Get more value from your data

Data Fabric

Smart Data Engineering, Operations, and Orchestration

Dave Wells

September 2019

Research Sponsored by

Infoworks

About the Author

Dave Wells is the Data Management Practice Director at Eckerson Group, a data analytics research and consulting organization. He is an internationally recognized thought leader in data management, a frequent speaker at industry conferences, and a contributing author to industry publications. Dave brings a unique perspective to data management based on five decades of working with data in both technical and business roles. He works at the intersection of information management and business management, where real value is derived from data assets.

Dave is an industry analyst, consultant, and educator dedicated to building meaningful and enduring connections throughout the path from data to business value. Much of his work today is focused on modernization—updating turn-of-the-century BI architecture to optimize for big data and analytics, sustaining the value of legacy data warehouses with cloud technology and data lake compatibility, and rethinking data governance to work well in self-service and agile cultures.

About Eckerson Group

Eckerson Group helps organizations get more value from data and analytics. Our experts each have more than 25+ years of experience in the field. Data and analytics is all we do, and we're good at it! Our goal is to provide organizations with a cocoon of support on their data journeys. We do this through online content (thought leadership), expert onsite assistance (full-service consulting), and 30+ courses on data and analytics topics (educational workshops).

Get more value from your data. Put an expert on your side. Learn what Eckerson Group can do for you!

About This Report

To conduct research for this report, Eckerson Group interviewed numerous industry experts and viewed a dozen or more demonstrations of data fabric technologies. The report is sponsored by Infoworks and Hitachi who have exclusive permission to syndicate its content.

Table of Contents

Executive Summary

Data fabric is a combination of architecture and technology that is designed to streamline the complexities of managing many different kinds of data, using multiple database management system, and deployed across a variety of platforms. A typical data management organization today has data deployed in on-premises data centers and multiple cloud environments. They have data in flat files, tagged files, relational databases, document stores, graph databases, and more. Processing spans technologies from batch ETL to changed data capture, stream processing, and complex event processing. The variety of tools, technologies, platforms, and data types make it difficult to manage processing, access, security, and integration. Data fabric provides a consolidated data management platform. It is a single platform to manage disparate data and divergent technologies deployed across multiple data centers, both cloud and on-premises.

The complexities of modern data management expand rapidly as new technologies, new kinds of data, and new platforms are introduced. As data becomes increasingly distributed across in-house and cloud deployments, the work of moving, storing, protecting, and accessing data becomes fragmented with different practices depending on data locations and technologies. Changing and bolstering data management methods with each technological shift is difficult and disruptive, and will quickly become unsustainable as technology innovation accelerates. Data fabric can serve to minimize disruption by creating a highly adaptable data management environment that can quickly be adjusted as technology evolves.

Key Takeaways

- Data fabric is an emerging solution to the complexities of modern data management. It combines architecture, technology, and services to automate much of data engineering, operations, and orchestration.
- Almost everyone today operates multi-cloud and cloud-hybrid systems. Managing across these systems needs a single, unified data management platform.
- Data fabric provides a single, unified platform for data management across multiple technologies and deployment platforms.
- No single vendor provides a complete data fabric solution today. Choose the right technologies to weave your data fabric. Interoperability is a key consideration.

Recommendations

- Don’t ask if you need data fabric. Ask when you’ll need data fabric.
- Make the case for data fabric in three dimensions—business case, technical case, and operational case.
- Identify your data fabric use cases. They may include managing complex data systems, modernizing data management architecture, migrating to cloud, moving to DataOps, and more.
- Leverage existing technology when weaving your data fabric, but don’t be anchored by it.
- Don’t forget the human side of data fabric. Data management has many stakeholders and you’ll need to engage them all.

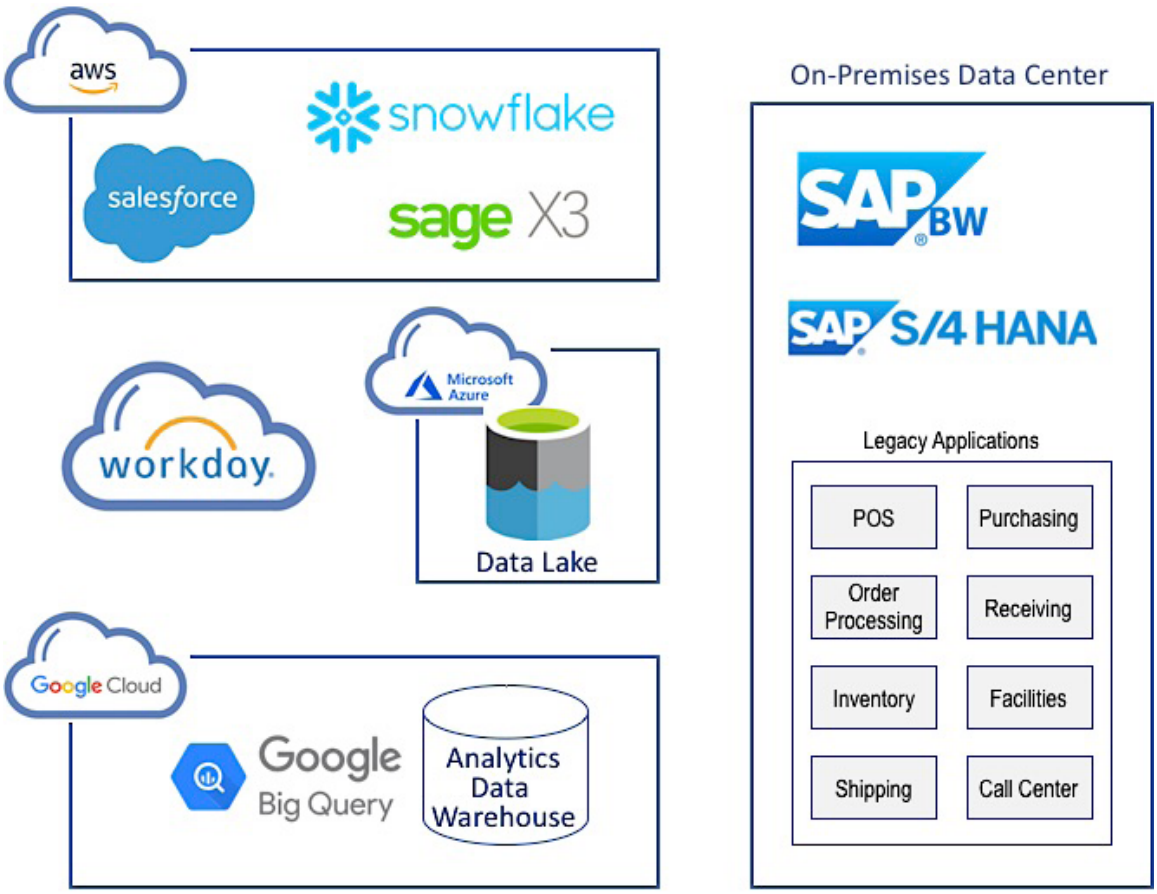
The Trouble with Data Management

Data management has become increasingly complex over recent years as the variety of data types, databases, deployment platforms, and data use cases expands. Today's data management challenges include data silos, data engineering bottlenecks, operationalization difficulties, and orchestration of data systems in runtime environments.

