while enhancing decision-making and agility for clients.

Connect with us!

GET IN TOUCH

Mani Kandasamy
Databricks Alliance CTO
Deloitte Consulting LLP
mkandasamy@deloitte.com

Jeff Lipkowitz
Partner Solutions Architects
Databricks
jeffrey.lipkowitz@databricks.com