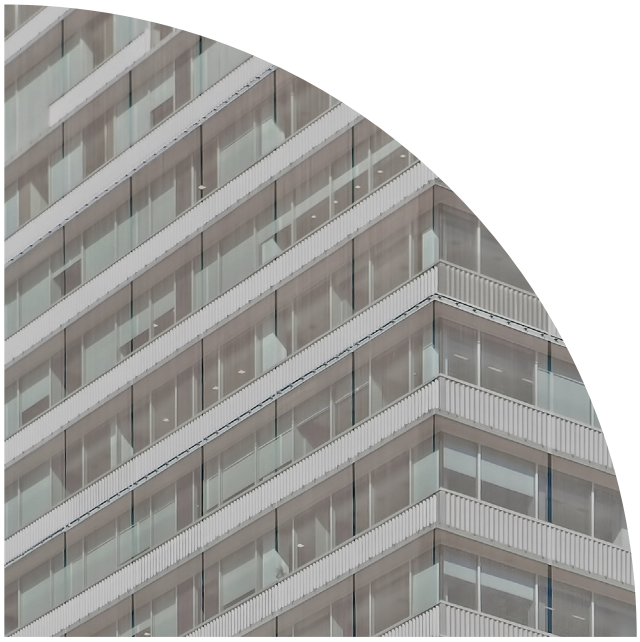
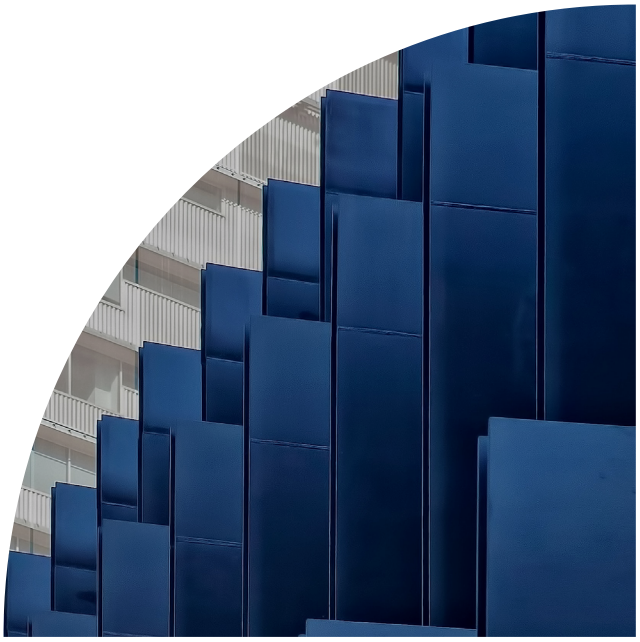
DataOS®

A Paradigm Shift in Data Management

Deploying A Data Fabric

ModernWhitePaper





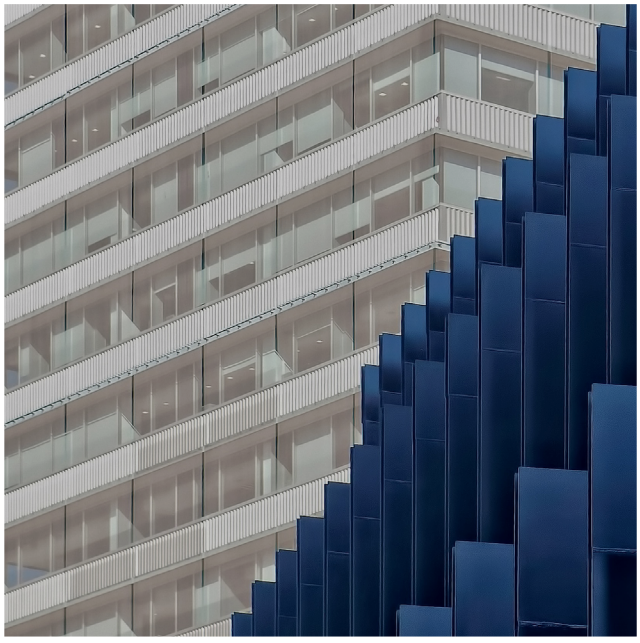


Table of Contents

TheModern DataCompany

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

Data Needs Are Changing in A Transformative Way

Until recently, the use of data was largely restricted to a central team who would prepare reports for management based on structured data such as customer data, sales data, financial records, and other data sources. Typically, everything was stored in enterprise warehouses, data stores, or marts before analysis. Data access was limited to the IT teams and a select few others. This meant heavy support requirements from IT, and every minute spent by IT just enabling the day-to-day distracted from work to continue to evolve the underlying infrastructure and technologies.

However, over the last two decades, there have been substantial scientific and engineering advances that allow the generation of vast amounts of data transported at blazing speeds. Organizations are now generating unprecedented volumes of data in a variety of formats at ever-increasing rates. At the same time, hardware and software innovations have driven the cost of storage and computation down, making them affordable and accessible by any organization. This has enabled architecture and infrastructure organizations to move more quickly and cost-effectively than in the past and helped to spur additional innovation within the end user base.

The emergence of cloud computing platforms coincided with the rise of data savvy. This led to the rise of citizen data scientists, and the number of data users multiplied exponentially. This posed challenges not just to the scalability of the underlying platforms, but also to the governance and security protocols that IT is required to maintain. New data and new use cases led to the need for ongoing assessment and tuning of governance and security standards.

This is one of a series of DataOS white papers that are available. One is a broad overview of DataOS while this and several others dive into topics of interest to specific types of users. This paper will focus on how architecture and infrastructure teams can get the most from DataOS.