Data Silos

In the age of data-driven business, data is everywhere. That is a good thing for data-hungry processes and analytics but a challenge for data management. When data is siloed across multiple cloud platforms and also stored in on-premises databases it becomes difficult to find, blend, and integrate when needed. The complex deployment landscape shown in figure 1 illustrates typical deployments today. This landscape has four separate cloud environments as well as several systems operated in an on-premises data center.

Figure 1. Data Silos across the Ecosystem



These kinds of data deployments can't be avoided today. With legacy systems, SaaS applications, data warehouses, and data lakes data is spread widely across platforms and technologies. Although data is abundant it is isolated and difficult to find, access, and integrate. One goal of data fabric is to "connect the dots" across data silos.

Data Engineering

Data engineering is a critically important part of analytics that receives little attention compared to data science. Recent research shows 12 times as many unfilled data engineer jobs as data scientist positions. Breadth and depth of required skills limits the number of qualified people to work as data engineers. Clearly the demand for data engineers outstrips the supply, and the gap continues to grow. The large number of unfilled jobs reflects the complexity of data engineering. Breadth of knowledge ranges from relational databases to NoSQL, from batch ETL to data stream processing, and from traditional data warehousing to data lakes. Depth of skills includes hands-on work with Hadoop; programming in Java, Python, R, Scala, or other languages; and data modeling from relational and star-schema to document stores and graph databases. The data engineer is part database engineer (building the databases that implement data warehouses, data lakes, and analytic sandboxes) and part software engineer (building the processes, pipelines, and services that move data through the ecosystem and make it accessible to data consumers). One goal of data fabric is to automate much of data engineering to increase reuse and repeatability, and to expand data engineering capacity.

Data Operations

Rapid, reliable, and repeatable delivery of production-ready data for reporting, analytics, and data science is an ongoing challenge. Operationalizing data pipelines is difficult. When data is spread across multiple platforms, pipeline processing must be able to span on-premises, cloud, multi-cloud, and hybrid environments. Sustaining data pipelines is equally challenging as business needs, data sources, and technologies continuously change. Fault tolerance is critical to data operations. Entire analytics supply chains are disrupted when a data pipeline fails, and repairs are especially slow and difficult when everything is done manually. Data operations must also be attentive to data protection, data governance, data lineage, metadata, and auditability.

DataOps is not practical without automation.

The emerging practice of DataOps holds promise, but it is a big shift. DataOps is a data management approach that is designed for rapid, reliable, and repeatable delivery of production-ready data and fully operational analytics. DataOps is not practical without automation. Robust DataOps technology offers features and functions for model orchestration, data pipeline orchestration, test automation, and deployment automation for

data pipelines and analytic models. One goal of data fabric is to fully support the automation needed for DataOps success, with ability to automate across on-premises, cloud, and hybrid data ecosystems.

Data Orchestration

Execution environments have many of the same challenges as data environments. Few organizations today have a single execution environment. Pushing processing to data locations results in on-premises, cloud, multi-cloud, hybrid, and edge environments for runtime processing of data. Separating computation from data and scaling each independently is fundamental to operate this extreme of distributed and parallel processing. End-to-end data pipelines often span multiple execution environments. Managing data access and processing across these complex environments requires attention to configuration and coordination, workflow and scheduling, cross-platform interoperability, fault tolerance, and performance optimization. Data orchestration is a multi-faceted and complex job that can't be done without automation. One goal of data fabric is to support automation across the many dimensions of data orchestration.

What is Data Fabric?

Data Fabric Defined

Data fabric is a combination of architecture, technology, and services designed to ease the complexities of managing different kinds of data, using multiple database management systems, and deployed across a variety of platforms. It provides a single, unified platform for data management across multiple technologies and deployment platforms.

Data fabric provides a single, unified platform for data management across multiple technologies and deployment platforms.

This simple definition captures the common concepts included in a variety of vendor definitions, each defining data fabric from the perspective of their products and contributions. Diversity of vendor solutions is a good thing that also brings diversity in the definitions. Some nuggets among the vendor definitions include the following:

- An architecture and set of data services that provide consistent capabilities across a choice of endpoints spanning on-premises and multiple cloud environments

