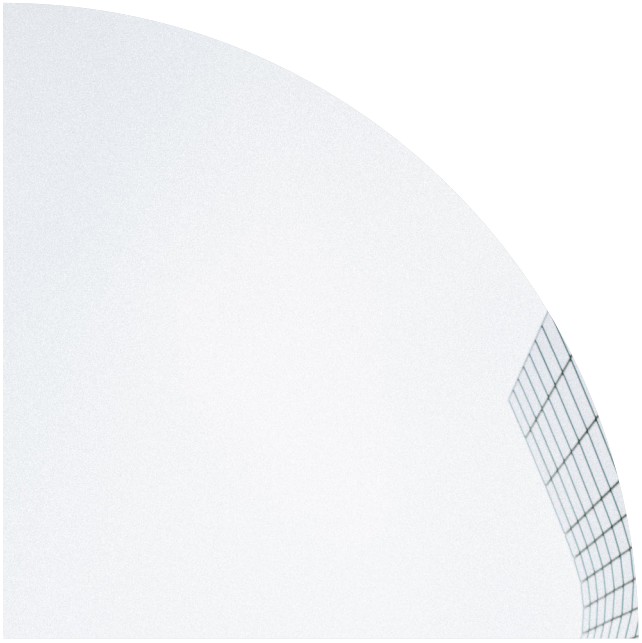
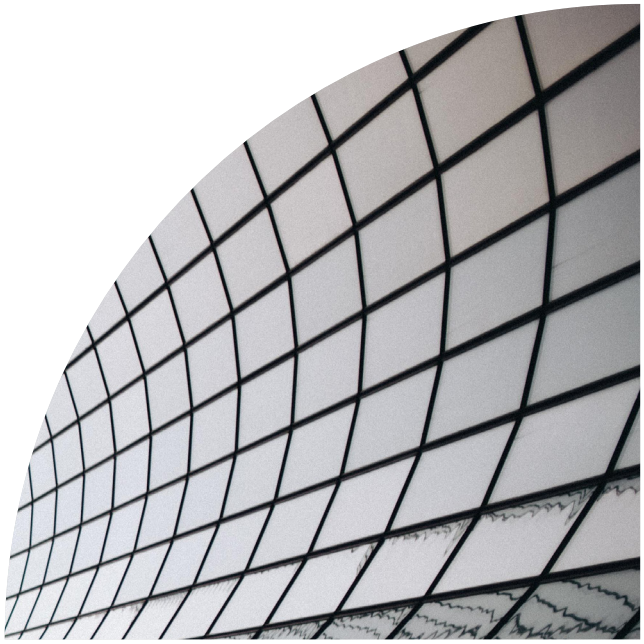
DataOS®

A Paradigm Shift in Data Management

ModernWhitePaper





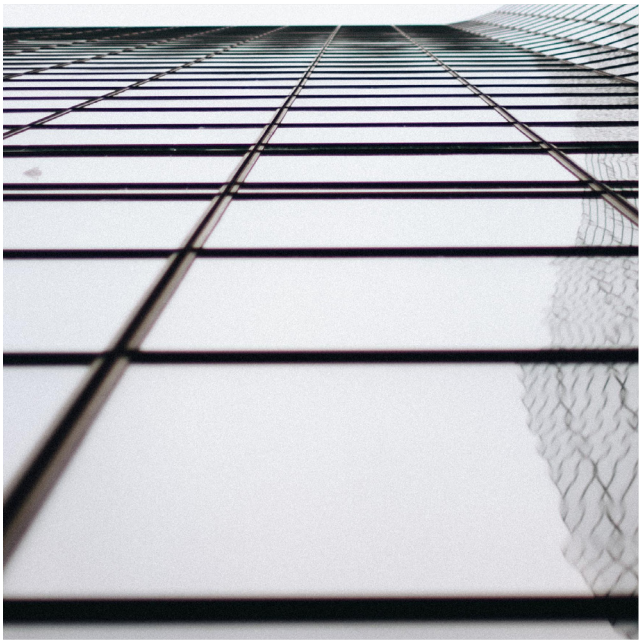


Table of Contents

Data Needs are Transforming 3

The Age of Data Driven Organizations 4

The Malady of Rigid Data Architectures 5

Are Data Fabrics the Panacea? 6

Data Operating Systems - A Paradigm Shift 7

Genesis of the Data Operating System 8

DataOS Capabilities Overview 9

Advanced Data Governance 12

Sample Use Case: Creating a Data Fabric with DataOS 14

Addendum: DataOS Features List 15

TheModern DataCompany

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.

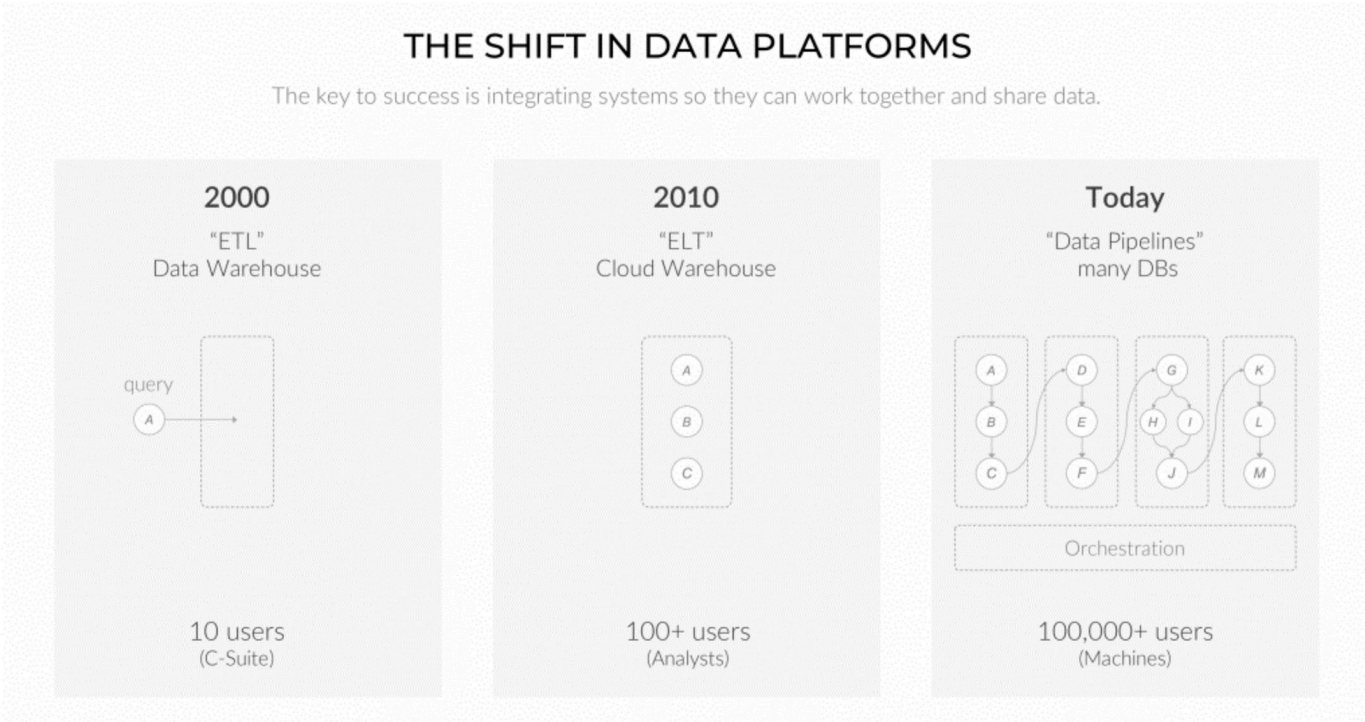
The emergence of cloud computing platforms coincided with a rise in data literacy. This led to the advent of citizen data scientists and an exponential increase in the number of data users. Thousands of humans and machines are now working simultaneously with an organization's data. The era of generating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

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.



Source: https://www.intermix.io/

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 Fabrics the Panacea?

