Eckerson Group | White Paper

The Rise of Dataware

An Integration-Minimizing Approach to Data Architecture

JOE HILLEARY

APRIL 2022

RESEARCH SPONSORED BY CINCHY

cinchy

About the Author

Joe Hilleary is a writer, researcher, and data enthusiast. He believes that we are living through a pivotal moment in the evolution of data technology and is dedicated to helping organizations find the best ways to leverage their information. The author of numerous articles and reports covering analytics, governance, and data sharing, he has also spoken widely on topics ranging from data management to business intelligence and DataOps. A firm advocate of innovation, he takes pleasure in keeping up with the latest developments in the world of data. He is a Senior Analyst at Eckerson Group.

About Eckerson Group

Eckerson Group is a global research and consulting firm that helps organizations get more value from data. Our experts think critically, write clearly, and present persuasively about data analytics. They specialize in data strategy, data architecture, self-service analytics, master data management, data governance, and data science. Organizations rely on us to demystify data and analytics and develop business-driven strategies that harness the power of data. Learn what Eckerson Group can do for you!

About This Report

Research for this report comes primarily from numerous briefings with software vendors. This report is sponsored by Cinchy, who has exclusive permission to syndicate its content.

Table of Contents

Executive Summary

Today, every enterprise has dozens, if not hundreds or more, software applications, both internal and customer-facing. Each application generates and stores massive amounts of data that is often shared with other applications to keep them in sync. In addition, companies leverage that data for analytics, bringing together insights from multiple sources to inform decisions and make predictions. Moving all this data around requires integration—and lots of it. Now, a radical new architectural approach called "dataware" redefines the relationship between data and applications and eliminates the need for data integration.

Dataware has the potential to revolutionize both software development and analytics. Just as the advent of software decoupled function from hardware and enabled the birth of the modern computer, dataware seeks to decouple data from software to enable a new paradigm of application-independent data. With dataware, applications don't have to maintain their own data store or ship data to other applications or analytics repositories. Instead, developers build applications on top of a centralized data store from the beginning. This approach will permit software to utilize data from other applications without point-to-point integrations. Widespread adoption of this approach requires buy-in from numerous stakeholders, especially software vendors, but on an organizational level, the technology needed to facilitate the transition a dataware paradigm is making in-roads.

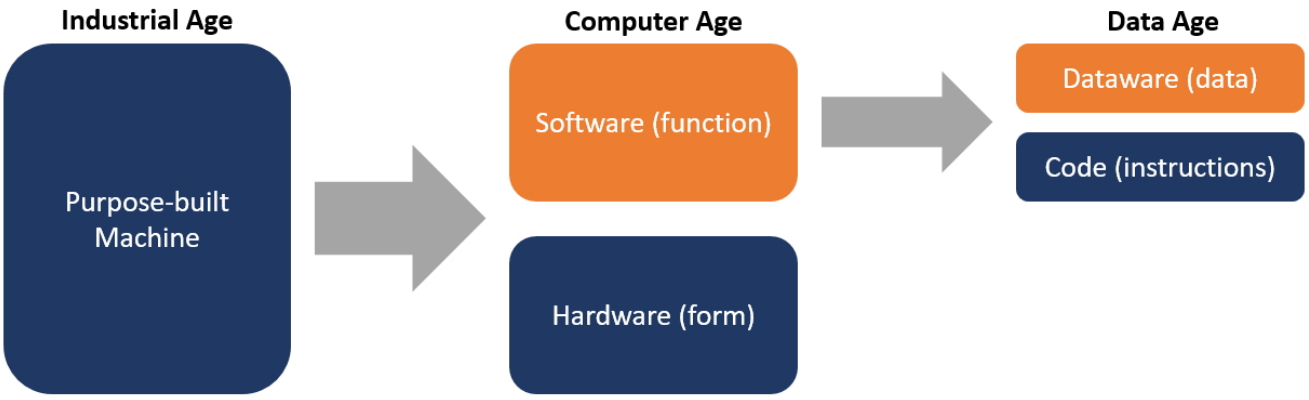
This report will define dataware and examine what it is and is not. In particular, it will compare dataware and other popular approaches to data centralization, such as data warehouses, data lakes, and data fabrics.