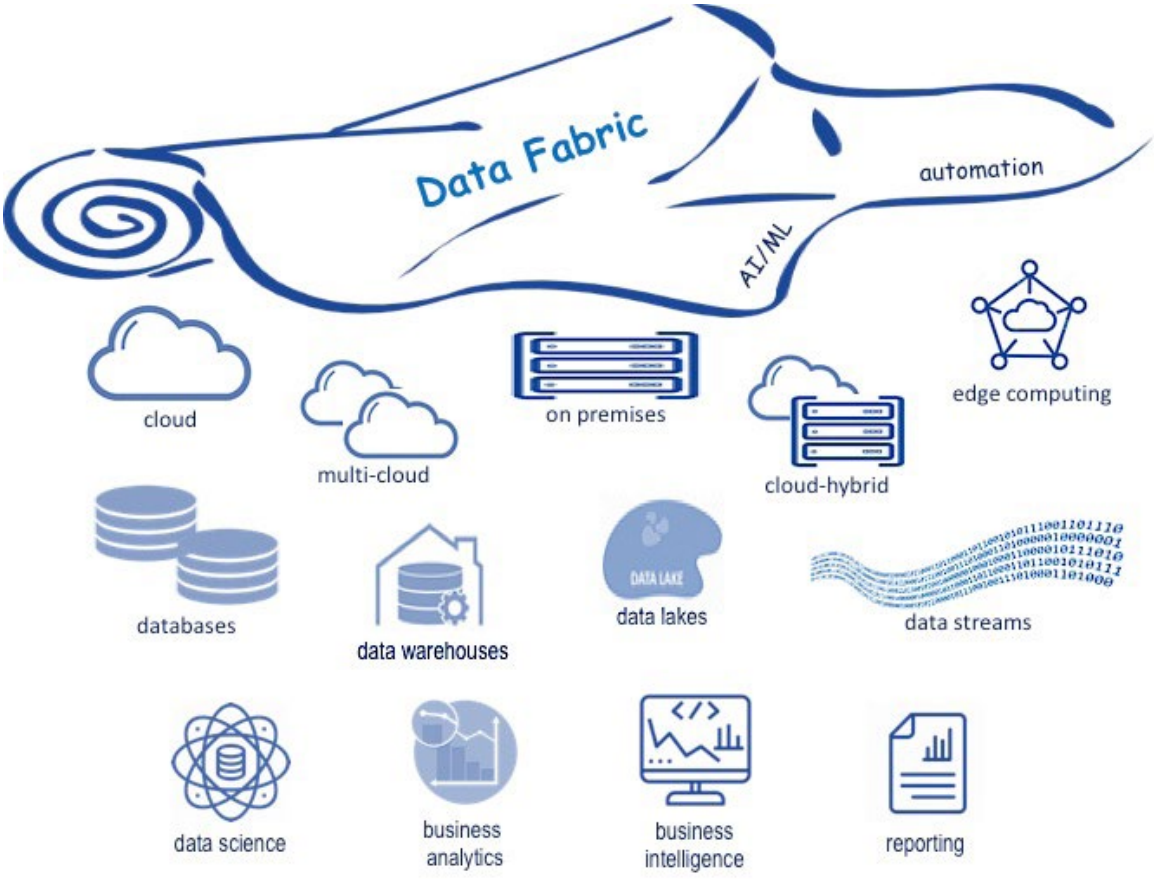
- A new way to manage and integrate data that promises to unlock the power of data in ways that shatter the limits of previous generations of technology such as data warehouses and data lakes
- Based on a graph data model … able to absorb, integrate, and maintain the freshness of vast quantities of data in any number of formats
- Automates and accelerates the data engineering, operationalization and ongoing management of BI, AI and machine learning workflows from source to consumption, for cloud, on-premise and hybrid environments
- Simplifies and integrates data management across cloud and on-premises to accelerate digital transformation
- A converged platform supporting the diverse data management needs to deliver the right IT service levels across all disparate data sources and infrastructure types
- A converged platform that supports the storage, processing, analysis and management of disparate data
- A system that provides seamless, real-time integration and access across the multiple data silos

Each of these items are quoted directly from software vendors offering data fabric solutions. In the interest of remaining vendor-neutral, specific vendor attribution is omitted here.

Data Fabric Concepts and Principles

Fabric is an apt metaphor for solutions to modern data management challenges. The concept of fabric applies in two ways. First, the scope of data, processing, platforms, and execution environments is so broad that a single tool can't provide comprehensive management. Fabric can be "rolled out" to cover current scope and continuously extended as the scope expands. (See figure 2.) The data fabric acts as a canopy covering the data management environment from end to end. Artificial intelligence and machine learning (AI/ML) provide the intelligence of smart data fabric, and automation reduces manual effort and accelerates the processes of data management.

Figure 2. Data Fabric Covers the Broad Scope of Data Management



The fabric metaphor also applies from the perspective that fabrics are woven from fibers. With data fabric we seek to weave data management into the activities and workflow of everyday business processes—to operationalize and orchestrate in a way that makes data an integral part of doing business.

Beyond metaphor and concept, the following principles are fundamental to data fabric:

- Ability to find, access, and combine data from all sources regardless of type and location.
- Ability to work with data of all types—structured, semi-structured, multi-structured, and unstructured.
- Support for the speed, scale, and reliability needed for enterprise-grade data systems.
- Ability to process across many execution environments including multiple data centers, multiple cloud platforms, and systems at the edge of the network.

- Ability to process and provision data at all velocities from streaming to batch.
- Support for multiple processing engines including Hadoop, Spark, Samza, Flink, and Storm.
- Ability to adapt to new processing engines and tools as technology evolves.
- Ability to move data from one platform to another without extensive refactoring.
- Ability to move processing from one execution environment to another without extensive recoding.

Why a Data Fabric?

Data Management Complexities

The complexities of modern data management expand rapidly as new technologies, new kinds of data, and new platforms are introduced. As data becomes increasingly distributed across in-house and cloud deployments the work of moving, storing, protecting, and accessing data becomes fragmented with different practices depending on data locations and technologies. Changing and bolstering data management methods with each technological shift is difficult and disruptive. As technology innovation accelerates it will quickly become unsustainable. Data fabric can serve to minimize disruption by creating a highly adaptable data management environment that can quickly adjust as technology evolves. A data fabric platform eases the pain of managing data with the following features:

- **Unified data management:** Providing a single framework to manage data across disparate deployments reduces the complexity of data management.
- **Unified data access:** Providing a single and seamless point of access to all data regardless of structure, database technology, and deployment platform creates a cohesive analytics experience working across data storage silos.
- **Consolidated data protection:** Data security, backup, and disaster recovery methods are built into the data fabric framework. They are applied consistently across the infrastructure for all data whether deployed in cloud, multi-cloud, hybrid, or on premises.
- **Centralized service level management:** Service levels related to responsiveness, availability, reliability, and risk containment can be measured, monitored, and managed with a common process for all types of data and all deployment options.

- **Cloud mobility and portability:** Minimizing the technical differences that lead to cloud service lock-in and enabling quick migration from one cloud platform to another supports the goal of a true cloud-hybrid environment.
- **Infrastructure resilience:** Decoupling data management processes and practices from specific deployment technologies makes for a more resilient infrastructure. Whether adopting edge computing, GPU databases, or technology innovations not yet known, the data fabric's management framework offers a degree of "future-proofing" that reduces the disruptions of new technologies. New infrastructure end-points are connected to the data fabric without impact to existing infrastructure and deployments.

Data Fabric Functions and Features

A Single Data Management Platform

Data fabric "stitches" together the many complex components of modern data ecosystems, providing a single platform to manage data of all types across multiple platforms, technologies, and topologies. On-premises, cloud, multi-cloud, hybrid, and edge deployments are all viewed and managed through a single lens. Data warehouses and data lakes are equally visible and can be managed together. Best-fit processing technologies—Hadoop, Spark, Storm, Kafka, BigQuery, etc.—can each be employed as needed despite disparity of data lineage, processing optimization, and scalability. All essential functions of data management, from ingestion to access, are woven together to provide a complete and cohesive data management platform.