Organizations Are Investing Heavily to Become Data-Driven

It is now possible for organizations to harness the power of big data, not just for reporting but for predictive analytics and solving operational needs. As a result, within an organization, we have seen the number of users of a data platform grow exponentially. Today, being data-driven is a competitive advantage. However, to be data-driven requires more support from the architecture team than ever before given the complex infrastructures and analytical requirements involved in achieving data-driven success. All of the largest and most successful global organizations leverage data in every step of their decision-making process.

Organizations compete fiercely on analytics and data. Their only recourse for capitalizing on rapidly growing data is to double down on more advanced analytics. They invest heavily into data and analytics infrastructure, machine learning, tools for democratizing data, and 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. The downsides to point solutions are that IT must support more and more solutions, and that there are greater risks of data leaks or security breaches. Data infrastructures full of point solutions require numerous integrations and lots of services to patch everything together and to keep updated and maintained for daily operations. Ensuring consistent governance and security becomes a difficult and time-consuming process.

Creating a data-driven organization is still a distant dream. Increasingly, as a result of incorporating point solutions, data architectures have become far too rigid. Most non-technology organizations are not getting value out of their data investments. They often struggle to integrate, govern, process, and syndicate data to external entities in order to generate the value they desperately need. Data practitioners and industry observers are dissatisfied. An emerging consensus is that organizations looking to be data-driven are falling behind the vast amount of data being generated. The pressure is on the architecture team to find a way to help the organization modernize its infrastructure and increase capabilities in a cost-effective way.

The Malady of Rigid Data Architectures

The rigidity of current data architectures means that they cause 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: inconsistent security and governance protocols, data silos/dark data, and data lock-ins.

Lack Of Flexibility

The pace of change in the technology realm is unprecedented. No sooner is an architecture implemented than one or more components are outdated as newer, better options become available. Companies can't afford to totally overhaul and redesign their entire architectures on an ongoing basis. At the same time, they can't fail to adapt to newer technologies to avoid falling behind. Some type of middle ground must be found.

What's needed is a flexible and composable platform that allows for individual components within the architecture to be changed without having to replace the others. This is done by creating connectors to each component within the architecture and allowing them to communicate with the rest of the components through a centralized intermediary.

Data Silos and Dark Data

Understandably, the structure of data ownership in an organization closely mimics its management structure. In most of these organizations, different 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, but no one outside department A knows where the data resides, how it is structured, or even that it exists. Apart from this, large chunks of data exist in any organization that are currently not owned by anyone across the entire organization (dark data).

Veritas estimates that 50%-55% of the data stored by most companies is dark data. There is a massive opportunity cost for not knowing the possible value of such data, not to mention its storage and management costs. The architecture must enable the identification, dissemination, and analysis of all of this data. 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 for 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. Vendors, for existential reasons, make it hard to move customer data off of their platforms. As gaps in needed functionality appear in the platforms, a rigid architecture provides limited options to address and close those gaps.

What is needed is a complete paradigm shift in thinking about data, data architectures, data experiences, data democratization and much more––one that contemplates the true owner of the data (the customer and the organizations that provide goods and services in exchange for which the customer volunteered their data).

Technology providers that solve the complexity of a data infrastructure while delivering organizational agility and performance and maintaining the role of stewards of customer and organizational data are "worth their weight in data." The more that users are able to bring their preferred toolsets to bear and have a wide range of functionality to take advantage of, the more productive they'll be. This fundamental shift demands a modern take on governance and a modern take on data engineering as a whole.

Are Data Fabrics the Panacea?

Modern organizations that are data driven have not become so by creating an ecosystem where data is treated as an asset protected within data prisons. Similar to financial and human capital, data is used as a form of capital (data capital) to empower teams. These teams can design processes that best leverage the shelf life of any data's value. This is critical, as any analysis done post this shelf life is fairly obsolete. With the tide of data being generated and peer organizations pushing the envelope, the shelf life of data is growing shorter. Moreover, with hundreds of thousands of humans and machines seeking data, organizations need a scalable governance framework. Of course, it typically falls on the infrastructure team to figure out how to implement systems to make it all work.

Modern organizations have systems that can deliver data capital to their humans and their machines at such a gold standard. These systems render data pipelines automatically and in a governed fashion whenever a new data request is generated by a human or a machine. Such systems are engineered with various modular components that use the same standards as fully integrated stacks. Such design patterns have inspired the term "data fabric," and they are being built into corporate architectures at a rapid pace.

The term data fabric, a contemporary design pattern, was coined around 2016 and is atop the "Peak of Inflated Expectations" in the Gartner Hype Cycle for Data Management 2021. A lot of marketing verbiage attempts to link products with the term. Sometimes, it may read as if "Data Fabric" is a single technology that underlies many products. However, data fabric is not a single technology, but a design concept, according to Gartner. In other words, an IT or infrastructure team can't simply "buy a data fabric." Implementing a data fabric is instead a process of integrating both existing and new infrastructure components in a modern fashion.