Introduction

We've spent the last several decades allowing software to fracture our data and then tasking data teams with putting the pieces back together. We integrate individual applications with one another and use centralized repositories to unify data for analytics. But what if we created a unified model of our data from the beginning, instead of after the fact? A new architectural approach called "dataware" does that—creating a shared data layer for applications that also provides a foundation for analytics.

Dataware is an evolution in both technology and methodology. Just as software liberated form from function, enabling the same hardware to perform multiple tasks, dataware liberates data from code. (See figure 1.)

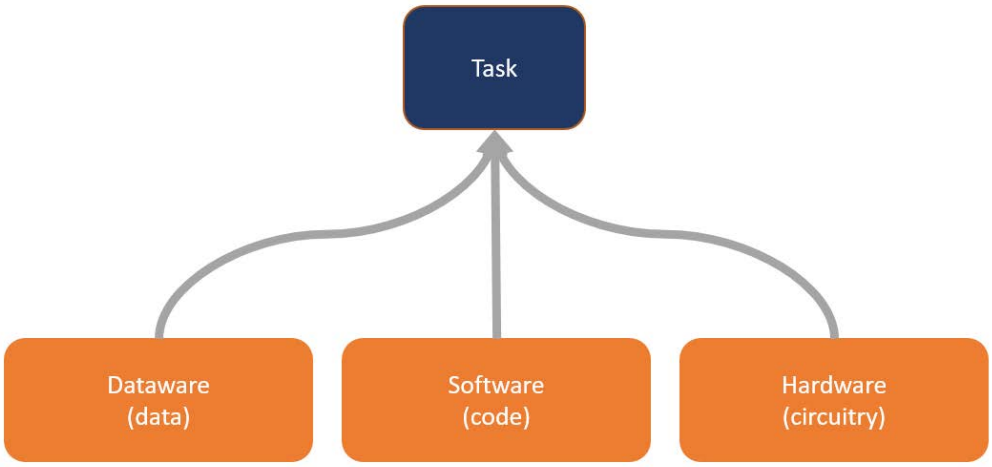
Figure 1. The Evolution of Dataware



The present configuration of software that contains both a database and code works when a single program runs in isolation. Once multiple pieces of software begin to interact, however, that architecture becomes a liability. Data from one piece of software must travel to other applications so that they can cooperate. Integration accomplishes this task by making a copy of the data in each application and moving it to all of the others. This approach results in the rapid proliferation of copies of data—all of which must be stored and maintained, increasing costs, complexity, and potential for error.

The dataware approach addresses this issue by taking the task of storing data away from individual applications. In this paradigm, software becomes responsible for the code alone. To accomplish a task, software provides the instructions, dataware provides the data, and hardware provides the circuitry to physically execute everything. (See figure 2.)

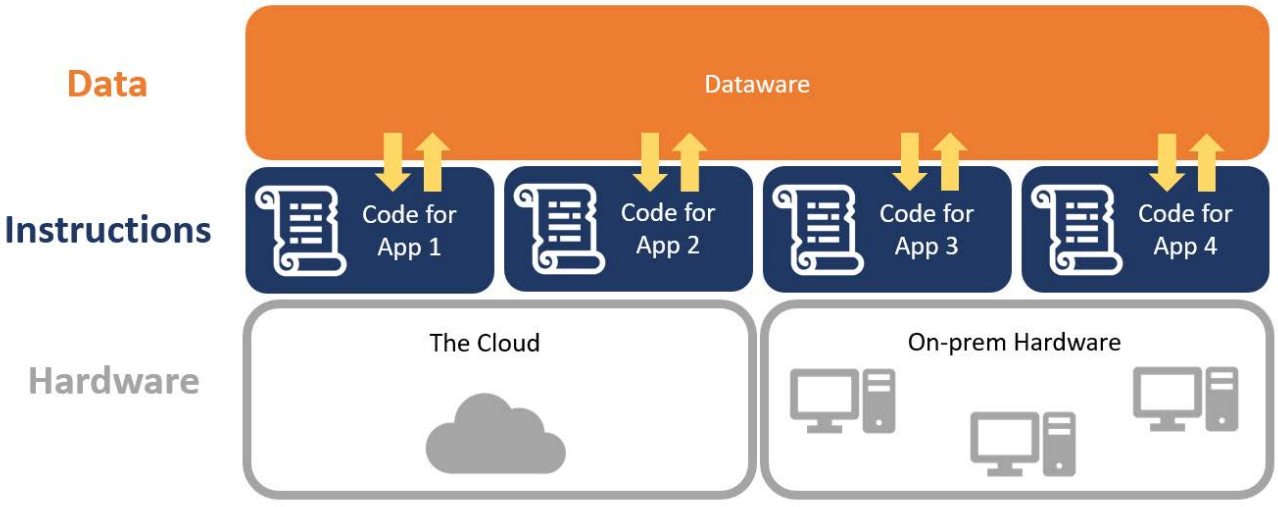
Figure 2. Accomplishing a Task with Dataware



