STITCHING TOGETHER LOGICAL DATA FABRIC WITH DATA VIRTUALIZATION

Abstract

In the current knowledge era, the volume of data has increased tremendously but the insights extracted from the data has not increased proportionately, which leads to a knowledge gap with the unutilized data. This data can be in the cloud, on-premises, or in mainframe computers. The data can exist in various formats, such as relational, non – relational, and NO-SQL databases. The biggest challenge most organizations face is to close the gap between potential derived value and current derived values from data.

denodo

Infosys®
Navigate your next

To manage the complexity of today's ecosystems, organizations are adopting newer architectural approaches such as logical data fabric, which helps organizations to manage data across hybrid/distributed data ecosystems and unifies disparate data in an intelligent way. This modern data management approach can optimize access to distributed data and intelligently curate and orchestrate it for self-service delivery to data consumers. With a logical data fabric, organizations can elevate the value of their enterprise data by providing users with access to the right data, just in time, regardless of where and in what format it is stored.
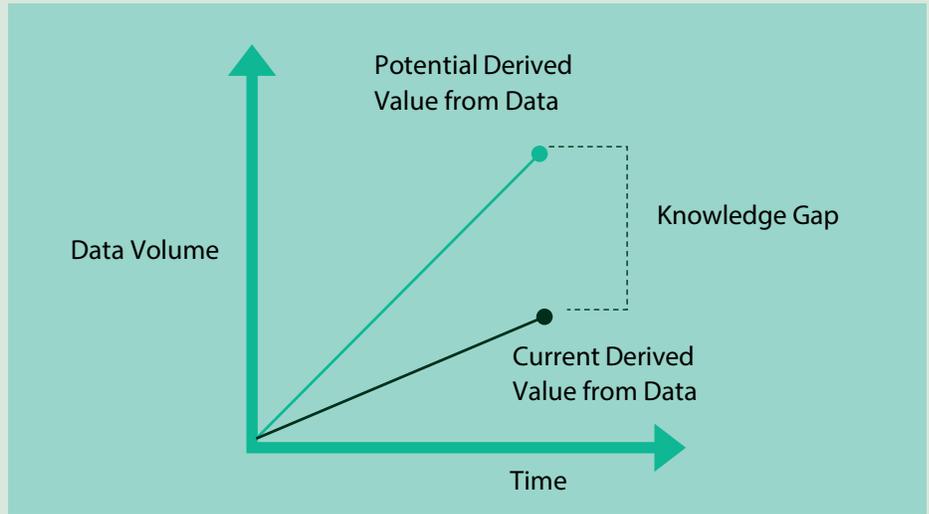

Figure 1: The knowledge gap

What is Data Fabric?

Data fabric is an architectural pattern that dynamically orchestrates disparate data across a hybrid and multi-cloud or on-premises data landscape to securely provide business-ready trusted data that supports applications, analytics, and business process automation. Data fabric is designed to help organizations to solve complex data problems and use cases by managing their data — regardless of where the data is stored and in what format it is stored.

Data fabric can be visualized as a fabric or a net, spread across the globe, wherever the organization's users are present. The user can be at any place in this fabric and still access data at any other location without any constraints, in real-time. In simple terms, data fabric:

- Provides a unified, consistent data view and access to data for any member of an organization worldwide and in real-time