A data fabric is also an emerging data management and data integration design concept:

- **for attaining** flexible, reusable, and augmented data integration pipelines
- **that utilizes** knowledge graphs, semantics, and ML/AI on active metadata
- **in support of** faster and, in some cases, automated data access and sharing
- **regardless of** deployment options, use cases (operational or analytical) and/or architectural approaches

In a similar vein, the concept of a "data mesh" was proposed by Zhamak Dehghani in 2019, built on four principles:

- Domain-oriented decentralized data ownership and architecture
- Data as a product
- Self-serve data infrastructure as a platform
- Federated computational governance

Data mesh is also a design concept. Both data fabric and data mesh are deeply advocated for in their respective forums. Both involve not only a technical but a "cultural" re-alignment as well, requiring companies to embrace a novel approach to using technology. For instance, one of the assumptions underlying data mesh is to treat data as product with no owners but rather custodians who would nurture, enrich, govern, and share their respective productized data artifacts with the rest of the organization. This would imply a significant cultural commitment to make sure that the investment in technology delivers a real ROI.

It appears that we may be far from an all-encompassing panacea for all things data. Even if the architecture team designs a powerful, modern environment, it still requires organizational buy-in and culture change to make things work.

Data Operating Systems – A Paradigm Shift

At Modern, we created a new way of thinking about data. This new paradigm creates a future where loosely coupled, tightly integrated building blocks enable organizations to compose the data architectures that they need. This would serve to deploy a data fabric or a data mesh––or an even more novel design in future––and future-proof an organization's data infrastructure needs with a composable platform that can accommodate all architectures, users, and systems simultaneously and includes both cloud and hybrid environments. This would be a huge advantage for an organization: to know that its underlying platforms and architectures can evolve without requiring massive effort.

The key to this paradigm shift is an operating system that consists of a set of primitives, services, and modules that are interoperable and composable. A good way to understand this would be a box of LEGO bricks where the same components can be used to construct either a car or a bus. There are LEGO bricks for the architectural, platform, tool, and connectivity components of a company's systems.

Such an operating system must fulfill the most essential function of any operating system: making users more productive so that organizations can attain the data-driven decision-making experience now reserved for data-first tech companies––and in days and weeks instead of months and years. To stay worthy of its name, a data operating system must do most of the things that our more familiar operating systems do––and more––to provide an enhanced data experience. Furthermore, a data operating system must support and make it easy for any architectural design to be implemented.

1. Present a consistent, devops-friendly interface to all resources
2. Work with current tools and infrastructure while being able to adapt to changes in architecture in the future
3. Support for cloud and hybrid environments, including scalable and on-demand provisioning
4. Interface to orchestrate resource allocation for complex scenarios
5. An intuitive and programmable shell
6. Requires less support and enables faster, less expensive implementations
7. Governance enabling the discoverability, encryption, and secure access control of assets

Just as computers use a standardized interface to show all files and applications, with a data operating system it should be easy to discover all available data assets and to use applications that utilize 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 and system administrators 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 operating system should be usable with minimal training. 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. In other words, the data operating system should be as easy for IT to use and manage as it is for end users.

One of the most important functions of an operating system is the sharing of data between applications. When an appointment is added through a laptop calendar, it immediately propagates to the calendars on all other connected devices. A user's contacts are usable on all their devices, whether to send email, text, or call. Similarly, a data operating system should facilitate interoperability between the different apps in the stack.

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

A Data Operating System Exists

DataOS® from The Modern Data Company, does all of these things across the entire data stack. Data architectures can now be deployed in weeks rather than months or years. DataOS is built to set organizations on the fastest path from data to decisions in a truly risk-free manner. It is also built to allow flexibility in how an architecture is designed today while enabling it to adapt to new requirements in the future.

DataOS does this by delivering a composable and agile data operating system that democratizes access to high quality, governed, and secure data in real-time. DataOS connects all of your structured, semi-structured and unstructured data assets across the enterprise and builds an intelligent semantic layer that enables business and technical users to discover, explore, and collaborate to deliver data products in days and weeks instead of months and years.

The unmatched composability of DataOS lets customers adapt it to any data architecture, be it a data fabric, data mesh, lake house, or something new. It also supports a wide range of toolsets, which means that there is no need to make the unpopular decision of telling users that they need to decommission some of their favored tools.

Speed and agility are instrumental to DataOS because it integrates the entire data engineering layer. This eliminates all of the point solutions that are required today and further eliminates all of the integration required to maintain such point solutions. This makes DataOS secure, scalable, and adaptable.

There are four overarching benefits that differentiate DataOS from anything else in the market today:

- **Composable, extensible, open and interoperable**
  - DataOS is completely composable and lets customers adapt it to any architecture or use case of their choice.
  - Agility in the infrastructure lets organizations react and adapt to new workloads and requirements without requiring major IT work.

- **Enhancing—not replacing––existing infrastructure**
  - Augment the functionality and ROI of your current investments.
  - There is no pressure to replace architectural components already in use.

- **All things data under one roof**
  - Building data products has never been faster or easier—weeks/hours instead of quarters/years.
  - Integrated architecture drives cost optimizations—Lower OpEx.