Modern, data-driven organizations create an ecosystem where data can move freely rather treating it like an asset to be guarded in a data prison. Similar to financial capital and human capital, data itself is used as a form of capital to empower teams. All data value has a unique expiration date; teams can design and implement processes that optimize the use of data at its peak utility. These processes are critical to ensuring that the greatest value can be extracted. Faster moving organizations and a continuous flow of information mean that data's shelf life is growing shorter. Moreover, with hundreds of thousands of humans and machines seeking data, organizations need a scalable governance framework.

Modern systems render optimized data pipelines in a governed fashion automatically whenever a new data request is generated by human or machine. Such systems are engineered with various modular components that use the same standards as fully integrated stacks. Such contemporary design patterns have inspired the term data fabric.

The term "data fabric" was coined around 2016 and is atop the "Peak of Inflated Expectations" in the Gartner Hype Cycle for Data Management 2021. Much marketing verbiage attempts to link this term to products.

However, Data Fabric is not a singular product or technology, but rather, 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 and 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 in their respective forums. Both involve not only a technical but also a "cultural" re-alignment, requiring companies to embrace a novel approach to using technology. One of the assumptions underlying data mesh is to treat data as a product with no owners of data; instead, data owners are custodians who nurture, enrich, govern, and share their respective productized data artifacts with the rest of the organization. This requires a significant cultural commitment to ensure that the investment in technology delivers a real return on investment.

It appears we may be far from an all-encompassing panacea for all things data.

Data Operating Systems - A Paradigm Shift

At Modern, we've 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 they need. This would serve to deploy a data fabric or a data mesh — or an even more novel design in future. This would future-proof an organization's data infrastructure needs with a composable platform that can accommodate all architectures, users, and systems simultaneously. The key is a data 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 LEGO set where the same components can be used to construct either a car or a bus.

Such a data operating system 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 operating system 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 operating system 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 operating system 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 operating system 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.

Genesis of the Data Operating System

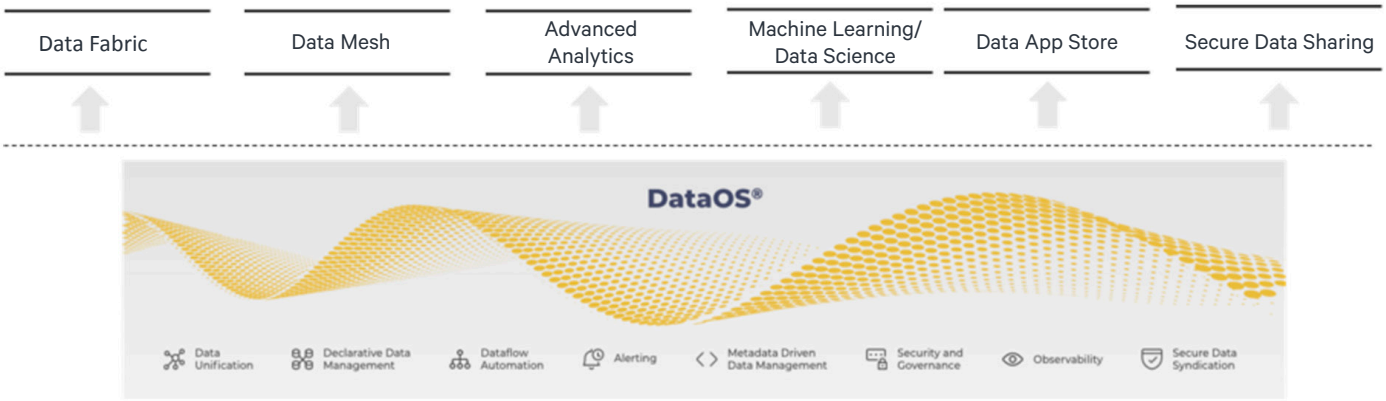
DataOS® from The Modern Data Company, does all of these things across the entire data stack. Data architectures can now be deployed in weeks rather than months or years. DataOS is built to set organizations on the fastest path from data to decisions in a truly risk-free manner.

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 structured, semi-structured and unstructured data assets across the enterprise. It builds an intelligent semantic layer that enables business and technical users to discover, explore and collaborate to deliver data products quickly and easily. The unmatched composability of DataOS lets customers adapt it to any data architecture, be it a data fabric, data mesh, lakehouse, or something new.

Speed and agility are instrumental to DataOS because the system 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.

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



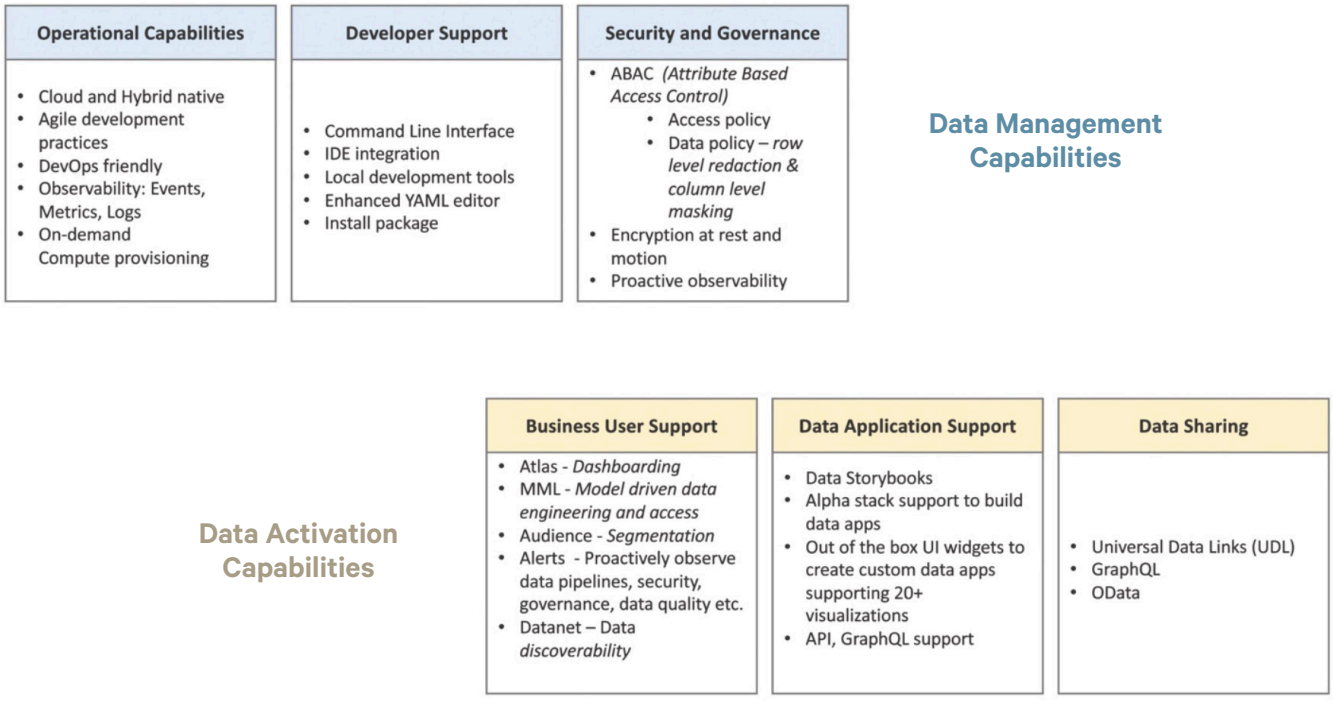
One operating system to quickly deploy multiple architectures

With these advantages, 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.

DataOS Capabilities Overview

It's important to understand the following DataOS-specific abstractions and applications to fully comprehend it.