Beyond a single data management platform, data fabric should provide a smart data management platform that uses algorithms to understand data characteristics, collect metadata, inform and advise data consumers, protect sensitive data, automate repetitive and complex processes, and much more. Machine learning should enable adaptive data fabric capable of adjusting to changing data, changing business needs, and changing user behavior.

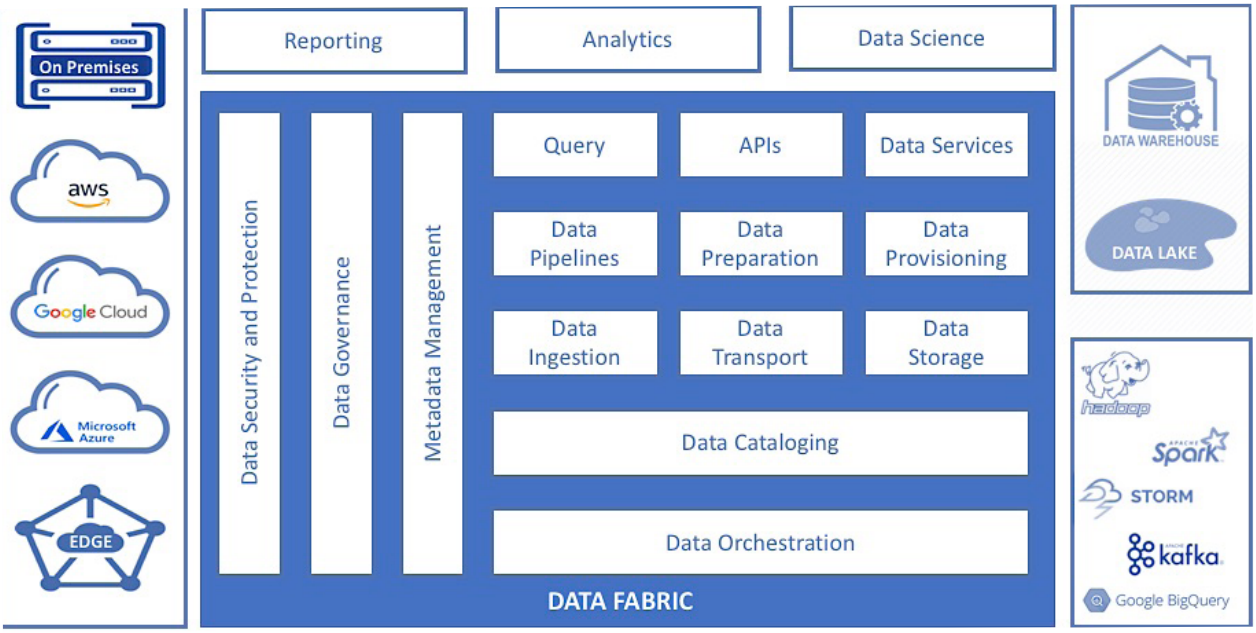
Data Fabric Components

To achieve the goal of complete and cohesive data management, a data fabric includes many components. (See figure 3.) External to the fabric, it must provide support for:

- Multiple platforms including on-premises data centers, cloud, cloud-hybrid, multi-cloud, and edge computing environments.
- Varied applications and use cases including reporting, analytics, and data science.

- Common shared data stores such as data warehouses and data lakes.
- Popular processing and data management technologies such as Hadoop, Spark, Storm, Kafka, and BigQuery.

Figure 3. Data Fabric Components



Within the data fabric, components include the following:

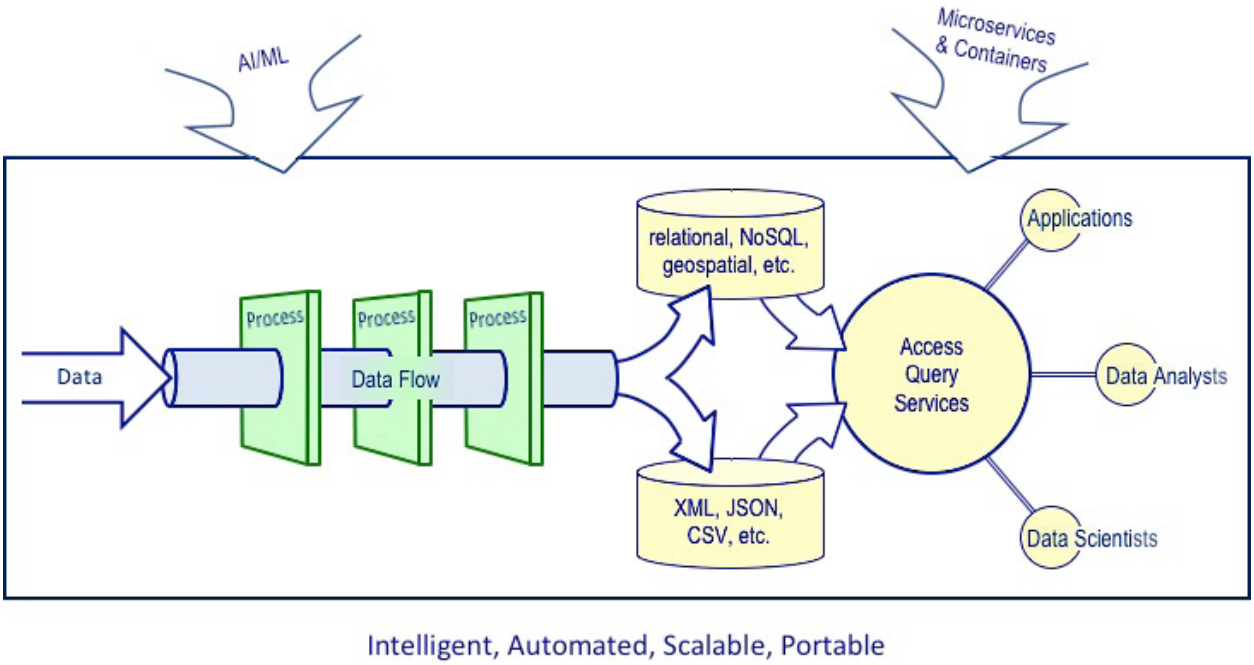
- **Data orchestration** coordinates participation by all stakeholders throughout the end-to-end data workflow. From ingestion and curation to preparation and consumption, workflows are orchestrated through execution of services that encapsulate reusable functions for data capture, storage, harmonization, governance, access, and application. Built on a foundation of AI/ML and automation, orchestration technology eases the pain of configuration, operationalization, and execution of data pipelines and analytics value chains. From a single control platform, orchestration provides the ability to activate any service, using any technology, across any network. Automated coordination of multi-services workflows across all execution environments—on-premises, cloud, multi-cloud, and hybrid—minimizes the manual effort of data operations and supports highly adaptable data management systems that readily adapt to change.
- **Data cataloging** has a central role in data management, collecting extensive metadata to support dataset searching, data understanding and evaluation, data governance, data sharing, and knowledge sharing. Smart data cataloging

includes algorithmic metadata discovery, semantic inference, relationship discovery, and automated detection and tagging of data sensitive to privacy, security, and compliance.