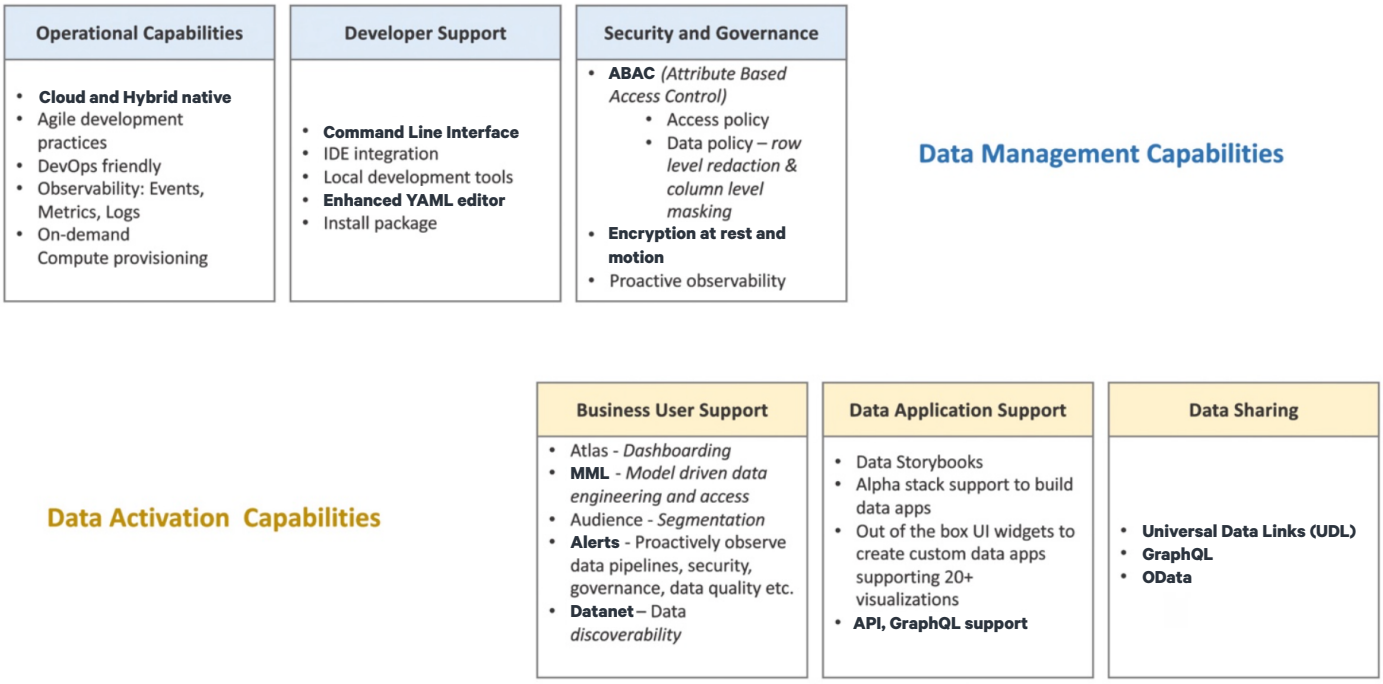
- **Model-driven Data Management**
  - Streamline data pipelines.
  - Prioritizes business-driven use and management of data.
  - Increases IT capacity and security.

With these advantages, DataOS delivers tangible value across many use cases:

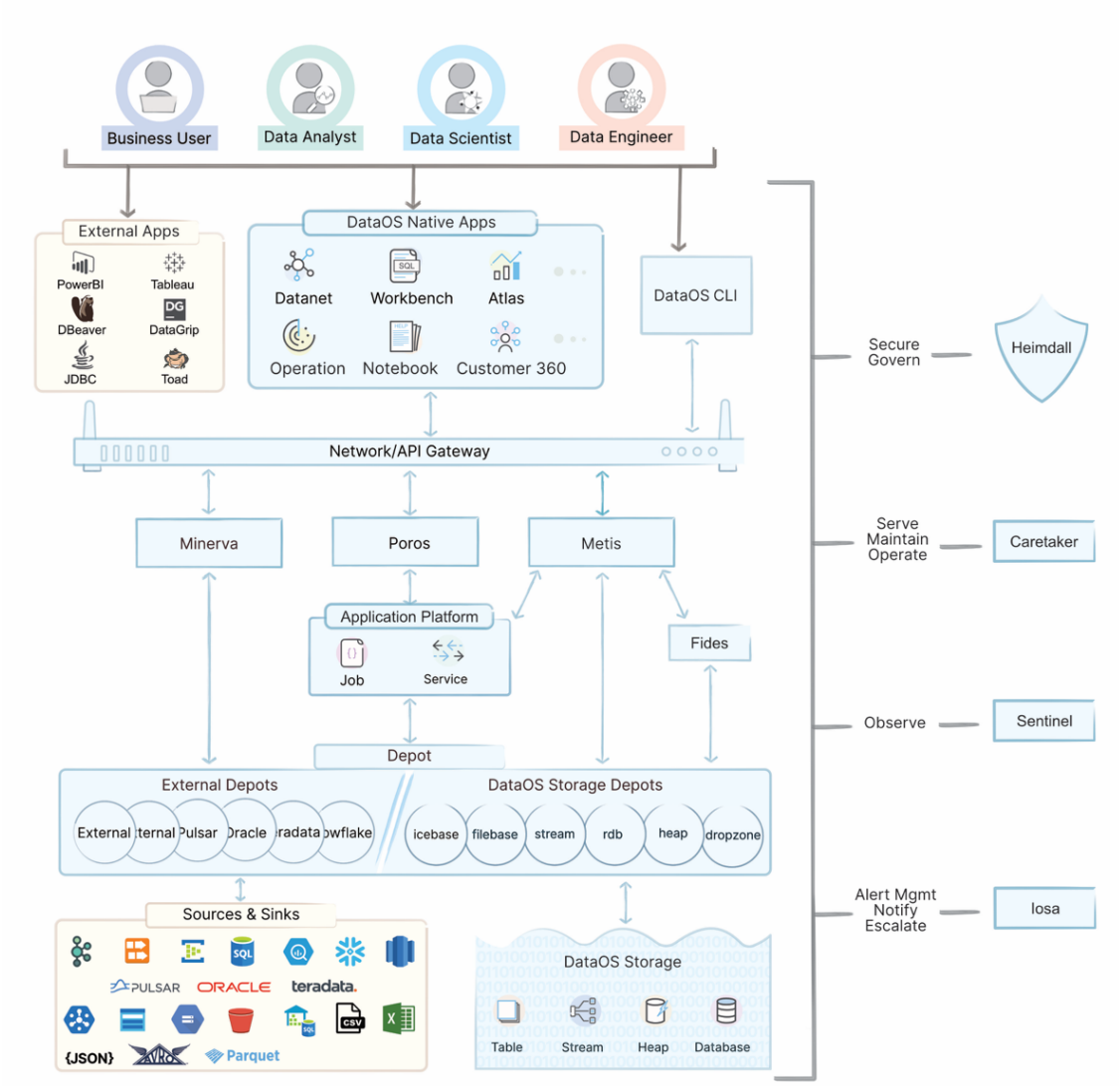
- Support any current or future architecture, including both cloud- and hybrid-based.
- Modernize your infrastructure, with flexibility to easily adapt in the future.
- Lower OpEx and architectural implementation costs.
- Built-in devops and administrative features simplify management of DataOS.
- Focus on value creation from data instead of integration and process work.
- System integrators can optimize the work of their engineers.
- Strengthening security and privacy controls for ALL of your data.