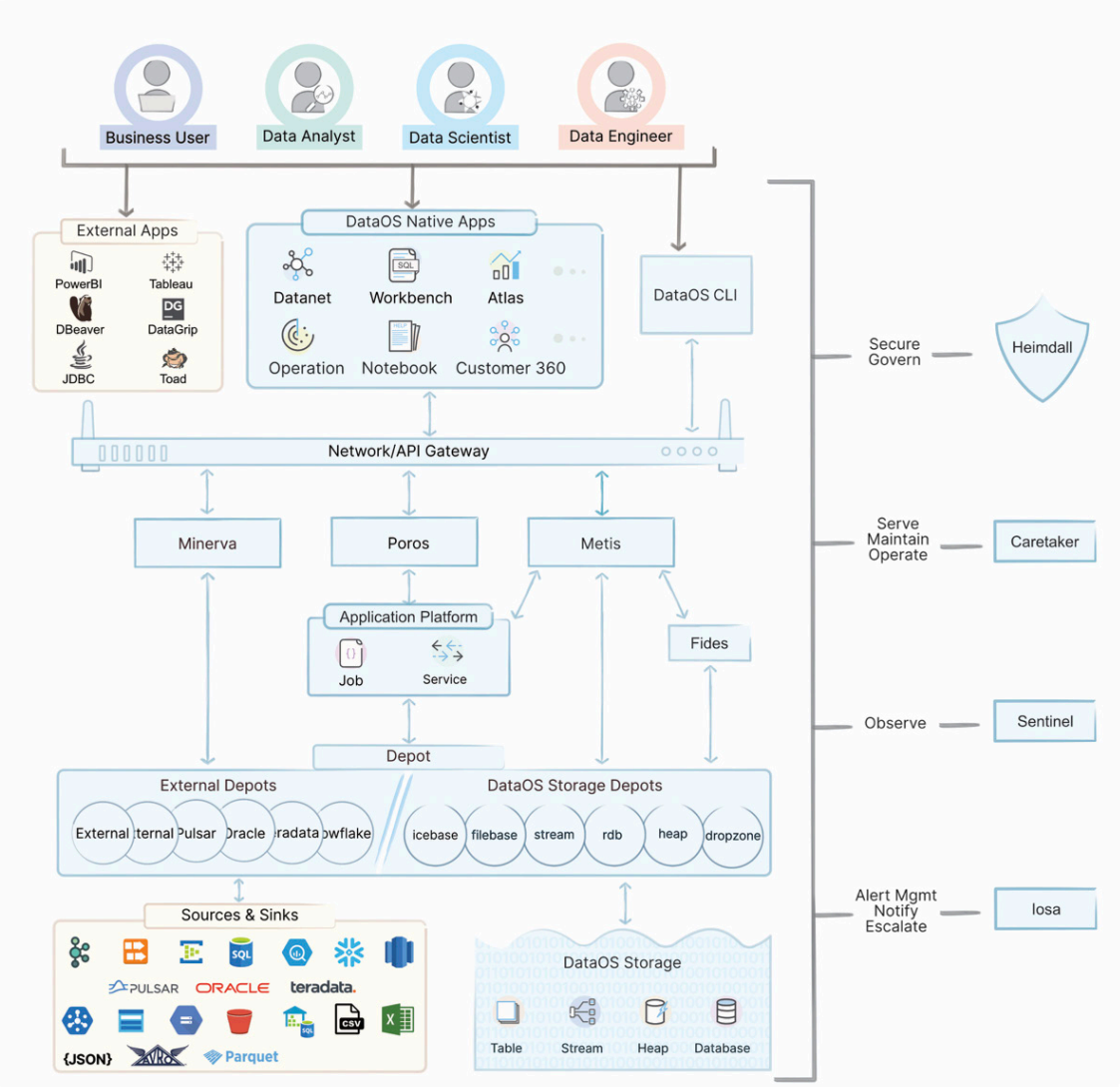


The addendum at the end of this paper is an exhaustive list of DataOS features.

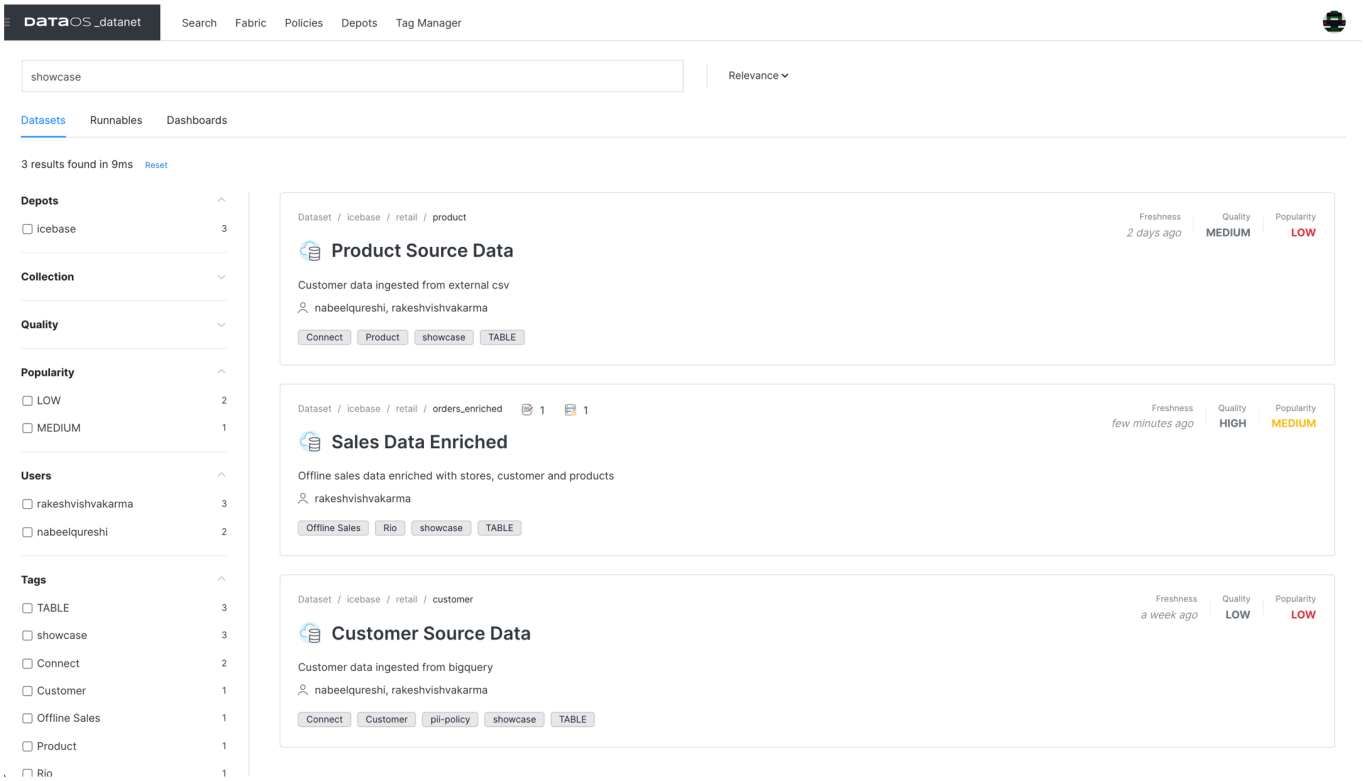
One of the most important abstractions to understand is a data depot. A depot provides a reference to the data source/sink; it abstracts the details of their configurations and storage formats. It helps a user easily understand the data source/sink, connect with it, and use the depot to access, process, and explore data within 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 job or services as well as other tools within the DataOS system.