- **Data ingestion** has undergone dramatic change with growth of data volume, variety, and velocity. Modern ingestion methods must support batch, real-time, and stream processing. Relational data must coexist with big data formats such as JSON, time-series, Apache Avro, Apache Parquet and, most notably of late, cloud object data formats—particularly AWS S3. Smart data ingestion tools will recognize each incoming data type and format, understand the structure of datasets with self-describing schema, recognize previously known data sources and process them as needed, and detect and respond to schema changes. When disruptive data source and schema changes require human intervention, the fabric will employ machine learning to improve its ability to automatically adapt to similar changes in the future.
- **Data transport** moves data across networks, from one storage location to another or from a storage location to a processing location. With cloud data storage or cloud-based processing, this typically means transporting data across the internet—exposing data on public networks. Securing data in motion is an important consideration. Ideally, smart data transport technology automatically recognizes sensitive data assets and protects them with encryption or secure data transport protocols.
- **Data storage** technology advances are among the most important drivers of change in data architecture today. In-memory and combinations of in-memory and disk methods have altered the economics of data stores for both structured and unstructured data. Ultralow-cost cloud object storage technologies are experiencing rapid adoption. Smart data fabric will support the full variety of data storage options, automatically applying the best mix of storage technologies depending on use cases. The trend toward stateful microservices and containers adds yet another dimension to data storage evolution, and emphasizes the role of data fabric in “future-proofing” data architecture and technology infrastructure.
- **Data pipelines** consist of data flow and processing that moves data from an origin (a data source or data store) to a destination (a data store or consuming application) and transforms the data to meet requirements of the destination. Modern data pipelines must handle high-velocity processing of messages, logs, and data streams, in real-time and with scheduled or triggered batch processing. The new generation of data pipeline technology applies AI/ML for smart data pipelines able to recognize incoming data, know how to process that data and where to deliver it, project data volumes and dynamically scale, and detect and respond to data and process anomalies. The pipelines often use microservices

and containers. They are decoupled from specific execution platforms and technologies, providing a high level of portability and ease of migration between execution environments. (See figure 4.)

Figure 4. Modern Data Pipelines



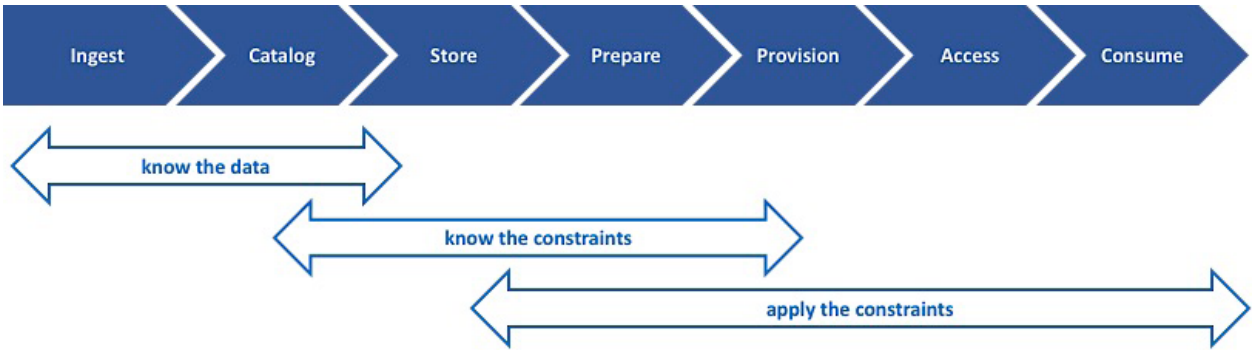
- **Data preparation** encompasses all of the transformations undertaken to create analysis-ready data including processing to improve, enrich, aggregate, format, and blend data. Self-service data consumers prepare data using visual, code-free data preparation tools. Data engineers may prepare data using code-free data preparation tools, scripting tools, or a combination of the two. Smart data preparation includes data relationship discovery, recommendations for data transformations, recognition and automatic masking of sensitive data, and collection of lineage metadata.
- **Data access** is provided through query, APIs, and data services. A wide variety of human and digital data consumers, combined with diverse use cases, requires support for many data access protocols. Query continues to be a common and widely used form of data access. Smart data fabric may include intelligent query optimization. APIs and data services, based on remote procedure call (RPC), SOAP, and REST protocols often build semantics, intelligence, and rules into data access mechanisms.
- **Data provisioning** is simply the work of providing data to consumers—getting the data where the consumer needs it, and in the required form and format. Sometimes provisioning delivers shareable data into a data store such as a data

lake or data warehouse. In other instances, it serves individual consumers and specific use cases such as a file delivered to a data scientist or data loaded into an analytics sandbox. Data provisioning depends on all of the other data fabric components—ingestion, transport, storage, pipelines, preparation, and access—to satisfy the variety of consumers and use cases that are possible in data- and analytics-driven organizations. The intelligence built into each component is the foundation of smart data provisioning.

- **Data security and protection** are ubiquitous themes throughout data management with direct relationships to all other fabric components. A distinct security and protection component is needed to provide cohesion and continuity across all security and protection touch points. Smart data security interoperates with existing authentication and authorization infrastructure. Smart data protection guards against risks from intrusion, corruption, and loss by automating detection and tagging of sensitive data assets and by providing data recovery capabilities.

- **Data governance** is similar to security and protection—a ubiquitous concern throughout data management. Security and protection are a subset of the scope of governance objectives, which also include data privacy, protection of personally identifying information (PII), regulatory compliance, data quality, data retention, and risk mitigation. At the macro level, data governance involves knowing when data is collected and cataloged, knowing data constraints when storing, preparing and provisioning data, and applying those constraints at all stages from point of storage to point of consumption. Smart data governance in the data fabric relies on the intelligence built into all of the components—ingestion, cataloging, transport, storage, pipelines, preparation, provisioning, and access.