- Integrates data from disparate data sources

- Securely delivers an integrated view of different data objects

- Automates the entire process using artificial intelligence/machine learning (AI/ML) capabilities
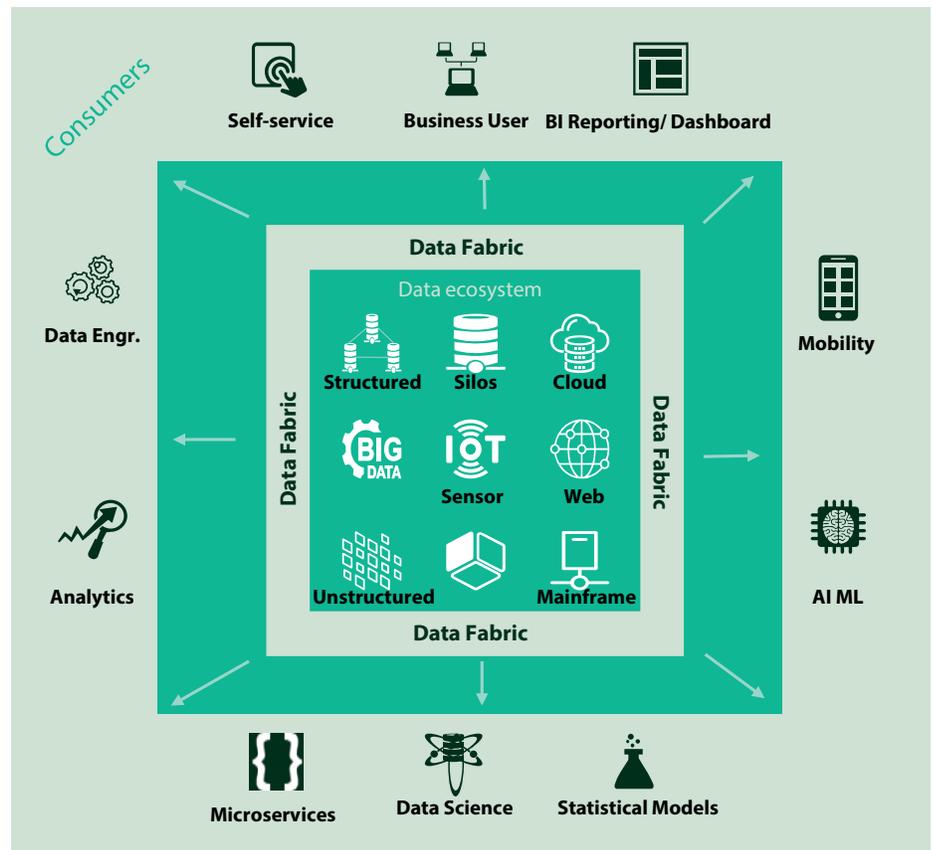

Figure 2: Reference Architecture – Data Fabric

What is Logical Data Fabric?

Logical data fabric is a vision of a unified data delivery platform that abstracts access to multiple data systems by hiding complexity and presenting data in business-friendly formats. A logical data fabric uses data virtualization, which avoids the traditional approach of physically replicating and integrating data. Logical data fabric helps organizations to:

- Minimize data movement and data replication, which can be time-consuming and expensive, and this can also pose security and compliance risks

- Virtually integrate, manage, and govern enterprise data across on-premises and cloud systems for insight generation and business decision making

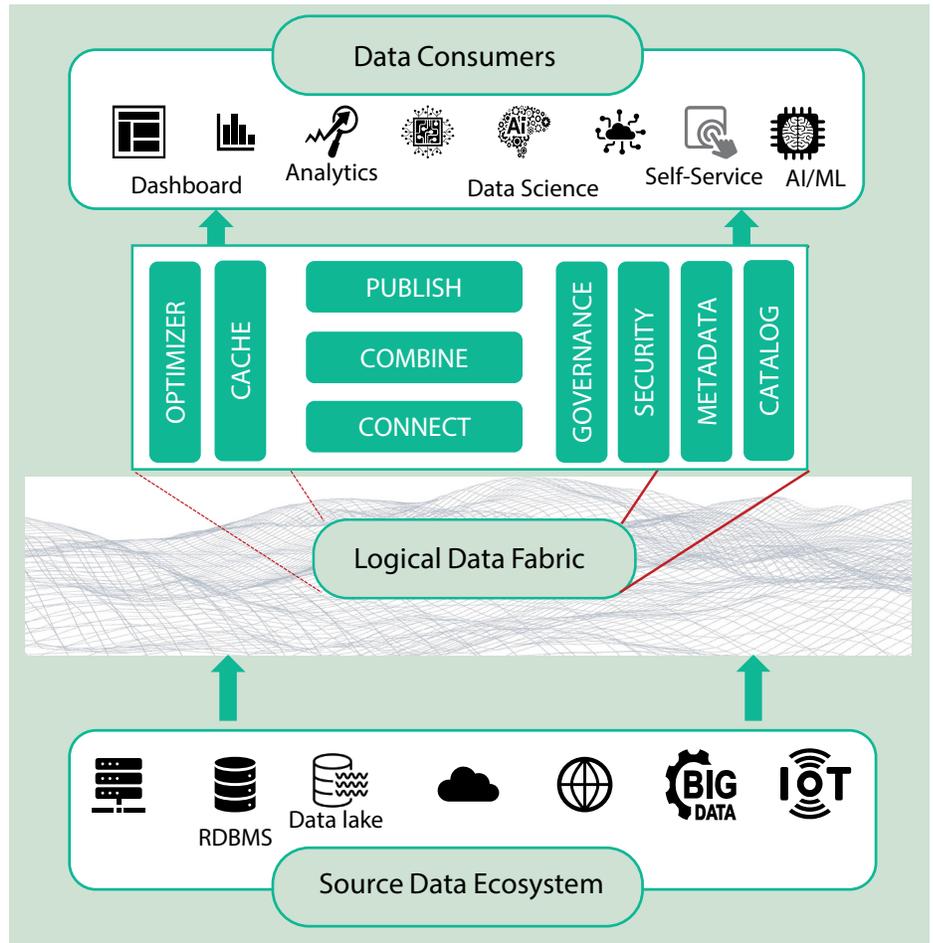- Establishes future-proof data architecture that can meet new demands