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

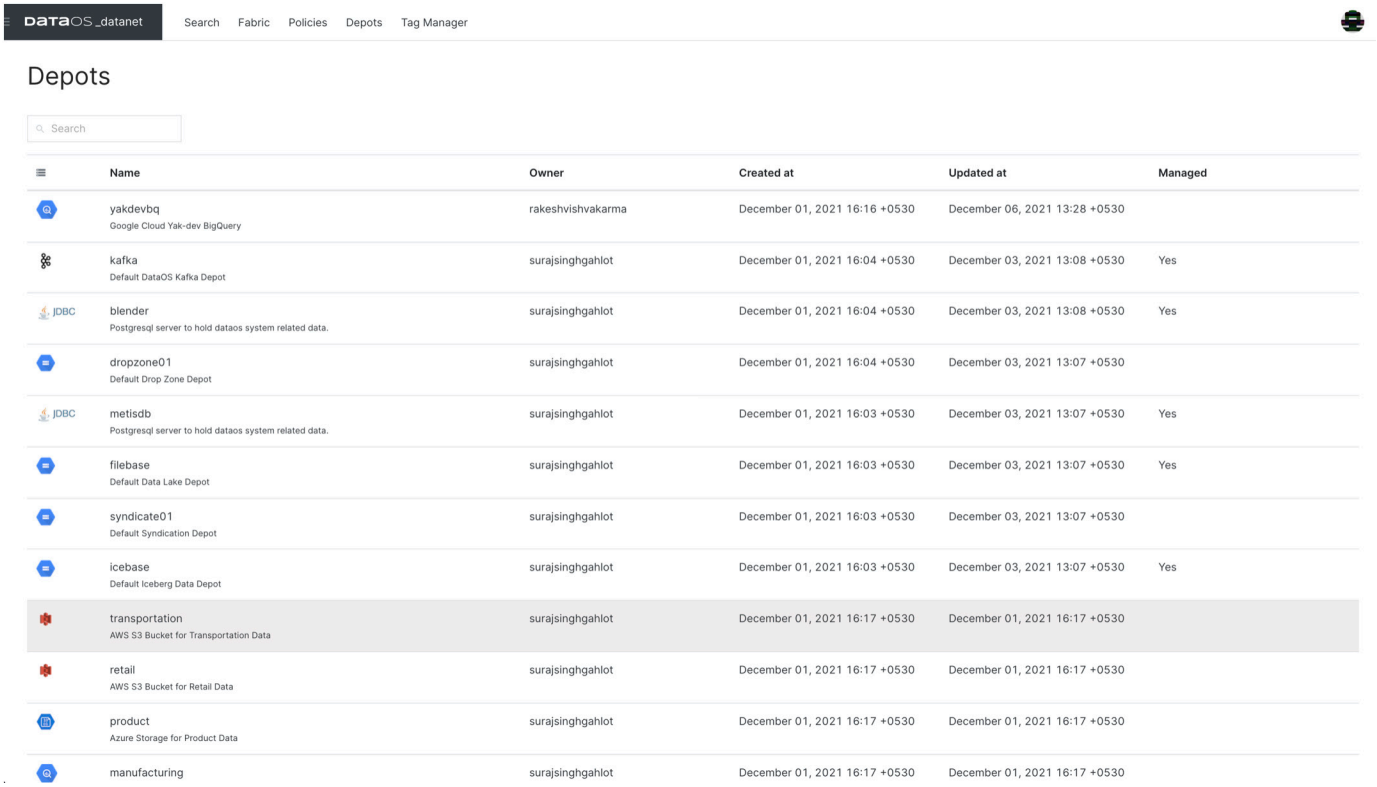
- Search datasets, executables and dashboards – In a single context view, see which users are working with data, what they are doing with it, and how often it is accessed
- Fabric with knowledge graph of the organization
- Create and manage access, masking, and filter policies
- Produce a list of depots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



Search datasets, executables, and dashboards



Depot abstracts away complexities of configurations and storage formats



Advanced Data Governance

While DataOS makes discovery seamless throughout the entire organization, 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.
- 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.

Data Quality

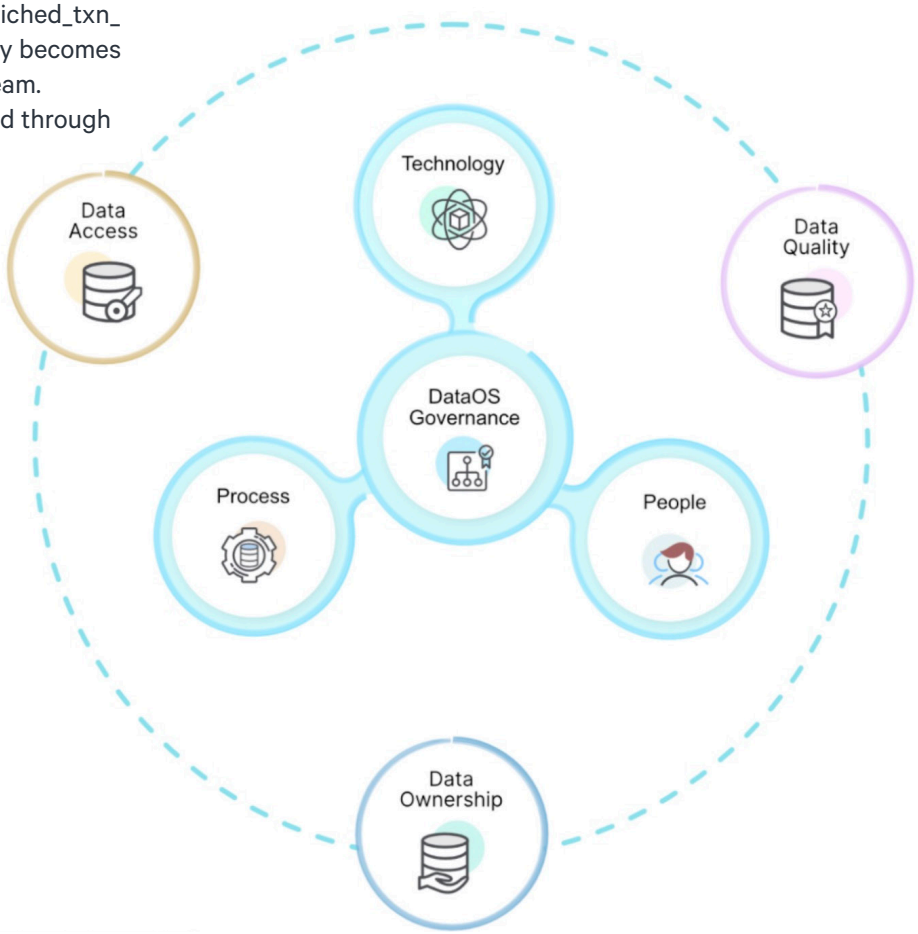
- Users can use Metis to learn about their own data and its landscape, such as quality, profile, lineage, dictionary, and much more.

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. Users can either read the data set or not.
- Users 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.

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 knowledge for most operations.

Through the Workbench, DataOS allows SQL queries on databases, then provisions data pipelines and workflows to enable anything from exporting the results 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 Tableau or any other analytics tool. Like any 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. This allows 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. 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.
- Setting up dynamic data integration in order to deliver integrated data through multiple delivery styles.
- Setting up automated data orchestration services.

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

The future is data-driven. With numerous design patterns to choose from, ever-expanding amounts of data, and data governance playing an increasingly central role, it's important to future-proof your investments. The key is DataOS: a true data operating system that can be molded into 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-generation data stack that also augments your existing infrastructure so building and delivering data products is faster and simpler for all users. DataOS is the data operating system delivering democratized access to quality data right when you need it, where you need it.

If you want to future-proof and transform how your organization thinks about data:

Think Agile
Think Composable
Think Data

Think Modern.

Addendum:
DataOS Feature List

1. **The Activation Layer**
    a. Data Sharing (Universal Data Link) - Share data with users via a web address fashion with the right governance policies
    b. SQL Workbench - Powerful workbench that enables a no-code option to querying data sets of all formats.
    c. Atlas - Quick BI. Create customized reports on any data sets 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.