DataOS Capabilities Overview

There is a set of DataOS-specific abstractions and applications that are important in order to fully comprehend the DataOS.



One of the most important abstractions to understand is a data depot. A depot provides a reference to the data source/sink and abstracts the details of their configurations and storage formats. It helps a user easily understand the data source/sink, connect with it, and use the depot to access, process, and explore data within the DataOS. A depot contains the data location information along with any credentials and secrets that are required to access data from an external source. This source could be anything—object storage, warehouses, streaming platform, files or JDBC/ODBC connections, etc. Once a depot is created, users can use it in any jobs or services and in other tools within the DataOS system.



DataOS® from The Modern Data Company, Based on open standards, DataOS offers native applications such as Datanet, which lets a user visualize data and data processes across the organization.

- Search datasets, executables and dashboards––In a single context view, see which users are working with data, what they are doing with it, and how often it is accessed.
- Fabric—knowledge graph of the organization
- Create and manage access, masking, and filter policies
- Produce a list of depots
- Manage tags in the system

Data Governance

While DataOS makes discovery through the entire organization seamless, it balances democratization with a best-in-class governance framework. DataOS enables users to define clear policies for data usage and authorized access, and to meet regulatory compliance requirements such as GDPR. This process encompasses the people, process, and technology that are required to ensure that data is fit for its intended purpose. Data governance deals with the following:

Data Ownership

- The DataOS data catalog engine (known as Metis) catalogs all data sources, data streams, data sets, and data sinks and tracks their ownership over time. This helps an IT organization keep on top of all the data assets present in the organization.
- Transitive ownerships will be automatically applied from the Jobs responsible for the Dataset. So, If James runs a job called “enrich_txn_with_product,” which creates a data stream called “enriched_txn_ with_product,” then James automatically becomes one of the owners for the new data stream.
- Ownership can also be manually curated through the DataOS GUI, which means that when IT determines that changes from the default are required, it is easy to make the necessary adjustments.

Data Quality

- Users can use Metis to learn about their own data and its landscape, such as quality, profile, lineage, dictionary, and much more.

Data Access

- Metadata of all datasets will be available for anyone in the organization to consume.
- DataOS will restrict access only at the data set level. You can either read the data set or not.
- You can always create new data sets with filtered columns and grant separate access to them.

There are two types of policies in DataOS:

- Access policy – the security measure to regulate who can view, use, or access a restricted DataOS environment/resource. It is implemented using an Attribute Based Access Control (ABAC) paradigm versus the archaic Role Based Access Control (RBAC).
- Data policy – a collection of statements that describe the rules controlling the integrity, security, quality, and use of data during its lifecycle and state change. For example, a data masking policy defines the logic that replaces (masks) sensitive data with fictitious data to maintain privacy.

Data Governance Continued

DataOS uses a completely configurable set of tags and attributes to set granular permissions using column-level masking and row-level redactions. These functions can be controlled, changed, and updated easily with a line or two of code. The output of a data query is automatically hashed or masked to fit the permissions of the individual user making the query. Once an individual's permissions are set, the operating system applies them to every query. Illustrating the power of this, for an organization that had more than 1,000 roles, DataOS was able to create the entire governance framework with just 15 policies.

Thus, governance policies can be applied consistently and immediately. Data requests no longer have to be routed through IT. Every user can run their own queries and make their own reports, with security handled transparently. Of course, DataOS has the tools necessary for IT to carefully monitor all activities.

DataOS is controlled at the command line by standardized YAML directives using a configurable set of primitives. This allows very fine-grained control of every function, from governance to data enrichment. But most functions can also be done using a standard point-and-click graphical user interface. Users need little or no coding or markup language for most operations.

Through the Workbench, DataOS allows SQL queries on databases, then provisions data pipelines and workflows that do anything with the results—from exporting them as CSV to sending them to any other tool in the regular data stack. Queries can be saved, shared, and modified by other users and the results presented through any of several built-in dashboards––or users can design their own.