Software provides the instructions, dataware provides the data, and hardware provides the circuitry to physically execute everything.

Like hardware, data becomes independent of a particular function. Just as a computer can run multiple applications, dataware serves as a shared data layer for multiple applications. Engineers simply specify a dataware platform during software configuration in the same way they would configure software to run in on-premises or cloud-facilitated hardware environments. The instructions from the code access data from the dataware and run on the hardware. (See figure 3.)

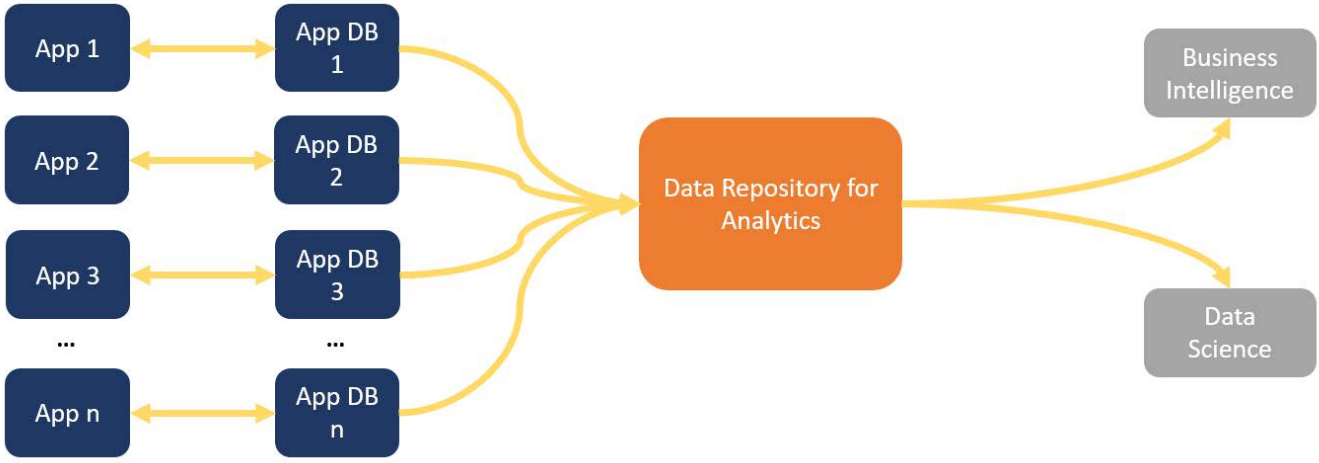
Figure 3. The Relationship Between Dataware, Software, and Hardware



Data Centralization

A key aspect of the dataware approach is the idea of data centralization. Companies have realized the value of bringing together data from multiple sources for decades in the world of analytics. Consider the basic model of data flow for data analysis: Using extract, transform, load (ETL) pipelines, organizations copy data from application databases to a comprehensive repository of data for analytics. From there, joins of different tables feed workflows for both business intelligence and data science teams. (See figure 4.)

Figure 4. The Basic Model of Analytics Dataflow



The concept of centralizing data for analysis forms the backbone of data warehousing and every other analytics trend that has followed. In some paradigms, pipelines integrate the data physically, in others, tools virtualize the data, centralizing it logically. Different technologies also offer support for different types of data. But ultimately the core idea remains.

Dataware extends this concept to applications, allowing the same repository that drives analytics to serve as the backend for software. In consolidating both analytical and operational data, dataware removes the need to copy data for either analytics or application integration. Software writes to and analytics tools read from the exact same repository.

Dataware allows the same repository that drives analytics to serve as the backend for software.