2. **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

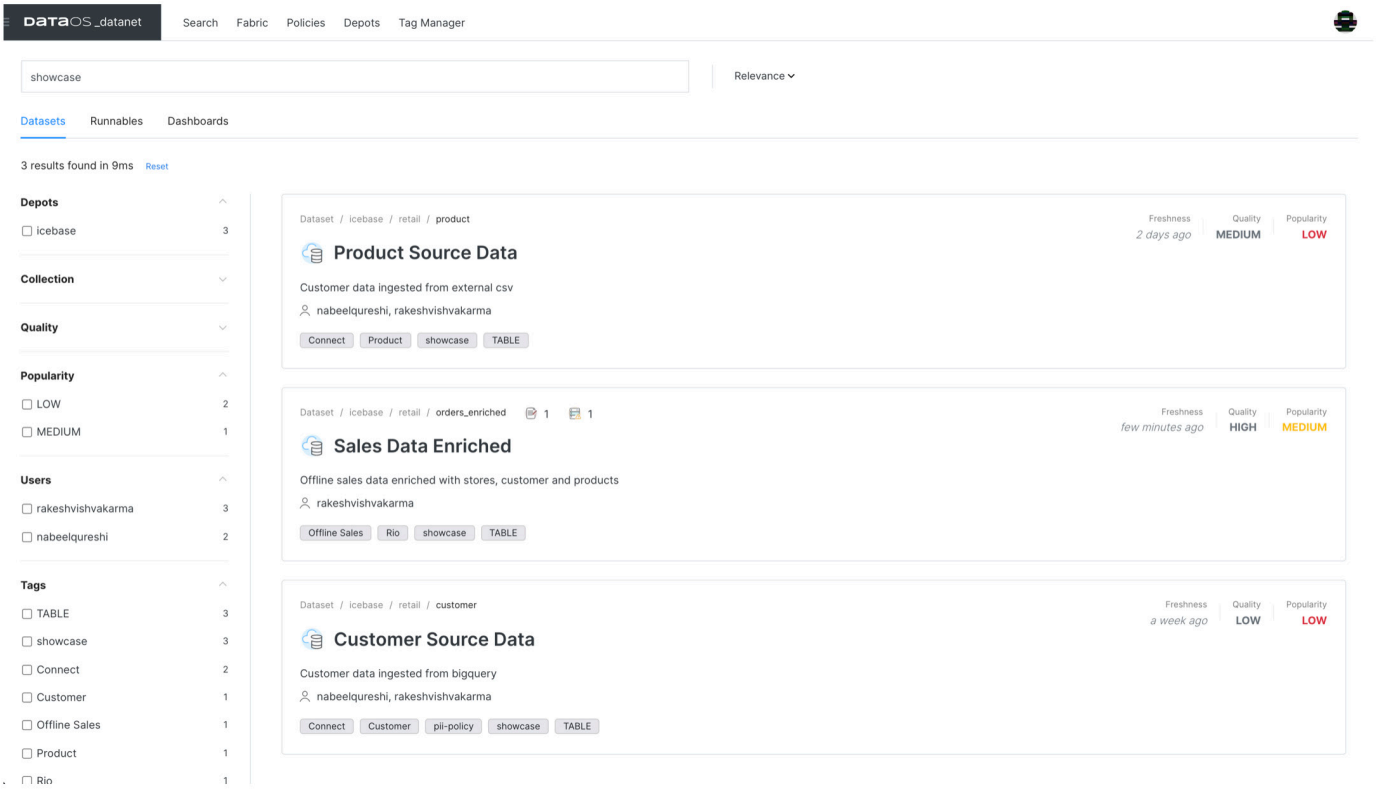
DataOS Feature List Continued

3. **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

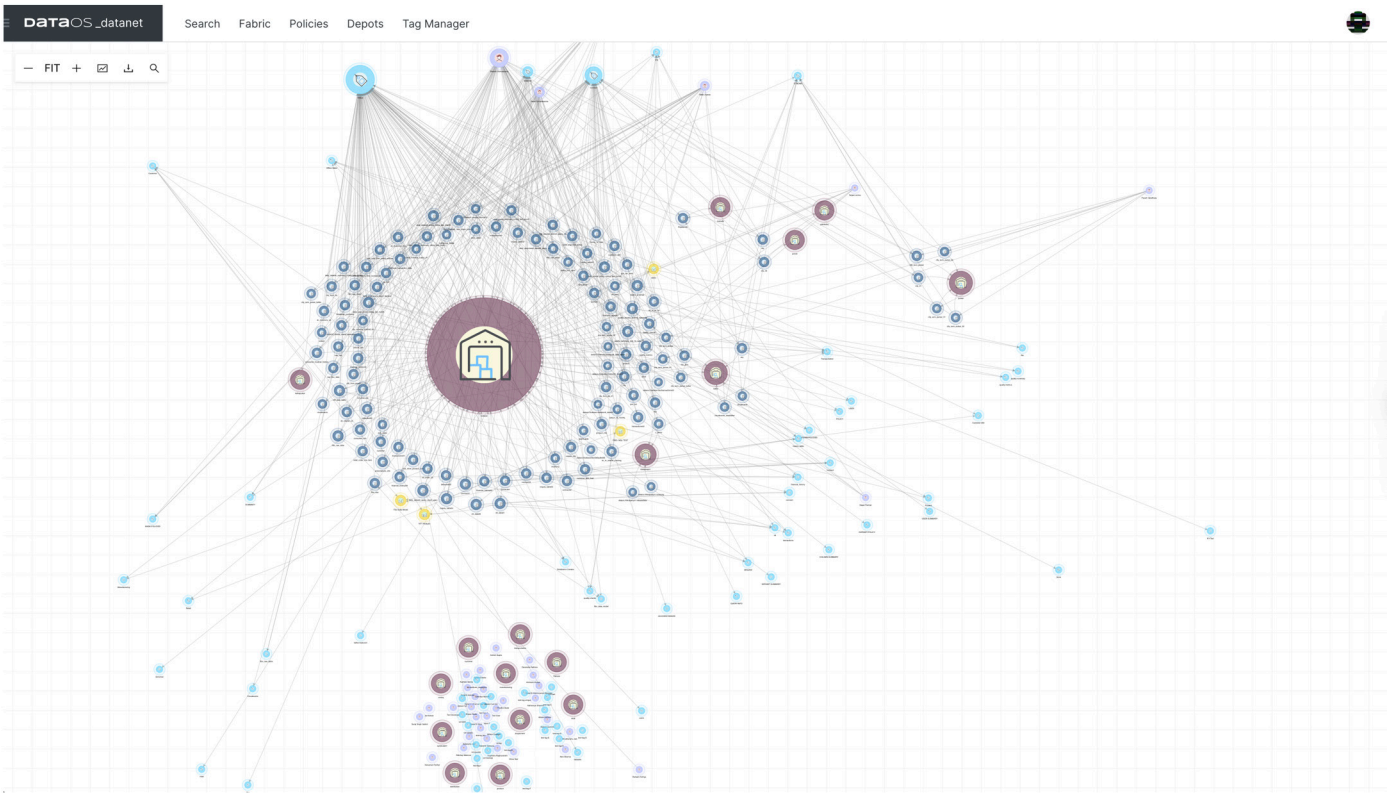
4. **Data & Dev Ops**
    a. Cloud and hybrid native
    b. Agile development practices
    c. DevOps friendly
        i. Version control
        ii. CI/CD
    d. Infrastructure observability to track and audit events, metrics, and logs
    e. On-demand compute provisioning
    f. Command line interface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