Figure 5. Data Governance across the Analytics Lifecycle



- **Metadata management** is another ubiquitous component with touch points throughout the data fabric. Metadata reduces friction throughout the processes of working with data. It is needed to search and understand data, assess data

quality, prepare and provision data, protect and govern data, trace data lineage, and trust data and analysis results. The following smart metadata opportunities exist throughout the fabric:

- Schema detection, relationship detection, and metadata extraction at ingestion
- Semantic inference and automated tagging when cataloging
- Relationship detection and knowledge graphing at ingestion and cataloging
- Profiling and quality estimation when cataloging
- Physical attribute metadata with data storage
- Process and lineage metadata as part of data preparation
- Delivery metadata with data provisioning
- Frequency and performance metadata with access
- User and usage tracking with consumption

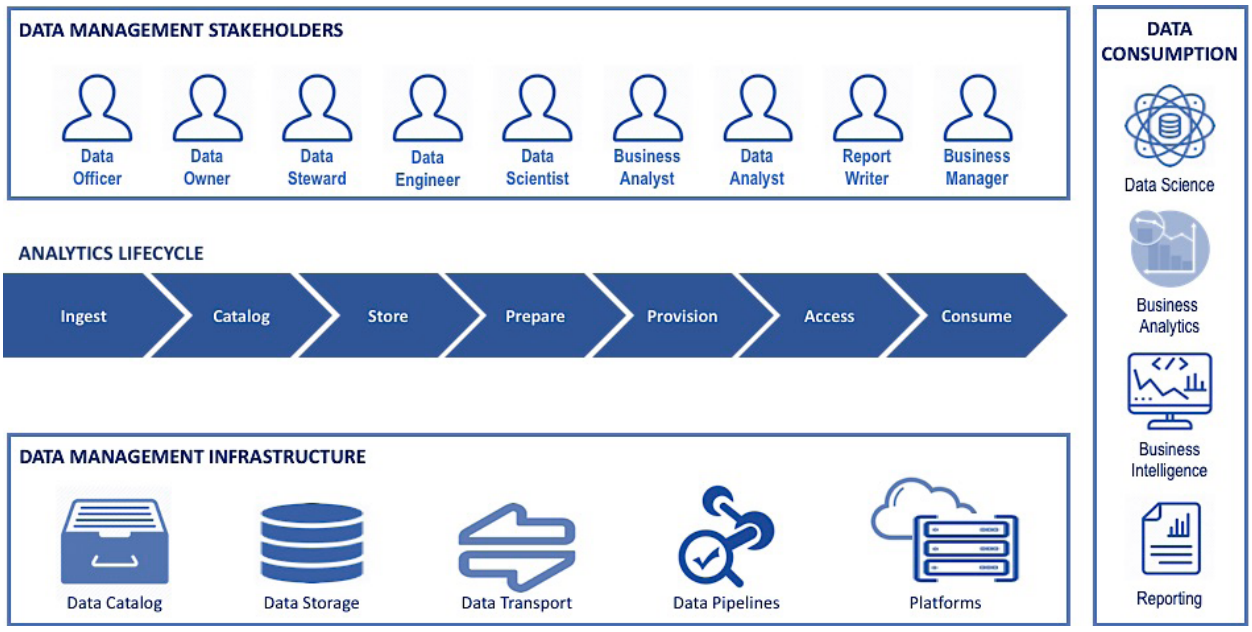
These are but a few of the many opportunities for automated metadata extraction and collection throughout the data fabric. Metadata is the means by which data managers and data consumers can know the data. Knowing the data is fundamental to deriving value from data.

Data fabric brings all of these components together as a single data management platform. Orchestration and cataloging are the foundation. Ingestion, transport, and storage manage data in motion and data at rest. Pipelines, preparation, and provisioning blend, harmonize, integrate, and otherwise make data ready for analysis. Query, APIs, and services make the data accessible. It is all supported by the critical administrative and oversight capabilities of security and protection, governance, and metadata management.

Data Management across the Analytics Lifecycle

From a cost and value perspective, data management is all cost. Value isn't realized unless the data is used to inform business processes and decisions, and to drive actions that produce positive business results. The data fabric reduces the cost of data management by automating complex processes, reducing manual effort, accelerating results, minimizing errors, and reducing waste and rework. Producing value begins with data consumption—data science, business analytics, business intelligence, and reporting. Every data management stakeholder (and there are many) has needs and expectations at one or more stages across the analytics lifecycle from ingestion to consumption. Fulfilling the needs and meeting the expectations depends on reliable data management infrastructure.

Figure 6. Data Fabric and the Analytics Lifecycle



Revisiting some of the data fabric components described earlier helps to understand many of the stakeholder needs and expectations.

- **Data ingestion**—Chief data officers expect that any data needed by the business, regardless of source and format, can be ingested at the speed needed by the business. Data owners and stewards expect that security- and privacy-sensitive data is recognized as soon as it enters the data ecosystem. Data engineers need the right technologies to ingest data of all types at any velocity from streaming to batch. Smart data fabric helps to ensure reliable data ingestion processes without disruption from schema and data source changes.
- **Data Catalog**—Data stewards need to have data cataloged and expect that all of the right metadata is recorded in the catalog to accurately describe the data and to help data consumers know how to use it. Data engineers use the catalog to find reusable datasets and reusable processes that promote consistency and reduce redundancy in data engineering work. Scientists, analysts, and report writers expect the catalog to provide robust search capabilities, and to help them to find, understand, evaluate, and access data. They also expect to find data wherever it resides without needing to know or care about deployment databases and platforms. Business managers expect faster analysis and greater analytics capacity with the catalog radically reducing the time that analysts spend finding and understanding data. Smart data fabric uses the AI/ML features of data catalogs to automate metadata extraction and collection.