More importantly, DataOS works with any type of data, no matter how it is formatted or stored. Using schema-on-read, the system works as easily with PostgreSQL as it does with Redshift and can pass the data (with or without additional processing) to any analytics tool. Like any other process in the operating system, this can be set up to run automatically based on an event such as new sensor data coming into an ingestion database. DataOS brings every tool in an organization's data stack together seamlessly.

DataOS also applies and stores rich metadata, allowing users to view a complete data lineage, beginning with the original data source and including every process that has been run on it since it was ingested.

Sample Use Case: Creating a Data Fabric with DataOS

Gartner recommends that organizations start investing in components of a data fabric now. The IT and infrastructure teams must lead that charge. A data fabric requires the following components, all of which are native to DataOS:

- semantic knowledge graphs
- active metadata management
- embedded machine learning

An organization can then take the following journey using DataOS:

- Deploying an augmented data catalog to access and represent all metadata, both active and passive
- Adapting knowledge graphs in order to advance metadata discovery
- Enabling the forming data fabric to collect, share, and analyze all forms of metadata over the connected knowledge graph to activate metadata
- Setting up machine learning models enriched with active metadata in order to simplify and automate data integration design
- Allowing IT to provide freedom to users and to automate many administrative tasks while retaining the ability to monitor and administrate all aspects of the data fabric
- Setting up automated data orchestration services

Summing up: Future proof with DataOS, the fastest path from Data to Decisions

Going forward, organizations must drive their decisions and operations based on data. In pursuit of this, the key to future-proof investments is to opt for a true data operating system like DataOS that can be molded to a design pattern of your choice.

DataOS is a paradigm-changing approach to set organizations on the fastest path from data to decisions. It is a composable operating system that forms a connective tissue between all your data resources, operationalizing data so users have access to quality, governed, and secure data in real time. It is a next-gen data stack that also augments your existing infrastructure so that building and delivering data products is faster and simpler for all users. It is the data operating system delivering democratized access to quality data right when you need it.

If you want this type of future-proofing and to think the Silicon Valley way of running data within your organization:

Think Agile
Think Composable
Think Data
Think Modern

Addendum Feature List

The Activation Layer

a. **Data Sharing (Universal Data Link) - Share data with users in a web address fashion with right governance policies**
b. SQL Workbench - powerful workbench that enables a no-code option to querying datasets of all formats.
c. Atlas - Quick BI. Create customized reports on any datasets with visualizations for different stakeholders
d. Data storybooks - Organize rich text, queries, and charts into storybook to easily document your analyses
e. Jupyter hub - preconfigured data science environment
f. **MML - Abstract away the complexity of data engineering with our model-based approach to building pipelines**
g. **On-demand compute clusters - Run sophisticated workloads with compute resources on demand. Isolate resources for different tasks/teams.**
h. **ODATA - Support for open data protocol for building and consuming restful APIs**
i. **GraphQL interface makes building data applications faster and easy**

The Knowledge Layer

a. **Re-usable and version-controlled artifacts**
b. Data Observability – monitor data pipelines and performance. Enhance data reliability by proactively reducing down time.
c. **Security & Governance**
    i. **ABAC - Tag-based governance enables flexible and granular policy creation that adapts to changing or new compliance regulations**
        1. **Policies that let you define access based on tag attributes**
        2. **Policies to define what data can be seen with row-level filtering and column masking**
    iii. **Federated authentication with support for Active Directory and Auth0**
    iv. **Encryption at rest and motion**

d. Data Discoverability
    i. **Get a holistic view of all the data, their relationships, and usage.**
    ii. **Lineage – track movement of data from source systems, understand transformations in detail**
    iii. Impact analysis – identify all downstream consumers like datasets, dashboards, models
    iv. Usage metrics
        1. Frequently queried by users
        2. Frequently used with datasets
    iii. **Fingerprints - Auto classification of PII information**

e. Quality
    i. Automated data profiling
    ii. Out the box data validation assertions, first-class support for custom assertions
    iii. **Incident management with alerting mechanisms**
    iv. Define SLAs for datasets, jobs and workflows

