ModernWhitePaper

DataOS® Series

A Paradigm Shift in Data Management 2nd edition

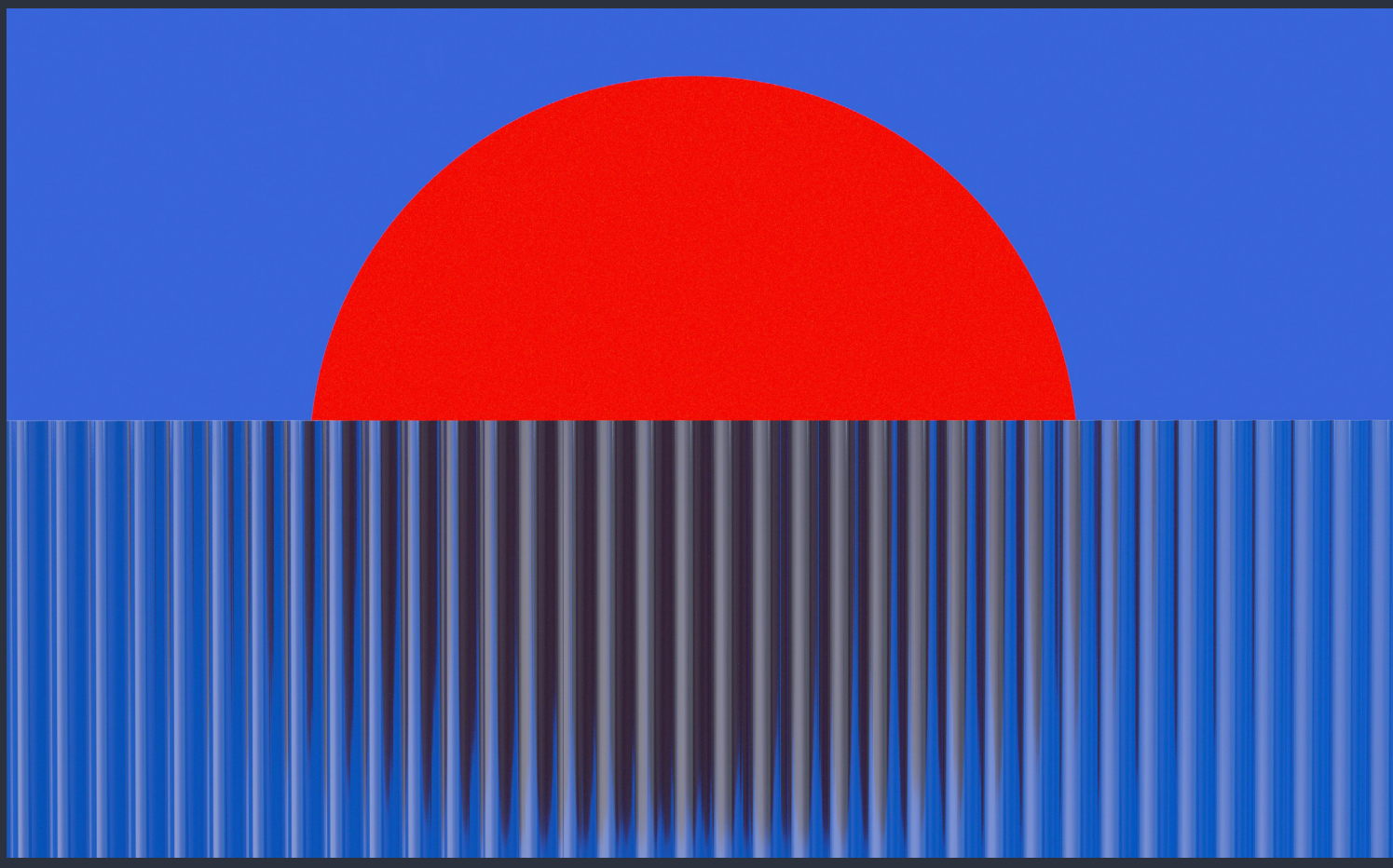


Table of Contents

TheModern DataCompany

Modern White Paper. DataOS® Series. A Paradigm Shift in Data Management. 2nd edition.

The Modern Data Company
306 Cambridge Ave
Palo Alto, CA 94306
TheModernDataCompany.com
info@TMDC.io

Data Needs Are Transforming

Until recently, data access was largely restricted. A central team would prepare reports for management based on structured data, such as customer data, sales data, and financial records. Data was stored in enterprise warehouses, data stores, or marts before analysis. Access to data was mostly limited to IT teams and a select few. Business teams spent inordinate amounts of time gathering and preparing data for analysis — leaving far less time for harnessing data insights.