- **Data Storage**—Data engineers need the ability to mix and match storage technologies depending on data types, data formats, and expected uses of the data. They need to store relational data, documents, images, geospatial data, property graphs, knowledge graphs, and more. Their expectations include storing data both on premises and in the cloud, using relational databases, NoSQL databases, and object stores. Smart data fabric helps data engineers make informed choices for data storage.
- **Data Preparation and Provisioning**—Data owners and stewards need to ensure that governance is an integral part of preparation and provisioning processes. Data engineers need robust capabilities to process data using a variety of technologies and processing engines, to deploy processing to multiple points across a network, and to deploy workflows across on-premises, cloud, multi-cloud, and hybrid environments. Self-service scientists and analysts need the ability to use self-service data preparation technologies without data access bottlenecks. Business managers care that their analysts get the data they need when and where it is needed. Smart data fabric connects data governance with data preparation and provisioning and supports process orchestration across all deployment platforms and execution environments.
- **Data Access**—Data owners and stewards need to ensure that data access works according to data security and protection requirements. Data engineers need quick and easy access to data for provisioning, and they need to take full advantage of reusable processes and pipelines when building query, API, and data services capabilities. Data consumers expect data to be readily accessible when needed. Smart data fabric supports authorized access connected with existing security infrastructure for authentication and authorization. For data consumers, it connects data catalog and data access to support seamless data searching, understanding, evaluation, and access.
- **Data Consumption**—Data stewards need to track data consumption to have full knowledge of who uses which data for what purposes. Data engineers track data consumption to identify unused and underutilized datasets and to retire obsolete data stores and data pipelines. Data scientists, business analysts, data analysts, and report writers expect a highly flexible consumption layer that quickly and easily adapts to the variety of ever-changing use cases. Smart data fabric collects usage metadata. It supports a wide variety of use cases through smart data preparation and provisioning.

Data Infrastructure Management

As data management has grown in multiple dimensions—volume of data, velocity of data, variety of data types, number and variety of consumers, number and variety of use cases, multitude of processing engines, and variety of deployment options—managing the data

infrastructure becomes increasingly complex. Sizing the infrastructure to fluctuating workloads and data volumes, managing distributed and parallel processing, adapting to continuous change, and adopting new technology innovations are at the core of infrastructure management complexities. As a unified platform for data management, data fabric must include infrastructure management features and functions.

Scalability and Elasticity

Growth is one of the biggest challenges of data infrastructure management. Expanding data sources and volumes, increased processing workloads, and growing user base all strain infrastructure to the point where scaling up is inadequate. Scaling out is the only practical approach, with compute capacity and storage capacity scaling independently. Scaling, however, is only part of the story. In a volatile data and analytics environment workload peaks and valleys can be quite extreme, driving the need for elastic infrastructure that can dynamically allocate and deallocate resources to match processing and data capacity with workload and storage requirements. Cloud technologies are inherently scalable and elastic, but getting advantage from cloud capabilities requires smart infrastructure management. An intelligent data fabric can anticipate workloads and forecast demand for processing and storage capacity, dynamically assign workloads to processors, and choose the optimum processing location and processing engine for a particular job.

Multi-Cloud and Cloud-Hybrid Support

As described earlier in this report, nearly every data management landscape includes multiple cloud platforms plus on-premises data storage and processing. (Refer back to figure 1.) Data fabric orchestration capabilities come to the forefront to meet this challenge. Data stored across multiple environments should not be isolated or siloed. Processing can't be confined to a single execution environment when data resides on multiple platforms. With data fabric orchestration, a single control platform can execute a sequence of services, built with diverse technologies and distributed across multiple execution environments. When data is distributed across multiple platforms, it makes sense to push the processing to the data location. Smart data fabric automates the coordination of multi-services workflows across all execution environments for comprehensive cloud, multi-cloud, and cloud-hybrid support.

High Performance and Optimization

As the world of analytics grows in every dimension—volume of data, volume of processing, complexity of processing, number of simultaneous users, frequency of queries, etc.—performance becomes a key consideration. Multiple data pipelines executing simultaneously can overload systems and create processing bottlenecks. Getting all of the work done quickly and efficiently, without impact to data freshness or user experience, is difficult when managing the execution environment manually. Data fabric automates runtime monitoring to predict, prevent, and detect processing logjams. Data fabric minimizes the impact of processing inefficiencies with early detection and automated optimization. Real-time optimization is practical when the logjam can be cleared by reallocating workloads and taking advantage of cloud scalability and elasticity. Optimization delays happen when the fabric

makes recommendations to increase network bandwidth, reallocate resources, optimize database queries, use data virtualization, or other solutions that require data engineer or systems administrator intervention.

Future-Proofing and Infrastructure Resilience

Future-proofing the technology infrastructure begins by anticipating future changes and using methods that minimize the disruptive effects of those changes. In data management, anticipating the changes is easy. Technologies will continue to evolve and innovate, widening the gap between leading and trailing technologies. Data volumes will continue to grow with increased collection from IoT and mobile devices. Analytics processes and technologies will become more complex and more compute-intensive as cognitive computing advances drive adoption of AI/ML throughout business processes. Numbers of users and use cases will grow as self-service adoption and generational shifts drive data democratization.

Yes, anticipating the changes is easy. Responding to those changes is more difficult and filled with hard questions. How do you continue to increase data storage capacity to keep up with data growth? How do you continue to increase compute capacity to cope with more data pipelines, more complex analytics, and more simultaneous processing driven by a growing population of self-service users? How do you fit new technologies into the existing technology landscape and architecture? How do you pull the trailing edge of technology forward as you push the leading edge? How to you commit to technologies that meet today's needs without experiencing vendor lock-in that limits your options for the future?

Obviously, these aren't easy questions to answer. But data fabric holds promise and eases the pain. With a unified platform for data management the changes need to be accommodated in only one place instead of across a disparate collection of data management technologies with limited interoperability. With smart data fabric and microservices/container architecture, processing is easily ported between environments and technologies, and the infrastructure becomes highly resilient.

State of the Market

Data Fabric Use Cases

The use cases for data fabric are abundant and somewhat overlapping. Anyone managing a complex data and analytics environment will certainly benefit by adopting data fabric. Architectural modernization—rethinking data management architecture for the age of analytics—is a good opportunity to adopt data fabric concepts and principles and to move toward data fabric technology. When migrating legacy data warehouses to cloud, expect to become a multi-cloud and cloud-hybrid environment where data fabric delivers great value. Building a DataOps organization is a good use case, because DataOps without automation

isn't practical. Organizations driving toward self-service and data democratization will experience many of the challenges that data fabrics address. Those pursuing data-driven digital transformation will find it difficult to achieve without smart and automated data engineering, operations, and orchestration.