Reducing the number of copies of data floating around solves several issues. First the creation and storage of copies costs money. Space may be relatively cheap in a world of cloud storage, but it's definitely not free. And why pay to store the same data twice or even dozens of times? Likewise, building new data pipelines to move those copies around requires expensive software and data engineers, and as the number of applications an organization uses grows, the number of pipelines needed to integrate data increases exponentially. What's more, copying and moving data creates opportunities to introduce errors. If two systems connected by a data pipeline desync, data can be lost or corrupted, reducing data quality. With one copy of the data used collectively by all systems, there's no chance of records appearing differently in different contexts.

Perhaps the most important impact of data centralization and copy reduction, however, is the increase in control over data. When all applications and humans use data from a single source, enforcing access and privacy polices becomes much easier. Governance professionals can programmatically grant access to different users or applications without fear that someone will get ahold of a deprecated copy of a table or one that contains sensitive fields. This reduces the risk of both data leaks and the costly regulatory penalties that accompany them. Furthermore, it eliminates any confusion over the "real" values of certain fields by creating a single source of truth.

The primary benefits of dataware stem directly from its ability to realize the vision of complete data centralization. It both physically and logically centralizes the data used not only for analytics but also for operational applications. Although the former has long been accomplished using integration, the latter is a technological leap forward that removes the need for integration altogether.

What Dataware Is

All dataware platforms should share the following attributes:

Basic Attributes

1. **Read and write capabilities.** Dataware must allow users and applications not only to access data, but also to record new data.
2. **Data collaboration.** In order for the entire approach to work, dataware must facilitate real time editing of data by multiple entities with share access. Think about collaborating with colleagues on a Google Doc. Dataware needs that type of functionality, but for data. Different applications must be able to edit the same data without conflict. The platform should also have an accessible user interface that allows business users to directly access and edit the data as easily as applications can.
3. **Operational and analytical base.** Dataware must serve as the backend for both operational applications and analytics workflows. Applications must interact with it in the place of an operational database, while data analysts need to be able to query it like a data warehouse or other analytics repository. This means it must support both transactional data and analytical data. The former consists of the individual business events generated by a company's systems, while the later is composed of highly-queryable aggregated records used for business intelligence.
4. **Active metadata layer.** Humans and applications both use different terminology when referring to the same things. As a result, dataware must have an active metadata layer that maps between the semantics of individual applications and the generic canonical model of the dataware. This layer serves to create semantic consistency. No matter what an application or individual might call a given field, the metadata layer will help the dataware platform provide the right information. In addition, the metadata layer will track collaboration, enabling traceback for edits.
5. **Change approval process.** Finally, dataware will provide a mechanism for approving changes to the data. In a user-based framework, dataware will enable data governance professionals to determine who or what applications may change which records. The dataware will permit organizations to implement access policies on a programmatic basis, so that the entire system can scale.

What Dataware Isn't

Often, it's easier to understand something new by understanding how it differs from other, more familiar, concepts. In this section, we will contrast dataware with a handful of other, related approaches. In that spirit, here is what dataware is not:

A Data Warehouse

	Data Warehouse	Dataware
Centralized Access	✓	✓
Networked Data	✗	✓
Centralized Governance	✓	✓
Federated Governance	✗	✓
Applies Schema	✓	✓
Feeds Applications (Transactional Data)	✗	✓
Replaces Application Databases	✗	✓
Feeds Analytics (Analytical Data)	✓	✓
Technology	✓	✓
Methodology	✓	✓
Requires Copying and Moving Data	✓	✗

A data warehouse is a structured central data repository that stores data in a pre-defined model for business intelligence. Unlike dataware, a data warehouse does not serve as a backend for applications. It also stores copies of data, which continues to reside elsewhere. Applications feed a data warehouse via pipelines, so instead of reducing the number of integrations needed. It adds an additional one per application.