Over the last two decades, substantial scientific and engineering advances have allowed for vast amounts of data to be transported at blazing speeds. Organizations are now generating unprecedented volumes of data in a variety of formats at ever- increasing rates. Concurrently, hardware and software innovations have driven down the cost of storage and computation, making them increasingly affordable and accessible.

The emergence of cloud computing platforms coincided with a rise in data literacy. This led to the advent of citizen data scientists and an exponential increase in the number of data users. Thousands of humans and machines are now working simultaneously with an organization's data. The era of generating data with Industrial IoT, transmitting it at 5G standard, and making sense of it in real time has truly arrived.

In this rapidly evolving landscape, the shift from traditional data management to a more dynamic and accessible approach has paved the way for the rise of data products. These products represent a transformative step in how organizations interact with, manage, and benefit from their data.

Unlike traditional data storage and analysis methods, data products embody a more agile, user-centric approach, enabling a broader range of users to engage with data in real-time and derive actionable insights. This paradigm shift is not just about technological advancement; it represents a fundamental change in organizational culture and processes, emphasizing data democratization and the empowerment of individuals across various departments.

This paper will explore how data products are redefining the landscape of data analytics, turning raw data into valuable assets that drive decision-making and innovation in the modern data-rich environment.

The Age of Data-Driven Organizations

It is now possible for organizations to harness the power of big data. Beyond reporting, data is empowering predictive analytics and solving for operational needs. For that reason, we have seen the number of data users in an organization grow from less than ten to over 100,000, including both humans and machines. Being data-driven offers a competitive advantage in this day and age. The largest, most successful global organizations leverage data at every step of their decision-making processes.

For organizations to remain competitive, they must capitalize on rapidly growing data by implementing more advanced analytics. They do so by investing heavily into data and analytics infrastructure, machine learning, tools for democratizing data, and improving access to analytics. This has helped spur a wave of tools across the data management ecosystem. These tools are considered “point solutions” because they are primarily built to cater to a specific use case or persona.

Such tools can expand in order to address adjacent commercial opportunities where no tool yet exists. Organizations attempt to stitch these point solutions together in order to create their data infrastructure. These data infrastructures require numerous integrations and a swarm of services to patch data together and keep it updated and maintained for daily operations.

For many, creating a data-driven organization is still a distant dream. Juggling an unwieldy number of point solutions has resulted in data architectures becoming overly rigid. Most non-technology organizations are not getting utility out of their data investments. They often struggle to integrate, govern, process, and syndicate data to external entities in order to generate desperately needed value. Data practitioners and industry observers are dissatisfied. An emerging consensus is that organizations aiming to become data-driven can’t keep up with the vast amounts of data being generated.

The Malady of Rigid Data Architectures

Rigid data architectures lead to significant and tangible adverse impacts rather than living up to their promise. Making data directly available to every user, while still governing it, has become a complex riddle. Getting value out of data is challenging with this sort of overhead. Three problems stand out: long analysis times, data silos/dark data, and data lock-ins.

Long Analysis Times

Data pipelines are foundational — their structure determines how long it takes to deliver everything from a simple dashboard to a predictive operational system. Data requests typically go to the IT department where they are added to a queue behind many such requests. The IT team must then locate the data, decide what is relevant to satisfy the original data request, and assess the quality and computations that need to be done. Then, they vet the data set to make sure that it does not contain any information that the user is not authorized to access. There is usually some back and forth with the user before everyone is satisfied with the resulting data. This all happens before the user starts working with their requested data.

In 2018, Kaggle found that data scientists spend 40% of their time cleaning and organizing the data they receive. Forbes puts that number at 60%, with another 19% spent collecting data sets. Even operational systems suffer from similar limitations related to real-time data.

Data Silos and Dark Data

Understandably, the structure of data ownership within an organization closely mimics its management structure. In most of these organizations, departments seldom collaborate because their systems are not set up to facilitate collaboration. For instance, department A may have data that would be very valuable to department B. However, no one outside department A knows where the data resides, how it is structured, or even that it exists. Additionally, burgeoning amounts of data mean that large chunks of unowned data exist across the entire organization. This is also known as dark data.

Veritas estimates that 50%-55% of the data stored by most companies is dark data. Not knowing the value of such data is a massive opportunity cost—not to mention the literal storage and management costs. Data discovery and governance are so critical that organizations in the business of core technology have simply chosen to create their own proprietary tools. These big tech organizations are the ones setting benchmarks of data-driven behavior.