DataOS Screenshots

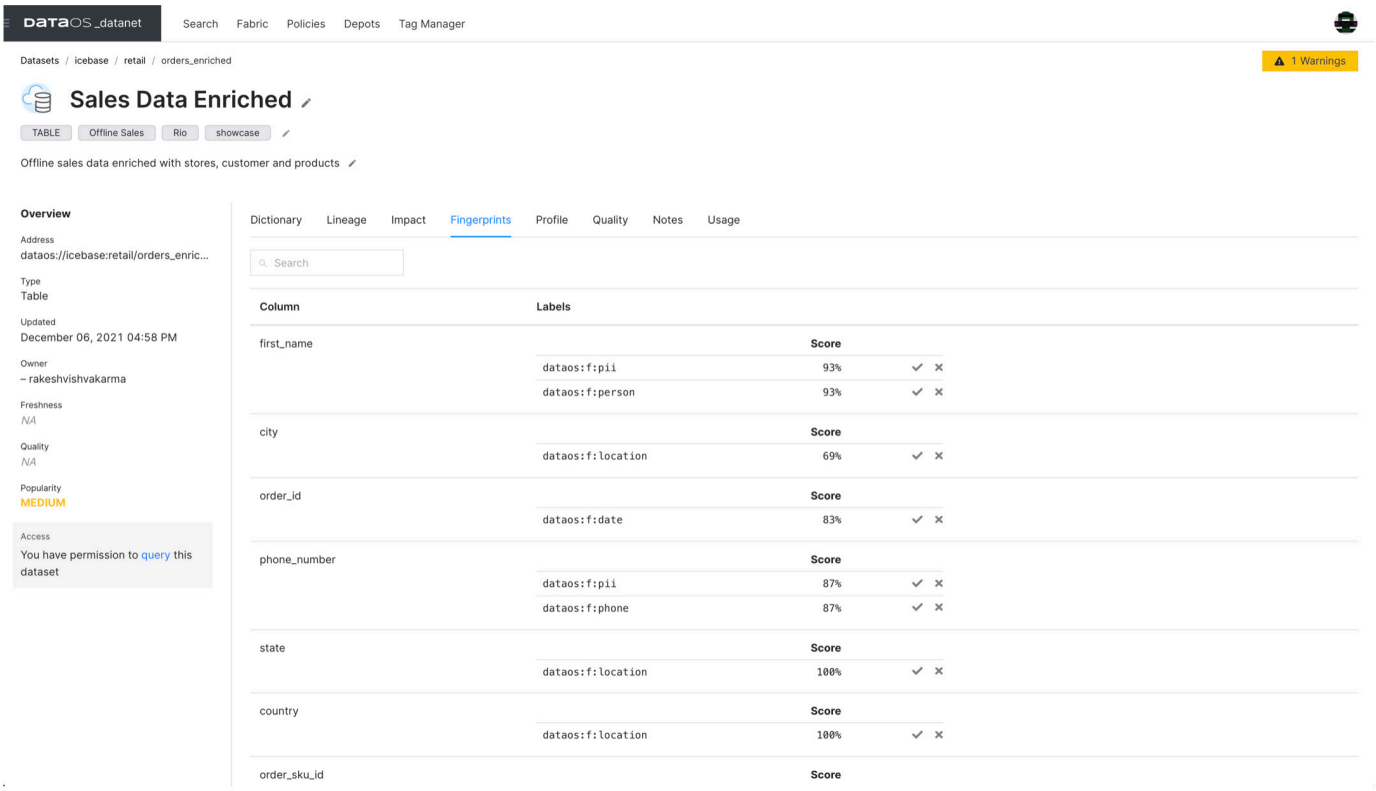
Datanet - Search



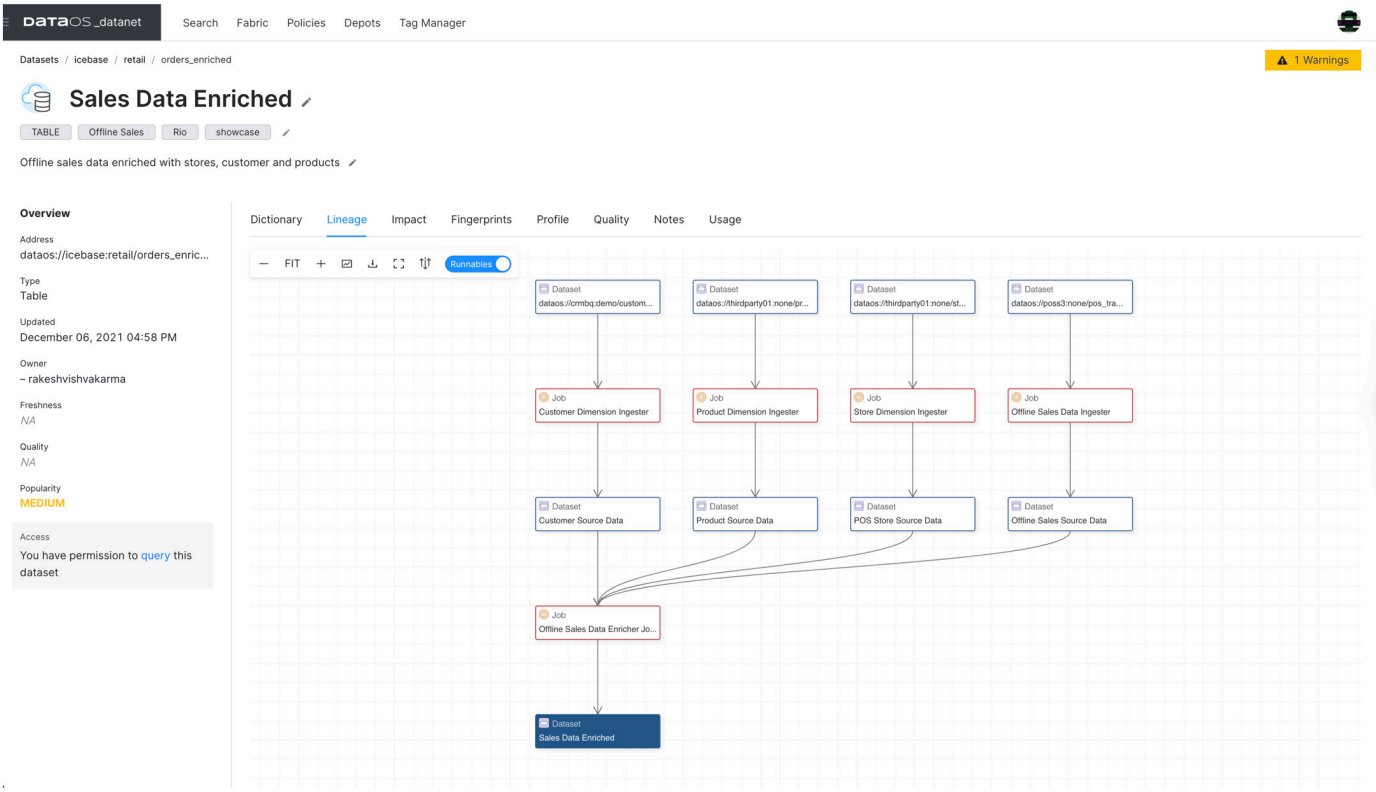
Datanet - Fabric



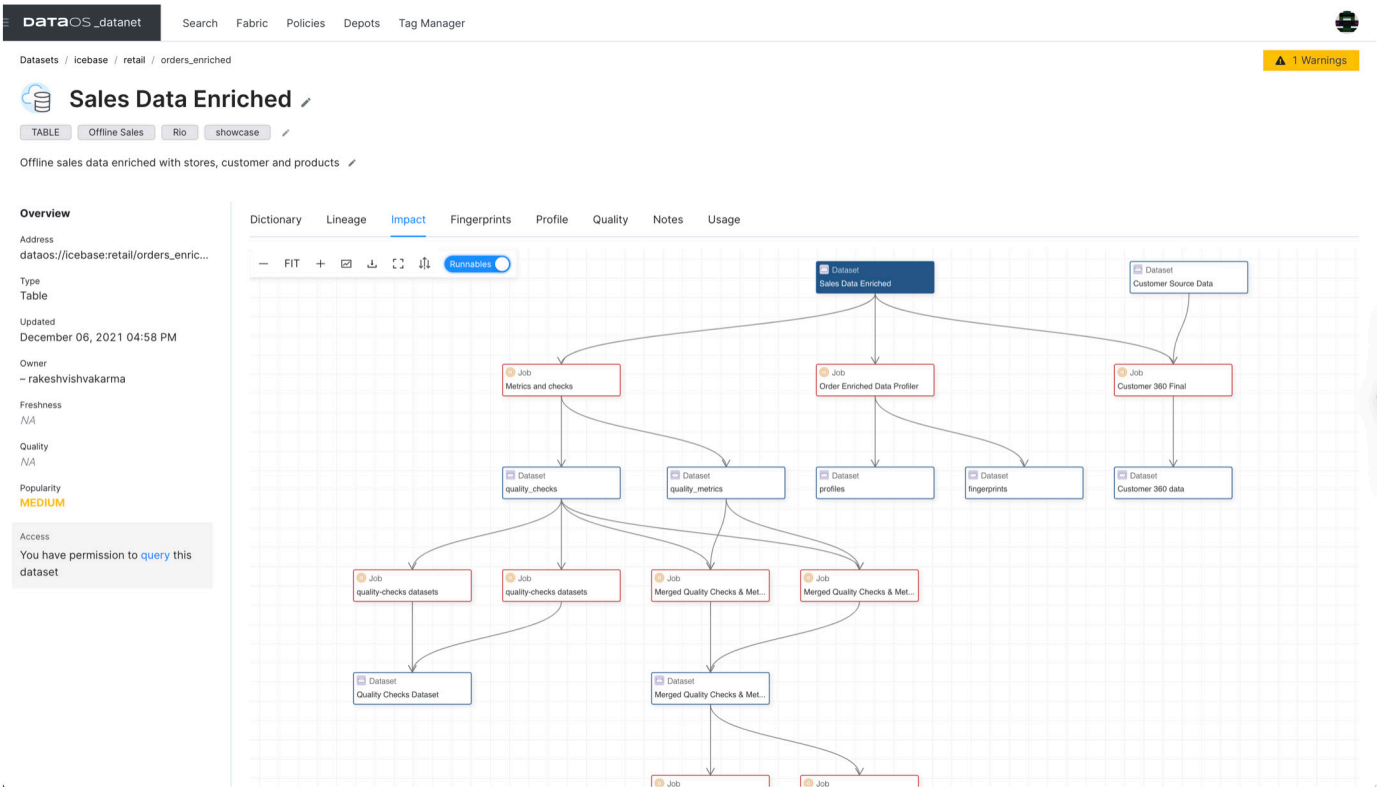
Fingerprint (Data Classification)