Data Virtualization

	Data Virtualization	Dataware
Centralized Access	✓	✓
Networked Data	✗	✓
Centralized Governance	✗	✓
Federated Governance	✓	✓
Applies Schema	✓	✓
Feeds Applications (Transactional Data)	✗	✓
Replaces Application Databases	✗	✓
Feeds Analytics (Analytical Data)	✓	✓
Technology	✓	✓
Methodology	✓	✓
Requires Copying and Moving Data	✗	✗

Data virtualization consists of a business view of distributed data that uses query federation to join data from heterogeneous systems in real time. Data virtualization shields business users from the complexity of backend systems, eliminates the need to move or copy data, and gives data administrators the ability to change those systems without impacting downstream reports and applications. Although, like dataware, data virtualization eliminates the need to move or copy data from applications for analytics, it does not replace operational databases for software. Instead, data virtualization allows for the creation of logical data warehouses. It allows users to query data for analytics without needing to physically integrate with a central repository, but it does nothing to impact application to application integration.

A Data Lake

	Data Lake	Dataware
Centralized Access	✓	✓
Networked Data	✗	✓
Centralized Governance	✓	✓
Federated Governance	✗	✓
Applies Schema	✗	✓
Feeds Applications (Transactional Data)	✗	✓
Replaces Application Databases	✗	✓
Feeds Analytics (Analytical Data)	✓	✓
Technology	✓	✓
Methodology	✓	✓
Requires Copying and Moving Data	✓	✗

A data lake is a type of repository that stores structured, semi-structured and unstructured data in its native format. They originated as on-premises repositories running on Hadoop, but have evolved to run in the cloud as object stores. In the same way as a data warehouse, a data lake requires that pipelines carry copies of data from source applications to a central location where analysts or data scientists can query it. Unlike a data warehouse, it can accept unstructured data and does not impose a schema, but nonetheless it doesn't support application development, nor does it reduce the need for integration.

A Data Lakehouse

	Data Lakehouse	Dataware
Centralized Access	✓	✓
Networked Data	✗	✓
Centralized Governance	✓	✓
Federated Governance	✗	✓
Applies Schema	PARTIALLY	✓
Feeds Applications (Transactional Data)	✗	✓
Replaces Application Databases	✗	✓
Feeds Analytics (Analytical Data)	✓	✓
Technology	✓	✓
Methodology	✓	✓
Requires Copying and Moving Data	✓	✗

A data lakehouse is a hybrid data repository that combines elements of a data lake and a data warehouse and supports multiple workloads including business intelligence, data science, and self-service analytics. At a minimum, it supports SQL constructs and cloud-native object stores. As with both the traditional data warehouse and the data lake, the data lakehouse focuses on analytics use cases. It cannot serve as the backend for an operational application in the way dataware does. It also requires even more integration than either of its component parts because data travels first to the lake and then structured data moves again to a data warehouse-like structure within the lake.

A Data Fabric

	Data Fabric	Dataware
Centralized Access	✓	✓
Networked Data	✓	✓
Centralized Governance	✓	✓
Federated Governance	✓	✓
Applies Schema	✓	✓
Feeds Applications (Transactional Data)	✓	✓
Replaces Application Databases	✗	✓
Feeds Analytics (Analytical Data)	✓	✓
Technology	✓	✓
Methodology	✓	✓
Requires Copying and Moving Data	✗	✗

Data fabric is a recent, catch-all term used to describe a combination of architecture, services, and technology that delivers a unified semantic layer that works across any type of data, any source or target system, and any data infrastructure. The goal is to simplify user access and data management in a large, complex ecosystem. Of all the present approaches, it is the closest to dataware. Like dataware,

a data fabric does not require the copying or moving of data. Instead, both operational application and analytical queries rely on a network of repositories (the individual databases for applications). The data fabric presents a unified view of this network using virtualization. Dataware actually acts as a data fabric for applications that are not natively built using dataware in the place of an operational database. A traditional data fabric will always require applications to have their own databases that connect to the network, however, while dataware allows developers to build new applications without dedicated operational databases.

A Data Mesh

	Data Mesh	Dataware
Centralized Access	✗	✓
Networked Data	✗	✓
Centralized Governance	✗	✓
Federated Governance	✓	✓
Applies Schema	✓	✓
Feeds Applications (Transactional Data)	✗	✓
Replaces Application Databases	✗	✓
Feeds Analytics (Analytical Data)	✓	✓
Technology	✗	✓
Methodology	✓	✓
Requires Copying and Moving Data	✓	✗

A data mesh is a distributed data architecture in which business units own, manage, and publish data for others to consume. In a data mesh, data doesn't move, instead consumers, such as corporate data analysts, get it from a department. Unlike the previous examples, a data mesh is not actually a technology. It's a governance framework that can be implemented using any number of different technologies.