Figure 3: Reference Architecture – Logical Data Fabric

Data Fabric Vs Logical Data Fabric

Factors	Data fabric	Logical data fabric
Approach	Physically moves the data from various on-premises and cloud sources using a standard ETL (extract, transform, load) process	Connects to the different on-premises and cloud sources, integrates only the metadata required, and creates a virtual data layer. This enables users to leverage the source data in real time.
Time to solution	There is a need to persist data for Analytical use cases or MDM needs.	Enterprise needs real-time data delivery without physically replicating data.
Data cleansing	Data quality is very poor, and data requires extensive, complex cleansing and data transformation.	Data is in good shape, but enterprise may need to perform basic validation checks and reformatting tasks.
Real-Time Data	Business users need only historical data for analysis and reporting.	Best fit for real-time data needs.
Limitations	Cannot be used as an enterprise data layer.	Cannot be used when data needs heavy transformation.

Table 1: Comparison between data fabric and logical data fabric

Why Logical Data Fabric

All top companies are data-driven, and they rely on insights for business decision-making. There are some barriers to leverage data to its full potential, some of them are:

• Heterogeneous data ecosystems

• Data volume growth

• The lack of access to source data systems

• Multiple data formats and data types

• Data integration complexities

Data-driven businesses demand real-time data connectivity, self-service capabilities, automation, and universal transformations, and these needs cannot be met by traditional data integration techniques.

Organizations today are by building data warehouses, data marts, or data lakes for data storage. Here are the challenges posed by data warehouses and data Lakes:

Data warehousing	Can solve integration challenges for much of our data but never actually integrate all the diverse data.
	Inflexible, and they do not offer the required agility to respond to business users.
	Applying AI/ML is a more data demanding task than most data warehouses can cope with, both in terms of complexity and data volume.
Data Lakes	A good option to process relatively large data amount of data inexpensively.
	Can hold structured and unstructured data, but it can be difficult to actually find and integrate different data sets as lakes continue to grow.
	They present more data management issues if data lakes become data swamps.

Table 2: Current data management Pros and Cons

These solutions were not designed with an ever-evolving data landscape in mind, and that makes it difficult to leverage them for a unified view of the data. Also, these techniques often lead to latencies and increasing cost. With the growing amount of data and time-constraints, delays in data access and processing cannot be afforded.

In such scenarios, a logical data fabric gives the advantage of extracting and processing data at the source point in real-time, enabling decision-makers to gain insight on the go. Some of the many benefits of logical data fabric are listed below:

Logical Data Fabric	Provides a single unified data-access point for all data: structured, unstructured data that is spread across the file system or traditional SQL/No-SQL databases
	Accelerates data delivery without the need to duplicate data physically, while enabling streamlined data discovery
	Increases flexibility and self-service, enabling multiple consumption patterns once data has been provisioned
	Provides data discovery through a data catalog.

Table 3: Benefits of logical data fabric

Key Capabilities of Logical Data Fabric

Data virtualization, which is a core component of logical data fabric, plays an important role in supporting the need to access, manage, and analyze data across disparate platforms for traditional reporting and BI — as well as modern use cases such as machine learning and artificial intelligence, integrated analytics for automated decision making, and analysis, and combining traditional data-at-rest with real-time streaming data sources.