Data Lock-ins

Rather than rely on standard, open data formats, many vendors choose specialized formats that cannot be used or understood by other systems. To ensure their own survival, many vendors make it hard to move customer data off of their platforms.

The industry needs a complete paradigm shift in its approach to data, data architectures, data experiences, and data democratization. One that appreciates the true owners of the data: the customer and the organizations they grant their data to in exchange for goods and services.

Technology providers should simplify complex data infrastructures while delivering organizational agility and performance, all the while maintaining the role of stewards of customer and organizational data. Providers that achieve this are “worth their weight in data.”

This fundamental shift demands a modern take on governance — a modern take on data engineering as a whole.

Are Data Products the Panacea?

Modern, data-driven organizations are evolving beyond merely safeguarding data in isolated silos. They now view data more dynamically, akin to financial capital and human resources, using it as a vital asset to empower teams and decision-making processes. With the understanding that the value of data has a distinct "expiration date," it's crucial to design and implement processes that harness data at its peak utility. These processes are not just about extracting maximum value but also about adapting to the rapidly changing pace of business and the ever-shortening shelf life of data. In environments where both humans and machines constantly seek data insights, the need for scalable governance frameworks becomes paramount.

In today's landscape, sophisticated systems are needed to manage data pipelines efficiently and in a governed manner, responding automatically to new data requests from humans or machines. These systems are often comprised of various modular components, adhering to standards seen in fully integrated stacks. This approach has given rise to the concept of "data as a product."

A data product is a refined and packaged form of data, designed to deliver specific value to its users. It goes beyond traditional data management and focuses on:

- Creating value-added data services or tools that can be easily used and shared within an organization.
- Utilizing knowledge graphs, semantics, machine learning, and AI to enrich and make data more accessible.
- Supporting rapid and sometimes automated access and sharing of data, tailored to specific operational or analytical use cases.
- Offering flexible deployment options and architectural approaches to fit diverse organizational needs.

The term “data fabric” often enters conversations around data products. Coined around 2016, it appeared in the Gartner Hype Cycle for Data Management 2021 and is often linked with marketing efforts to various products. However, it's important to note that data fabric is not a singular product or technology, but an emerging concept in data management and integration. Similarly, the concept of a data mesh, proposed in 2019, builds on the principles of treating data as a product. It focuses on decentralized data ownership, self-serve data infrastructure, and federated computational governance.

The concept of data fabrics, though innovative, also falls short of being a complete solution. Data fabrics aim to create a unified, integrated layer of data across an organization, but they face their own set of challenges. One significant issue is the complexity involved in integrating various data sources and types across an organization, which can be a daunting task, especially in large enterprises with legacy systems. Furthermore, while data fabrics can facilitate access to data, they do not inherently address the quality or usability of the data. This can lead to scenarios where data is accessible but not necessarily in a form that is immediately valuable or actionable for business purposes.

The implementation of a data fabric requires substantial investment in technology and expertise, which may not be feasible for all organizations. Therefore, while data fabrics contribute to the broader landscape of data management solutions, they are not a panacea for all data challenges and need to be considered as part of a more comprehensive data strategy.

Treating data as a product involves a cultural shift within organizations. It requires stakeholders to act not just as owners but as custodians of data, nurturing, enriching, governing, and sharing their productized data artifacts for the collective benefit of the organization. This shift is a substantial commitment, underscoring the need for a cultural realignment in how technology is utilized to ensure a tangible return on investment.

While we are exploring various avenues to optimize data use, it seems we are still on the journey to find a comprehensive solution for all data-related challenges. Data products, as an evolving concept, offer a promising path but without a comprehensive approach—and the tools that enable it— many companies cannot implement data products to manage the complex world of data management and utilization.

Data Product Platforms: A Paradigm Shift

At Modern, we've pioneered a new approach to data, envisioning a future where data is not just stored and managed, but transformed into dynamic, valuable products. This innovative paradigm relies on loosely coupled, yet tightly integrated building blocks, enabling organizations to craft the precise data architectures they need. This approach is ideal for developing and deploying data products that are both versatile and tailored to specific organizational needs. Such a strategy future-proofs an organization's data infrastructure, providing a composable platform capable of accommodating diverse architectures, users, and systems in unison. The cornerstone of this approach is a data product platform composed of interoperable and composable primitives, services, and modules. Imagine a LEGO set, where the same pieces can be used to construct various end products, from a house to a car to a spaceship. This flexibility and modularity are what make data products an essential component of any forward-thinking data strategy.