Master Data Management

	MDM	Dataware
Centralized Access	✓	✓
Networked Data	✗	✓
Centralized Governance	✓	✓
Federated Governance	✗	✓
Applies Schema	✓	✓
Feeds Applications (Transactional Data)	✓	✓
Replaces Application Databases	✗	✓
Feeds Analytics (Analytical Data)	✓	✓
Technology	✓	✓
Methodology	✓	✓
Requires Copying and Moving Data	✓	✗

Master Data Management (MDM) is a set of practices and tools used to maintain a consistent set of trusted records with standard attributes and terms for key business entities, such as customer, product, and locations. Most dedicated MDM tools do not replace application databases, nor do they remove the need for integration. In fact, they generally increase the amount of integration needed as pipelines copy the new data created in applications into the platform and copy the mastered data back out to ensure consistency. That said, an organization could accomplish MDM using dataware. The active metadata management layer and data collaboration features of a dataware platform facilitate the same functionality as an MDM platform, only without the integration.

Microservices

	Microservices	Dataware
Centralized Access	✗	✓
Networked Data	✓	✓
Centralized Governance	✗	✓
Federated Governance	✗	✓
Applies Schema	✗	✓
Feeds Applications (Transactional Data)	✓	✓
Replaces Application Databases	✗	✓
Feeds Analytics (Analytical Data)	✗	✓
Technology	✓	✓
Methodology	✓	✓
Requires Copying and Moving Data	✓	✗

A microservice architecture is a design pattern for applications in which the functions of the application exist as smaller, discrete components that communicate via API. This approach makes it easier to test,

maintain, and add additional functionality. Unlike dataware, microservices affect the code layer of applications. Essentially, they facilitate a networked approached to code that's similar to what a data fabric accomplishes for data. Because it doesn't focus on the data layer, a microservices architecture does not reduce the amount of data integration needed.

Remaining Hurdles for Dataware

Dataware has the potential to disrupt the entire data integration industry, but several major hurdles stand between it and the ability to realize its complete vision. In the short term, most businesses already have hordes of applications designed with traditional operational databases. Companies have invested heavily in these solutions and the means to integrate them. As a result, dataware platforms only truly act as such for new applications designed after an organization implements dataware. For existing applications, the dataware platform simply works like a data fabric. It connects the current environment in a network with a common semantic layer that will allow future applications to communicate without individual integrations to every legacy application.

Even for new applications, dataware platforms often remain relegated to the role of a data fabric. Sure, any new applications an organization develops in-house can benefit from the full power of the dataware approach, but most companies rely heavily on third-party solutions. Until the wider community of software vendors adopts dataware-friendly architectures, off-the-shelf applications will only be able to interact with dataware as a data fabric. This still reduces the overall amount of integration needed to add new applications to the data environment but limits the overall impact of the approach.

The Need for Dataware Standards

Wide-spread adoption of dataware requires not just buy-in from the entire community of software developers, but also the creation of new standards. Even if software vendors wanted to design applications for dataware, how could they make them compatible with all the variations that exist?

One organization, the Data Collaboration Alliance, is already working with the CIO Strategy Council of Canada to hammer out a standard for Zero-Copy Integration. This standard would define a framework for decoupling data from applications, thus creating a best practice for dataware. Although this effort is national, it could provide a starting point for a more global standard. Unless, this or some other standard catches on, it's difficult to imagine third-party software developers being willing to develop applications that run on dataware. It's not dissimilar to the early days of computing when software had to be written for specific hardware configurations because of a lack of standards in chip design.