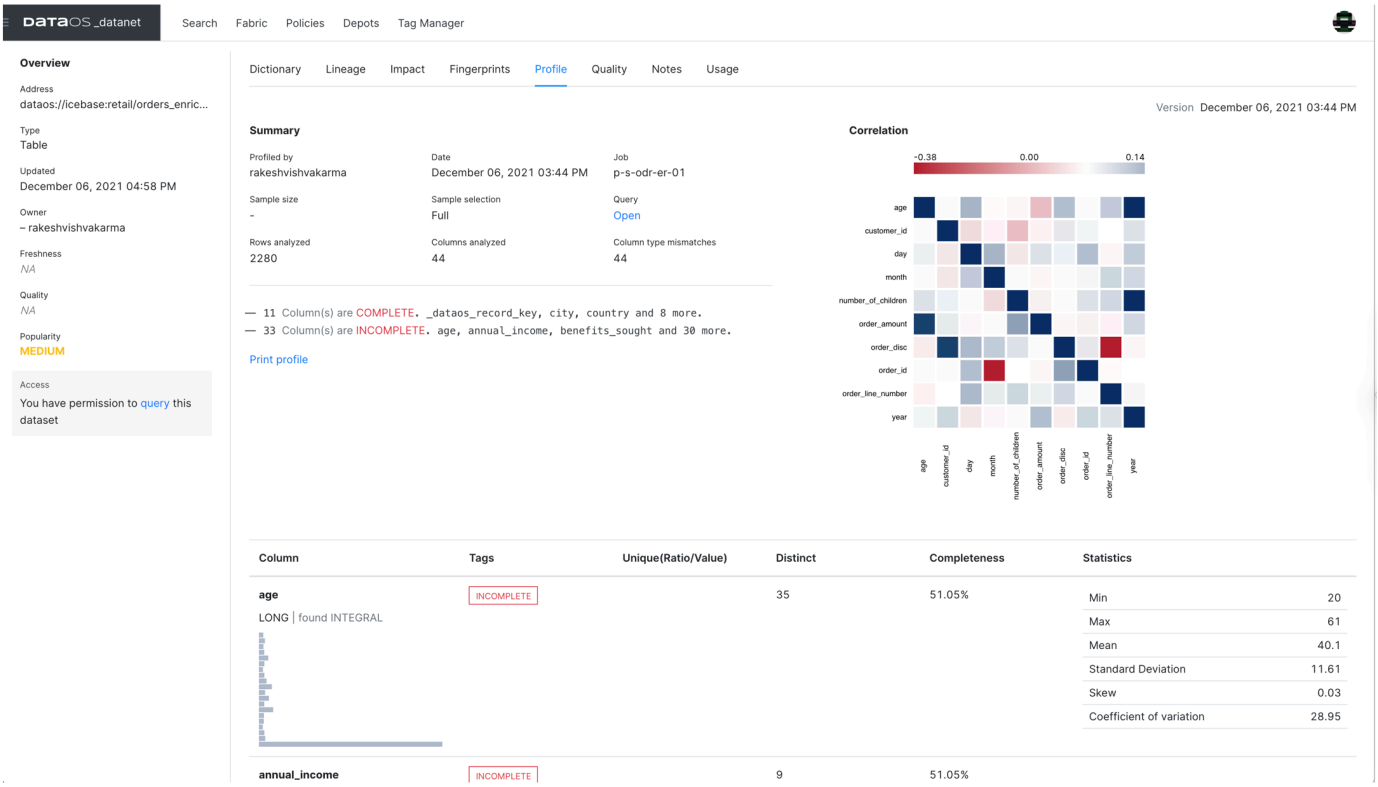
Lineage



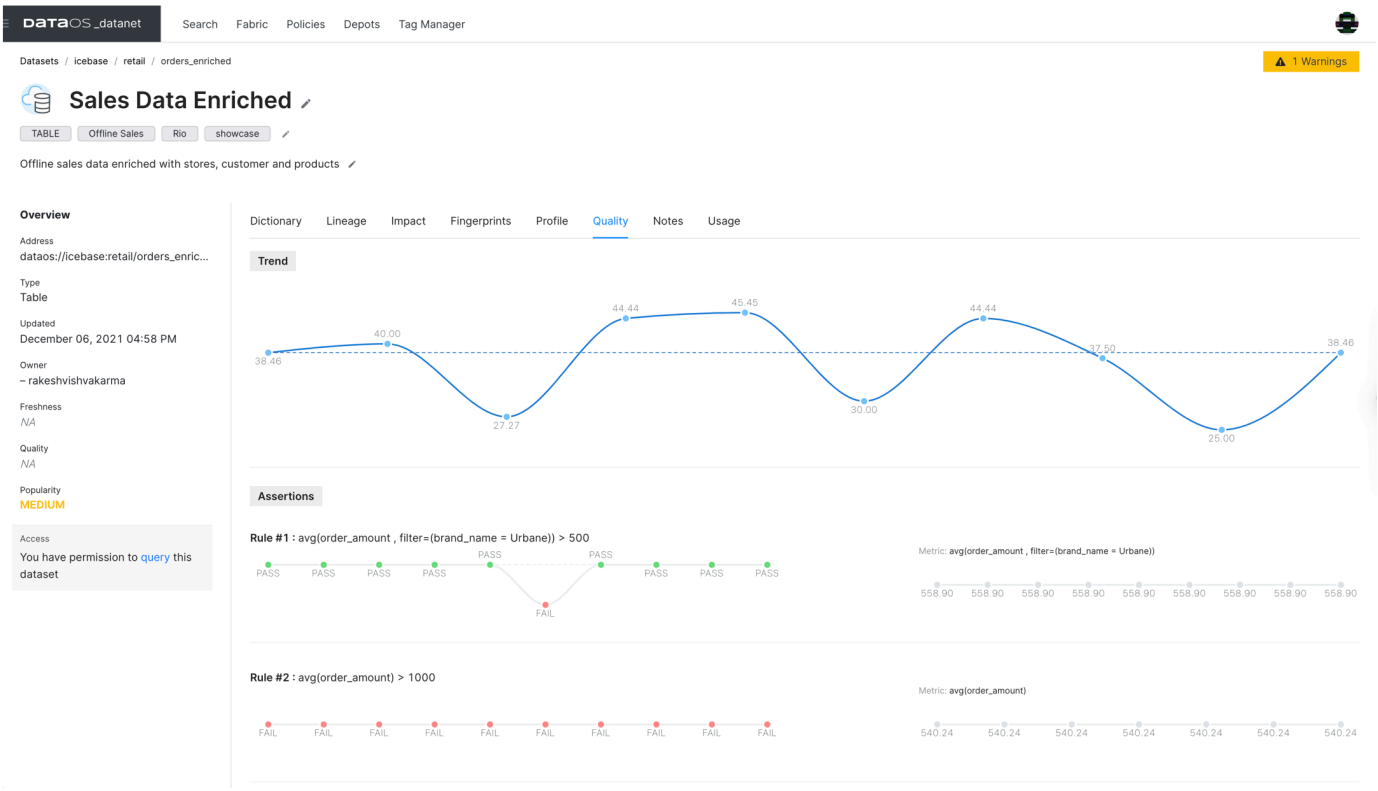
Impact



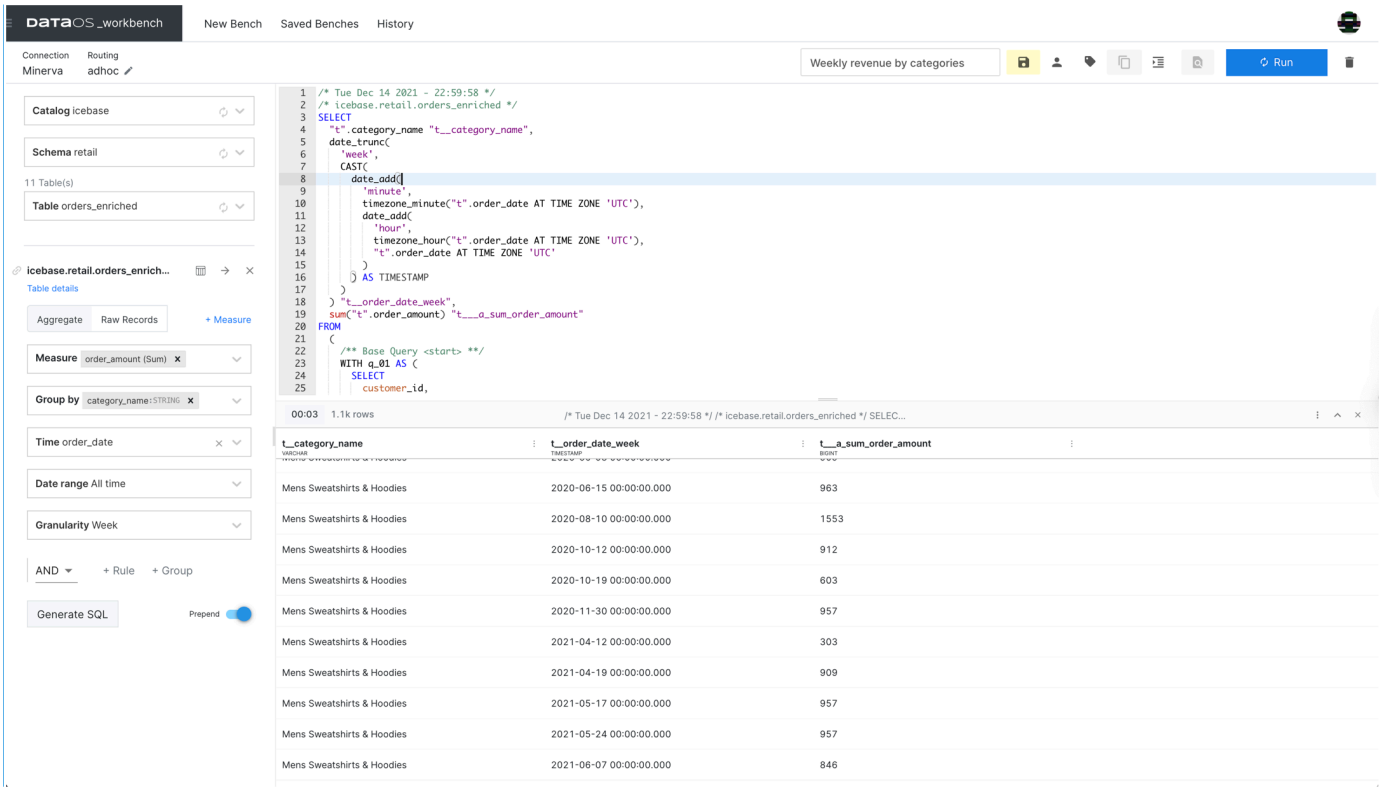
Data Profile