Such a platform must fulfill an essential function: streamline processes so that data-driven decisions can be made in near to real time; an experience previously reserved for data-first tech companies. The setup should take days or weeks instead of months or years. To stay worthy of its name, a data product platform must do most of the things that our more familiar operating systems do, and more — provide an enhanced data experience.

1. Present a consistent, DevOps friendly interface to all resources
2. Interface to orchestrate resource allocation for complex scenarios
3. An intuitive and programmable shell
4. Interface for data/resource sharing
5. Governance
    a) Discoverability of assets
    b) Secure access control of assets

Just as computers use a standardized interface to show all files and applications, a data product platform should make it easy to discover all available data assets and to use applications that draw on them. It should treat all legacy data and new data transparently, regardless of how it is stored or formatted. It should be open to new applications, without the patching and fixing that is currently required whenever a new tool is added to a data stack. Everything should be logged so that data managers can see how and when data has changed over time, which systems or processes have touched the data, and which tools are used most.

A data product platform should be usable with minimal training. At the same time, more advanced users who want to get under the hood should be able to use a command-line interface to automate and improve their work. This command-line interface should require only a little more specialized training than the GUI, and it should use a standard command syntax.

One of the most important functions of an operating system is the sharing of data between applications. A good analogy is when an appointment is added through a laptop calendar and is immediately propagated to the calendars on all other connected devices. Similarly, a data operating system should facilitate interoperability between the different applications in the stack.

A proper data product platform should have built-in systems that can be configured to control access to and permissions for all devices on a network. Additionally, it should keep logs of all activity to be reviewed as needed. It should secure the data in the local environment.

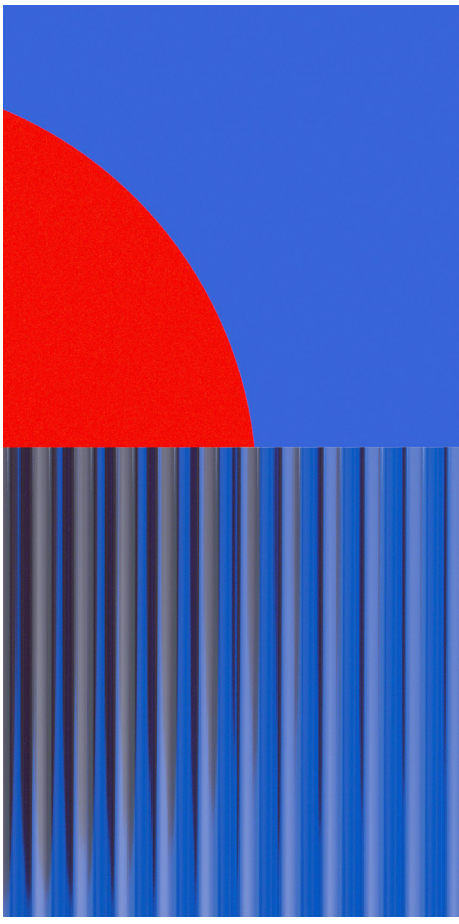
The Value of DataOS®

DataOS delivers tangible value across many use cases:

- Lower OpEx costs.
- System integrators can optimize the work of their engineers.
- Build out solutions easily without an expert talent pool.
- Build data products faster with superior data experience — in hours or weeks instead of quarters or years.
- Drive new revenue models by sharing or collaborating with secure, governed data.
- Focus on value creation from data instead of integration and process work.
- Modernize your infrastructure.
- Democratize access to data and insights.
- Strengthen security and privacy controls for all of your data.
- Dictate the insights you need for your business, instead of the data dictating the insights you get.

The Modern Data Company

DataOS® from The Modern Data Company is a groundbreaking data operating system that radically simplifies the intricacies of data. Treating data as software, it empowers organizations to build comprehensive data products that drive outcome-based, trusted decisions. Seamlessly integrating with any existing data architecture, DataOS® provides a unified and composable data ecosystem. This facilitates the smooth integration and operationalization of data products at scale, ensuring business processes remain uninterrupted. It enables cross-organizational data collaboration that is not only safer but also more efficient than ever before. Dive into the transformative world of data products to future-proof your data initiatives.



The Modern Data Company
306 Cambridge Ave
Palo Alto, CA 94306
TheModernDataCompany.com
info@TMDC.io

TheModern DataCompany

Modern White Paper. DataOS® Series. A Paradigm Shift in Data Management. 2nd edition.

E22D