DataOS Feature List Continued

The Data Layer

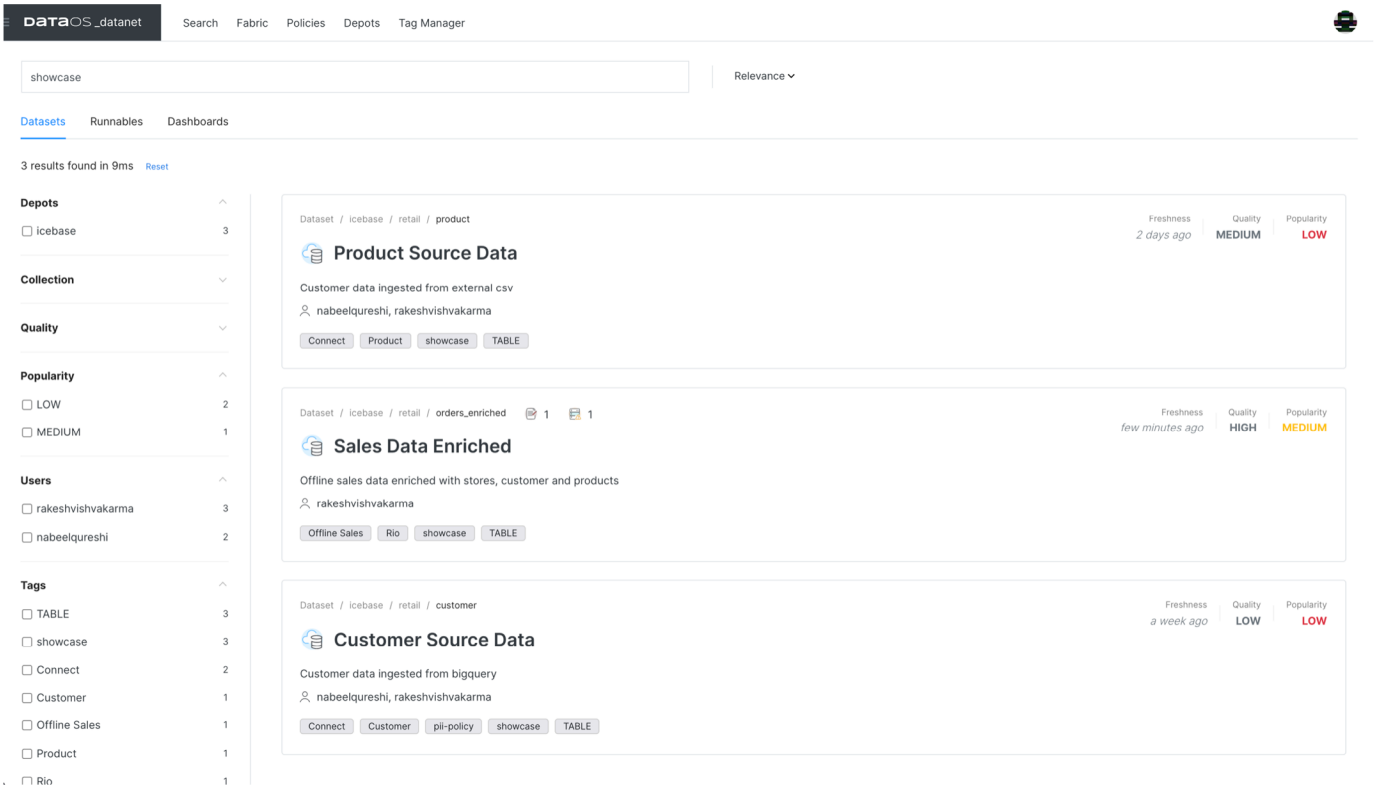
a. Data interface to access data stored across relational, non-relational, streaming and object data sources
b. ACID compliant data lake
c. Schema evolution
d. Time travel - Data Replay
e. Orchestration engine to deploy, monitor, schedule, or trigger sophisticated data pipelines
f. Data transformation
    i. Support for both batch and streaming data
    ii. Horizontal auto-scaling
    iii. Stateful and stateless stream processing engine
    iv. Deploy custom data apps written in python, java, scala etc.
    v. Rest APIs on demand
g. Decoupled storage and compute

Data & Dev Ops

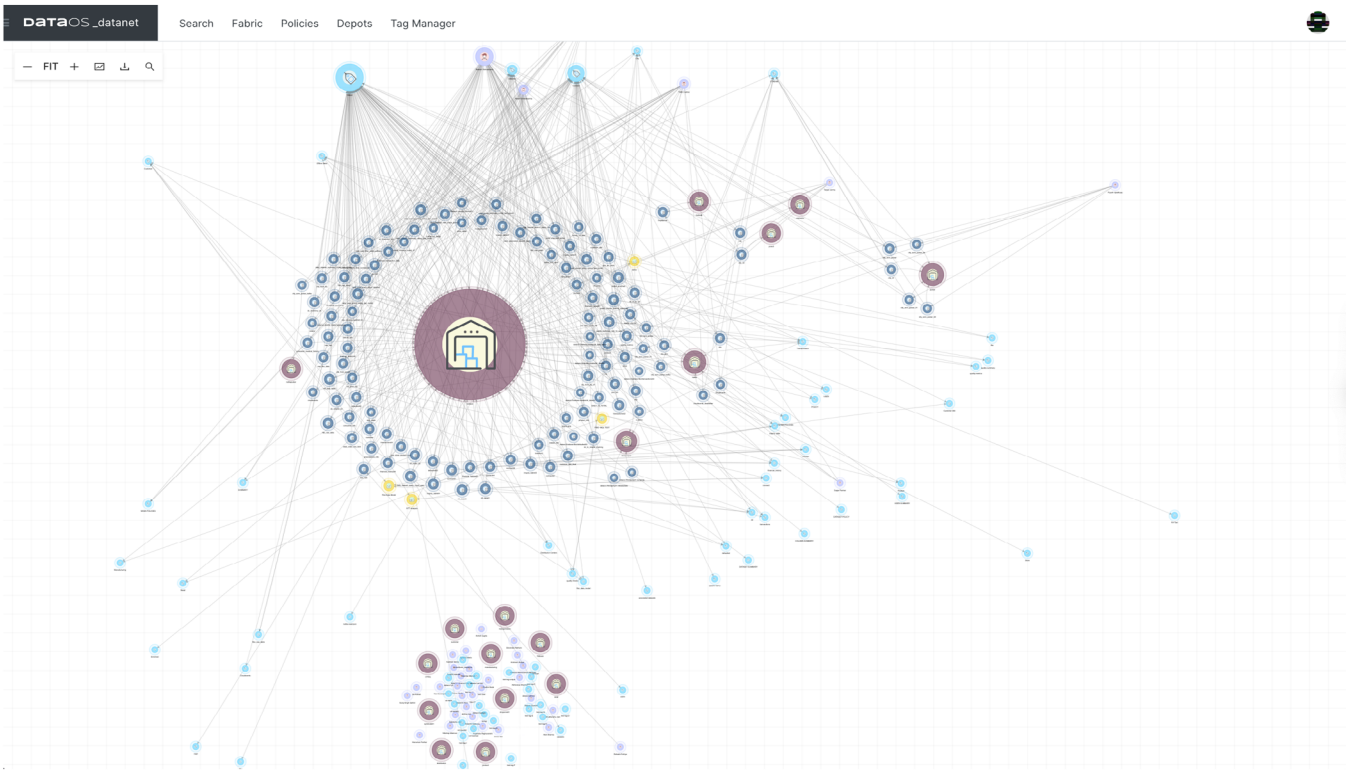
a. Cloud and hybrid native
b. Agile development practices
c. DevOps friendly
    i. Version control
    ii. CI/CD
d. Infrastructure observability to track and audit events, metrics and logs
e. On-demand Compute provisioning
f. Command Line interface
g. IDE integration
h. Local development tools
i. Enhanced YAML editor
j. Install package