An enterprise grade logical data fabric should offer these critical capabilities -

Integrate data across on-premises and multi-cloud environments

- Transparently access and integrate data from different on-premises or cloud-based sources

- Aggregate data from across different cloud environments into a logical view

Automate manual tasks using augmented intelligence

- Use machine learning to continually monitor data sources and track changes to data structures to automatically adapt semantic/virtualized models

- Enable seamless access despite changes to the underlying source(s)

- Provide data analysts with automated recommendations

- Automate data caching to improve and maintain high performance

Boost performance of analytics with rapid data delivery

- Leverage a dynamic query optimization technique to partially push computation down to the appropriate source system

- Automate caching local copies of frequently accessed data

- Provide intelligent query federation

Support data discovery and data science initiatives

- Enable data scientists to gain an enterprise data awareness

- Provide self-service data access to authorized enterprise data stakeholders

- Deliver enterprise data sets to different data science projects through a variety of channels such as a BI front-end, through APIs, or even using notebooks

Catalog all data for discovery, lineage, and associations

- Provide comprehensive data catalogs across the entire enterprise data infrastructure

- Capture the full lineage of any enterprise dataset as well as all applicable associations

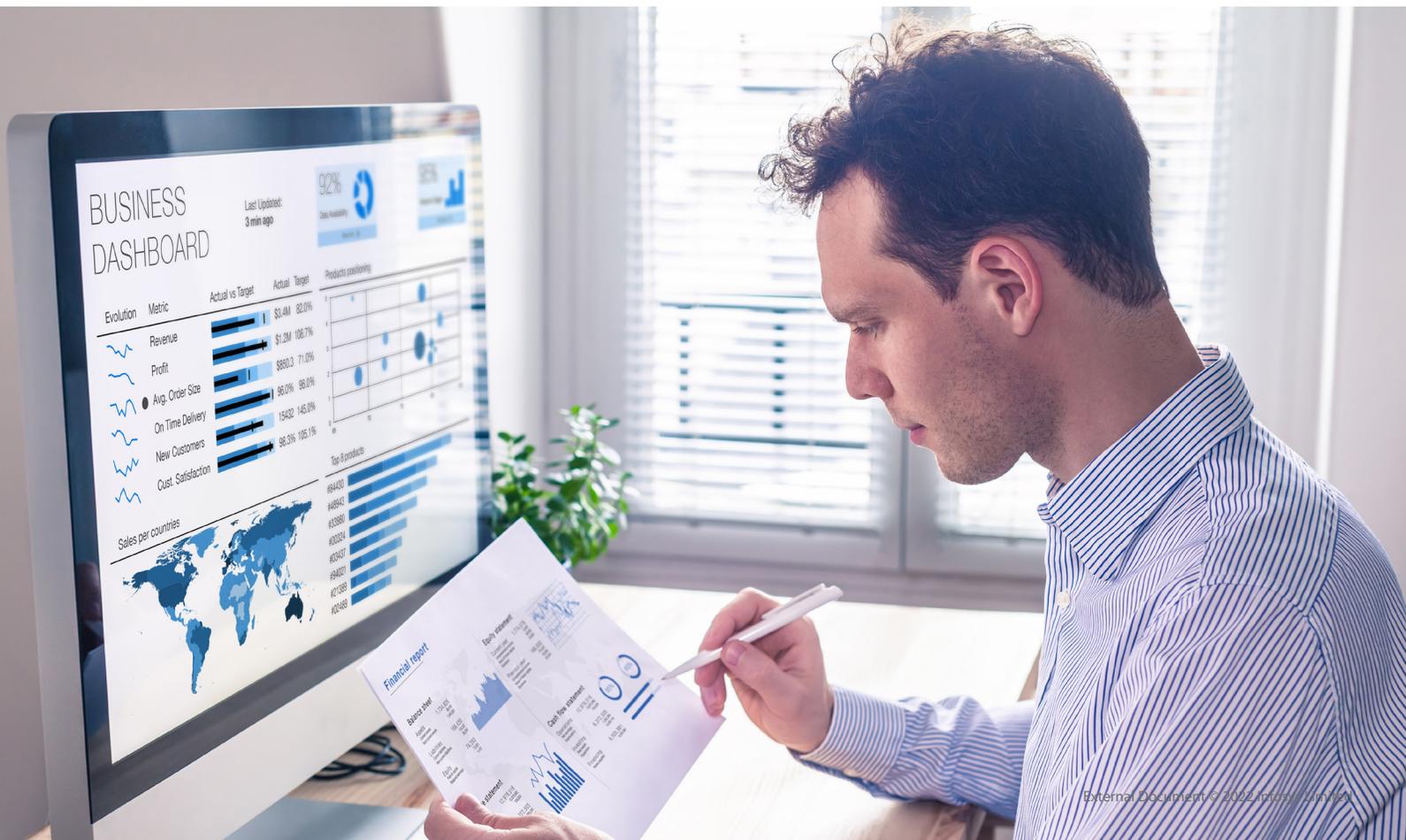# The Business Benefits of Logical Data Fabric