Current State

Data fabric is a relatively new concept and the technology market is best characterized as emerging and evolving. Today no single vendor provides a complete data fabric solution in a single product. Data fabric vendors fit into the following three main categories:

- **Single Product**—A few vendors offer a single product that provides much, but not all, data fabric functionality. These are good choices if they offer the functions that are most essential for your organization and have a strong roadmap for the future.
- **Single Vendor**—Some vendors offer multiple products, each providing a subset of data fabric functionality. They emphasize interoperability among the products as the approach to delivering some (again, not all) data fabric functionality. These may be good choices if the product suite provides the most needed functions, especially for those who already use some of their products.
- **Multiple Vendors**—Many vendors offer products that deliver a subset of data fabric functionality. These can be good choices for those organizations whose technology infrastructure includes diverse products that each provide some data fabric functionality, and who have the technical ability to resolve interoperability among disparate products.

Data fabric architecture varies with vendors and products. One vendor sees orchestration as the umbrella under which all other functions operate; orchestration is the glue that holds it all together. Another vendor drives data fabric from a graph perspective where the combination of relationships and AI/ML are the foundation of data fabric capabilities. One positions the data catalog as the centerpiece, shifting from passive catalog to what they call "active data hub." Yet another vendor bases data fabric capabilities on data virtualization, with an abstract semantic layer as the foundation. Data lake management vendors are also evolving to become data fabric providers, building on their pipeline management and storage management functions with expanded automation and orchestration features.

The Future of Data Fabric

Data fabric is relatively new, but much needed. The world of data management will continue to grow in scope and complexity. Managing without data fabric will become impractical within a few short years, and those who fail to adopt data fabric will fall behind in areas where data is a competitive differentiator. The technology will continue to expand, improve, and innovate. Expect to see some convergence of architectural perspectives—graphs blended

with orchestration, for example, or a catalog and graph combination. Ultimately, every data management product will have a role in data fabric, but today's leaders will shape the future.

Getting Started with Data Fabric

Do You Need Data Fabric?

Whether you need data fabric is not in question. The real question is *when* you will need data fabric. How severe is the silo effect of your data resources today? Will it intensify in the future? Is data engineering an analytics bottleneck today? How quickly will demand for data engineering capabilities expand? How can you grow data engineering capacity? Do you struggle to operationalize data pipelines? Are production data pipelines difficult to manage and maintain? Is your runtime environment complex and fragile? Are execution conflicts and logjams commonplace? Do errors and exceptions disrupt routine workflows?

Don't ask if you need data fabric. Ask when you'll need data fabric.

If these questions reflect the data management challenges you face, then it is time to look closely at data fabric and how it will fit into your data management processes and practices.

Recommendations

The top-level conclusion from this report is that you will need data fabric to build and sustain a data-driven organization. Getting to data fabric is a journey that involves several activities.

- Begin by looking at the business case for data fabric. What is the cost/value ratio of data management in your organization today? How much could you reduce cost through automation? How many value opportunities go unrealized? How much can data value be increased? What benefits would unified data governance offer?
- Take a look at the technical case. Do you need multi-cloud support? Do you need cloud-hybrid support? Is your data management scalability and elasticity challenged? Does growth of data, processing, and users routinely degrade system performance? Do you need technical infrastructure agility? Are you concerned about vendor lock-in?
- Also look at the operational case. How difficult is it to operationalize data pipelines? Of all analytic models developed, how many are actually sustained in operations? How difficult, complex, and labor intensive is scheduling and workload management? How disruptive are errors and exceptions to routine workflow?

- Consider your data fabric use cases. Is your data management environment highly complex and difficult to manage? What data initiatives do you have planned or in process? Are you modernizing data architecture? Are you migrating to cloud? Are self-service and data democratization underway or in your future? Do you aspire to become a DataOps organization?
- Itemize the tangible benefits of data fabric. What will you gain with unified data management? How will unified data access provide benefits? What results do you expect from consolidated data protection? What value will central service level management provide? What needs can be met with cloud portability? How will infrastructure resilience help you?

Now you know the *why* of data fabric. The next steps are to answer *what* and *how.*

- Look at your existing technology infrastructure. What products offer some data fabric functionality? How limited are those capabilities? What level of interoperability is possible with the product mix? Where are the gaps in data fabric functionality?
- Determine your preferred approach to data fabric. Do you want a single product solution, a single vendor solution, or a multiple vendor solution?
- Plot the roadmap to implementing your preferred solution. If single product, what is the path from product evaluation to implementation? If single vendor, what is the path from vendor and product evaluation to implementation? If multiple vendor, how will you select products, in what sequence will you implement, and how will you stitch them together as a fabric of interoperating technologies?
- Don't forget the human side of data fabric. Who are the stakeholders? How will you get them engaged and committed? What participation is needed from each? What organizational changes might be needed?

Getting to data fabric isn't quick or easy. But for every data-driven organization it is necessary, either now or in the future. Begin today by asking the questions, exploring the possibilities, and looking toward a future of smart data engineering, operations, and orchestration.

About Eckerson Group

Wayne Eckerson, a globally known author, speaker, and advisor, formed Eckerson Group to help organizations get more value from data and analytics. His goal is to provide organizations with a cocoon of support during every step of their data journeys.

Today, Eckerson Group helps organizations in three ways:

- **Our thought leaders** publish practical, compelling content that keeps you abreast of the latest trends, techniques, and tools in the data analytics field.
- **Our consultants** listen carefully, think deeply, and craft tailored solutions that translate your business requirements into compelling strategies and solutions.
- **Our educators** share best practices in more than **30 onsite workshops** that align your team around industry frameworks.

Unlike other firms, Eckerson Group focuses solely on data analytics. Our experts each have more than 25+ years of experience in the field. They specialize in every facet of data analytics—from data architecture and data governance to business intelligence and artificial intelligence. Their primary mission is to help you get more value from data and analytics by sharing their hard-won lessons with you.

Our clients say we are hard-working, insightful, and humble. We take the compliment! It all stems from our love of data and desire to help you get more value from analytics—we see ourselves as a family of continuous learners, interpreting the world of data and analytics for you and others.

Get more value from your data. Put an expert on your side.
Learn what Eckerson Group can do for you!

About Infoworks

Infoworks provides the first Enterprise Data Operations and Orchestration software system to automate the development and operationalization of data pipelines from source to consumption in support of business intelligence (BI), machine learning (ML) and artificial intelligence (AI) analytics applications. Infoworks' code-free development environment allows organizations to develop and manage end-to-end data workflows without requiring an army of big data experts. The software system automates and simplifies development of data ingestion, data preparation, query acceleration and ongoing operationalization of production data pipelines at scale. Infoworks supports cloud, multi-cloud, and on premise environments, enabling customers to deploy projects to production within days, dramatically increasing business agility and accelerating time to value.

Learn more at www.infoworks.com