DataOS Screenshots

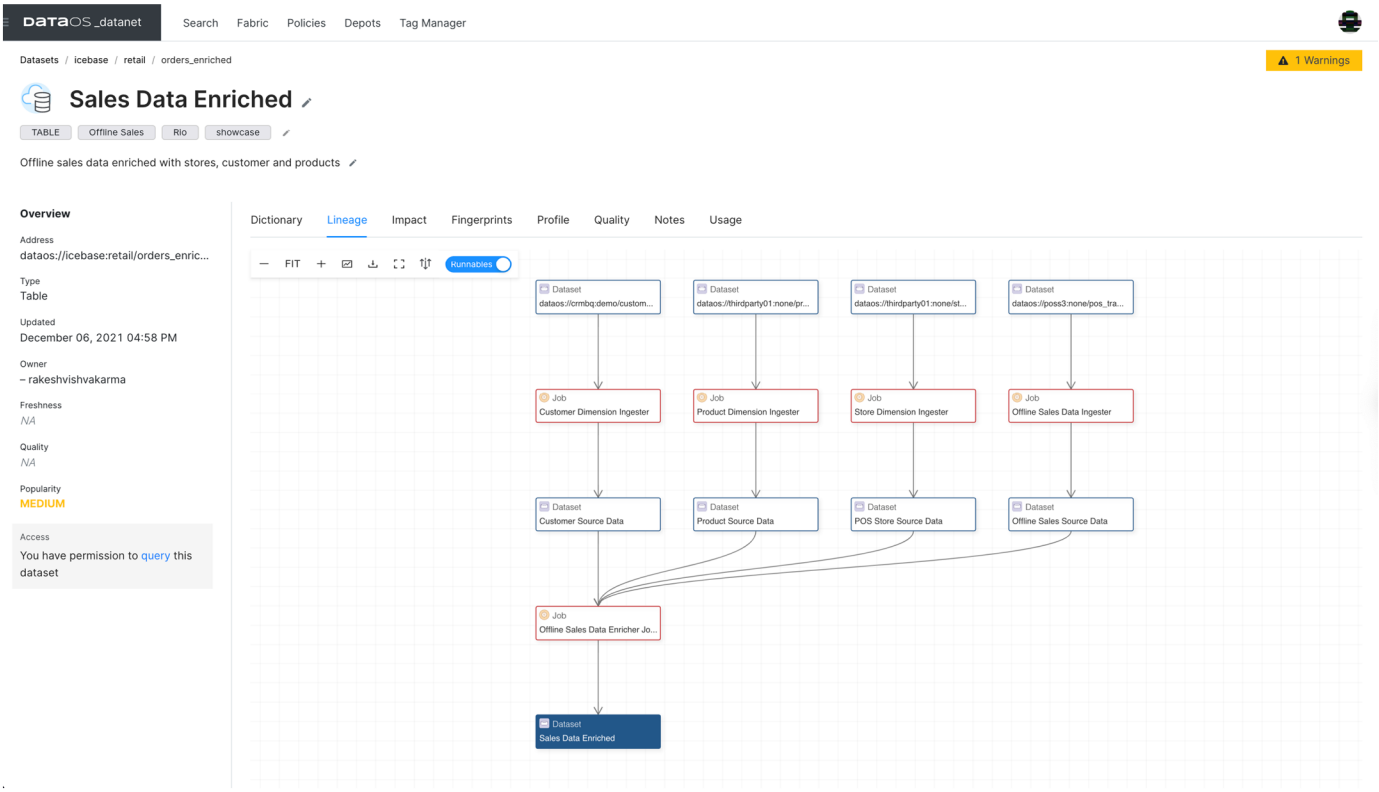
Datanet - Search



Datanet - Fabric



Lineage



About DataOS®

DataOS® is an operating system that consists of a set of primitives, services, and modules that are interoperable and composable. These building blocks enable organizations to compose various data architectures and dramatically reduce integrations. Enterprises can have the same data-driven decision-making experience akin to data-first tech companies in days and weeks instead of months and years.

About The Modern Data Company

Founded in 2018, The Modern Data Company began with the realization that enterprise-wide data access has been siloed. Data engineers and database administrators have been the longstanding data gatekeepers who funneled data to analysts and data scientists. We aim to change that by freeing enterprises to make better data driven decisions by democratizing access to data. When all employees, irrespective of their technical skills or background, can easily explore and analyze enterprise data, then both productivity and market expansion are realized at a faster pace.

TheModern DataCompany

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

DataOS®: A Paradigm Shift in Data Management – Deploying A Data Fabric

G27M