At Infosys, we have provided business benefits to our clients by implementing data virtualization across a variety of use cases. Below are a few case studies that showcase our recent data virtualization implementation:

Facilitate Integrated Data

Objective	A large US Based investment bank was on-boarding another mid-sized Investment bank business as part of a merger & acquisition to gain new capabilities, eliminate competitors, and expand its market reach.
Problem Statement	• During the transition, business users still need access to all of the company's global products data in a consistent and timely manner for all regulatory compliance, financial, and management reporting activities. • Analysts and data scientists need unified access to data for critical use cases like AML, fraud, etc. • Access to data needs to be seamless; existing, manual, error-prone, time-consuming processes need to be eliminated.
Solution	To accelerate the transition, we leveraged the Denodo Platform, which serves as a data virtualization layer above the acquired bank's systems and its own core internal systems. • Created a custom-built semantic layer that enables self-service access to data (logical data fabric): • Abstracts underlying technology, data models, and geographic complexities of data sources from consumers • Provides a single point of access to data entities to enable data-as-a-service or a subscribe-deliver model. Used data integration techniques such as data virtualization and data federation. • Access discrete data formats from sources, process user defined transformations, and prepare customizable o/p formats using data wrangling/munching • Outbound data delivered through channels like files and messages streams, scheduled by days, business calendars, business events
Metric Wall	• Integrate with 65 global data sources • Expanded threshold of outbound data to approximately 600 GB s per day, through channels like message streams, files scheduled by time, business calendar, and business events • 3,500+ events/day for subscribing data • 55 consumption models • Supports Batch, micro-batch, and real-time data delivery
Business outcomes	• $300M savings through automation • 200X improvement in time-to-market • 100X improvement in data consistency • 80X reduction in time for application integration • 150X improvement in data processing and ingestion

Facilitate Federated Data -

Objective	The client was a non-profit mutual benefit corporation and health plan founded by the California Medical Association. It is based in California and serves 4.5 million health plan members and more than 65,000 physicians across the state. The client was on the path to improving the current Find a Doctor Portal with a key focus on achieving high performance and fixing a few data integrity/inconsistency issues.
Problem Statement	The client was facing data challenges such as: • Fragmented user experiences due to inaccurate, poor data quality • A lag in data retrieval due to complex ETL processes • Mobile and Web used different channels to retrieve same data, resulting different experiences
Solution	Infosys fully leveraged an emerging, nimble solution based on data federation and data virtualization using the Denodo Platform. • Created a virtual canonical logical layer integrating disparate source systems to satisfy "Find a Doctor/ Wise choice Portal" needs. • Performed quick consolidation from various systems to check requirement feasibility and certified data quality. • To address performance challenges, cached data and created cache indexes at various levels to achieve high search query performance.
Metric Wall	• Achieved search query response SLA of 2 seconds
Business outcomes	• With data virtualization, we helped client to meet its SLA of a high-performant "Find A Doctor" portal. • Data virtualization made finding a doctor, and the work of medical groups, vision care providers, specialists, hospitals, pharmacies, urgent care centers, and providers, faster and easier.

Conclusion

Logical data fabric is gaining more popularity in stitching all the diverse data sources together with the help of data virtualization and providing a unified view. Physically replicating data in a common repository will not only increase storage costs but will also consume unnecessary amounts of time. In the meantime, data gets stale or out-of-sync, and such issues need to be addressed before business users can consume it further. Logical data fabric is an appropriate solution that provides real-time access to the source data much faster and easier than traditional approaches.

References

https://www.denodo.com/en/partner/infosys

https://www.denodo.com/

About the Author

Nilesh Bhatti
Senior Technology Architect Data & Analytics Practice

For more information, contact askus@infosys.com

Infosys.com | NYSE: INFY

Stay Connected