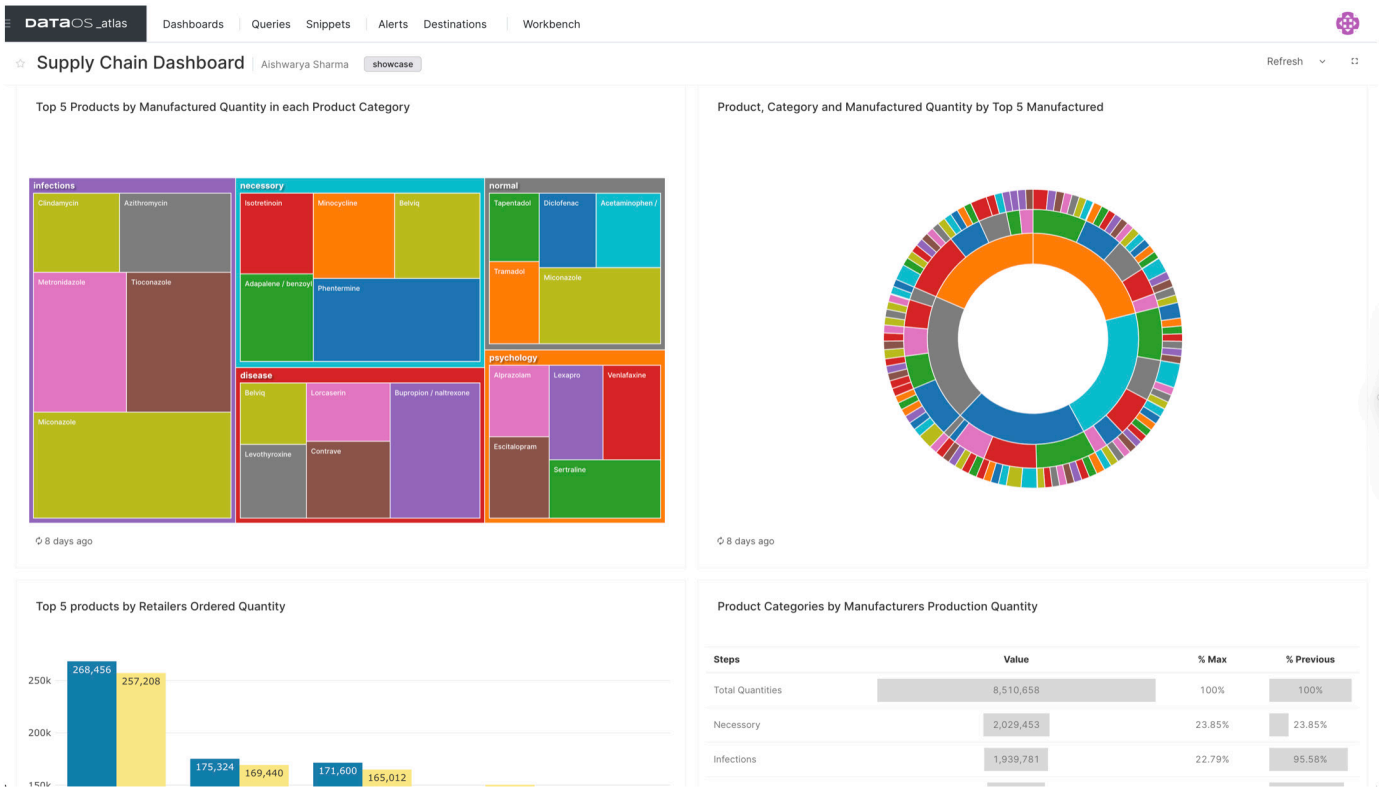
Data Quality



Workbench – Data Studio



Atlas - Dashboard



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

G26M