Deciding to Buy or Build

Another major challenge for dataware is the difficulty of implementing the approach. Creating a platform that meets the requirements of dataware takes a lot of effort. Currently, only Cinchy, K2View, and Microsoft, in the form of their dataverse product, offer dataware platforms off the shelf. As the space matures, more vendors are likely to step into the ring, but for now the options to acquire dataware are limited. Below are pros and cons for both building a custom dataware platform and buying one of the few third-party platforms:

Building Dataware

Pros:

- Don't need a universal solution, just one that works for your data and applications
- Can tailor the platform to your environment for better performance
- Avoid vendor lock-in

Cons:

- Complex architecture with very few examples to draw from
- Takes developer time away from other priorities
- Long-term responsibility for maintenance
- Won't necessarily meet future standards

Buying Dataware

Pros:

- Much faster time to implement.
- You can receive advice on how to use dataware from experts.
- There is more down-the-line flexibility.
  - The vendor has the incentive to make the platform compatible with standards and as many third-party tools as possible

Cons:

- Every future application you build is locked into the same vendor because of the shared proprietary backend.
- No control over costs—the price is set by the vendor and potentially subject to change over time.
  - As critical infrastructure, there's little room to negotiate once your entire environment feeds into the platform.

Conclusion

Dataware is a bold reimaging of the relationship between data and applications. It promises to reduce the burden of integration by providing a shared data layer for all of an organization's applications while simultaneously serving as the base for analytics. The approach remains in its early days, but concept of separating data from applications has major ramifications not only for the data world but also for software development. If dataware can overcome the initial hurdles of adoption and gain widespread acceptance, it will flip the data technology industry on its head. So many of the tools we use today focus on bringing data together from multiple applications, but, with dataware, the data is never separated to begin with.

Even before those broader hurdles are overcome, however, dataware solutions can deliver immense value. In the present environment, dataware platforms enable the creation of robust data fabrics. The data collaboration and active metadata features of a dataware platform help it serve as an intermediary layer between the applications and data stores companies already possess, while simultaneously reducing the need for additional integration down the road. Ultimately, an investment in dataware today is an investment in data fabric that carries a potential upside as the overall approach of separating data from applications gains traction with the developer community.

About Eckerson Group

Wayne Eckerson, a globally-known author, speaker, and consultant, formed Eckerson Group to help organizations get more value from data and analytics. His goal is to provide organizations with expert guidance during every step of their data and analytics journey.

Eckerson Group helps organizations in three ways:

- **Our thought leaders** publish practical, compelling content that keeps data analytics leaders abreast of the latest trends, techniques, and tools in the field.
- **Our consultants** listen carefully, think deeply, and craft tailored solutions that translate business requirements into compelling strategies and solutions.
- **Our advisors** provide one-on-one coaching and mentoring to data leaders and help software vendors develop go-to-market strategies.

Eckerson Group is a global research and consulting firm that focuses solely on data and analytics. Our experts specialize in data governance, self-service analytics, data architecture, data science, data management, and business intelligence.

Our clients say we are hard-working, insightful, and humble. It all stems from our love of data and our desire to help organizations turn insights into action. We are a family of continuous learners, interpreting the world of data and analytics for you.

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

Eckerson Group

RESEARCH CONSULTING EDUCATION

GET • MORE • VALUE • FROM • YOUR • DATA

About the Sponsor

Cinchy is the world's first comprehensive Dataware platform that unlocks data from enterprise apps and connects it together in a universal data network. Built for real-time data collaboration, the Cinchy Dataware Platform addresses the root cause of data fragmentation and data silos, eliminates the need for costly and time-consuming data integration, and mitigates risks of data duplication.

Cinchy simplifies the way organizations connect, control, and collaborate on their massive and growing volume of data by:

> Preventing the need to manage and maintain data for every application

> Sharing data without point-to-point integration or copying of data

> Automating consistent data controls across all applications

With Cinchy, midsize and enterprise organizations gain agility to accelerate digital transformation, reduce the time and cost to build applications by more than 50%, decrease project delivery risks, improve data governance, and enable effortless sharing of quality data across systems and users.

Customers including PwC, TD Bank, YMCA, Colliers International, Investment Quebec, InsuranceGIG, Fiera Capital, Burnco, Reformulary, OmniCable, Rayond James and and Natixis use the Cinchy Dataware Platform to:

> Centralize, manage, and maintain critical data between systems

> Deliver an integrated eCommerce platform

> Enable data controls and governance

> Automate controls, process and workflows

> Gain a 360 view of everything - customers, employees, products, partners

> Give business users a central easy way to access, manage, and update data

> Power analytics and business intelligence

> Replace the use of spreadsheets for critical business processes

> Retire legacy systems and reduce licensing of SaaS applications

> Quickly deliver new solutions aligned to business needs

Request a demo with one of our Dataware experts to see how Cinchy can help your organization