

DataOS[®]

A Paradigm Shift in Data Management

Creating Scalable Analytics

 Modern White Paper



Table of Contents

Data Needs Are Changing in A Transformative Way
Organizations Are Investing Heavily to Become Data-Driven
How Rigid Data Architectures Inhibit Analytics
Are Data Fabrics the Panacea?
Data Operating Systems – A Paradigm Shift
A Data Operating System Exists
DataOS Capabilities Overview
Data Governance
Sample Use Case
Addendum Feature List



DataOS®: A Paradigm Shift in Data Management – Creating Scalable Analytics
© 2022 The Modern Data Company. All trademarks are properties of their respective owners.

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. Business teams spent far more time gathering and preparing data for analysis than on the real work of analytics. Analytics and data science teams spent too much time locating and compiling data for their models while often fighting for more resources and permissions.

However, over the last two decades, there have been substantial scientific and engineering advances that allow for 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. In turn, this makes it easier than ever to create high impact, scalable analytical processes that can transform a business.

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 analytics and data science professionals 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 data platform users grow exponentially. At the same time, more and more demand has arisen for complex and scalable analytics processes that go well beyond basic reporting. Today, being data-driven is a competitive advantage. All 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 are 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 and analytics ecosystem. These new tools, however, are often point solutions that encourage experts building analytical processes to disregard many of the tools they already use. Instead of direct transfer of most skills to the new environments, users may have to spend substantial time learning new interfaces and protocols.

Creating a data-driven organization is still a distant dream. Increasingly, even as analytical capabilities expand, data architectures have become far too rigid. Most non-technology organizations are not getting value out of their data investments. Analytics practitioners and industry observers are dissatisfied. An emerging consensus is that organizations looking to be data-driven are falling behind, overwhelmed by the vast amount of data being generated and the growing demand to analyze it.



How Rigid Data Architectures Inhibit Analytics

The rigidity of most data architectures causes significant and tangible barriers to analytics rather than delivering on their promise of making it easier to analyze the data. While analytics experts know how to wrangle data to get what they need, rigid architectures make it an inefficient, time-consuming, and frustrating process.

Three problems stand out: long analysis times, data silos/dark data, and toolset lock-ins.

Long Analysis Times

The time to deliver anything from a simple dashboard to a predictive operational system depends on the way data pipelines are structured. Typically, any data requirements that are not already in place must be routed through the IT team. The IT team must then locate the data and create a process to load the data and make it available for analysis—while applying the appropriate access and security protocols. There is usually some back and forth with the user before everyone is satisfied with the resulting data and setup. This all happens before the analytics expert can start working with the new data.

Rigid architectures can also at times cause problems with execution. If too much data must be moved back and forth for a process, this can seriously reduce performance. A flexible architecture will allow for changes that reduce the need for data movement. A rigid architecture often forces complex analytical processes to be executed in ways that take far too many system resources and far too much time to be practical.

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 work together 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).

As analytical experts build their processes, they often add to the problem by acquiring additional data and using it without officially logging it and making it more broadly available. At the same time, data that would be extremely useful for the experts is unknown to them. Having automated and easy-to-use tools to find, catalog, and make dark data available across the enterprise will enable data scientists to build more powerful models.

Toolset Lock-ins

Professionals who build complex data science and analytics processes for a living have a wide range of tools to choose from today. While only a decade ago a handful of tools dominated the market, there are dozens of tools and programming languages being used today. At the same time, each professional strongly prefers his or her toolset to the others. Many people will even filter job opportunities based on the toolsets in use at a company.

As a result, it is imperative that organizations have an architecture in place that allows analysts and data scientists to use their tools of choice to access and analyze data. Systems that only provide proprietary tools or that lack meaningful integration with more than a few toolsets will not enable the level of productivity required. The modern and scalable approach is one that allows people to bring their own toolset and use it against all the relevant data available.



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 capital and human capital, data is used as a form of capital (data capital) to empower teams. Analytics 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 obsolete from the start. 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 to support their analytics.

Modern systems render optimized data pipelines in a governed fashion to those building analytical processes. Such systems are engineered with various modular components and have inspired the term “data fabric.” The term data fabric, which refers to 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. 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. Similarly, “data mesh” is another popular design pattern used today.

Both the data fabric and the data mesh paradigms are deeply advocated for in their respective forums. Both involve not only a technical but a “cultural” re-alignment, requiring companies to embrace a novel approach to using technology. In the end, an analytics and data science team can make use of either of these approaches effectively. However, as will be discussed next, there might be a better, third option to these two approaches. It appears we may be far from an all-encompassing panacea for all things data.



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 and create the analytical processes that they need. This approach can support deployment of a data fabric or a data mesh—or an even more novel design in future. This future-proofs an organization's data infrastructure needs with a composable platform that can accommodate all architectures, users, and systems simultaneously. A good way to understand this would be to imagine 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 components as well as LEGO bricks for the analytical components of a company's requirements.

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.

1. A consistent interface to all data and analytical resources within the system
2. Features that allow automated data discovery and profiling
3. Orchestration tools supporting resource allocation for complex analytical requirements
4. Access to a range of machine learning and artificial intelligence functionality
5. Support for the delivery of insights with features such as data story books
6. Ability to use one's preferred toolsets while creating and managing reusable artifacts
7. Enable creation of data products quickly in both cloud or hybrid environments

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 build analytics processes that utilize them. 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. Support for automating the execution of an analytics process and distribution of its results is central. Delivery can be a highly scalable back-end delivery from one process to another. Or it can be a data storybook built for a business sponsor. All this should be possible for a data scientist to do with minimal additional training.

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 Data Operating System Exists

DataOS® from The Modern Data Company, does all these things. Data architectures enabling robust and scalable analytics and data science capabilities can now be deployed in weeks rather than months or years. DataOS is built to set organizations on the fastest path to analytical success 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 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. In addition, DataOS supports a wide range of toolsets, which allows data scientists to use their preferred interfaces and languages while making full use of the power of DataOS.

There are four overarching benefits to analytics and data science professionals and organizations that set DataOS apart from anything else in the market today:

- **Composable, extensible, open and interoperable**

- DataOS is completely composable and lets customers adapt it to any use case of their choice, using any data, and applying any algorithm.
- The agility of DataOS lets organizations react and adapt to new analytical requirements.

- **Enhancing--not replacing--existing tools**

- Augment the functionality and ROI of your current investments in analytical skills and toolsets.
- There is no pressure to replace components in use. Simply use them to access DataOS.

- **All things data under one roof**

- Building data products has never been faster or easier--weeks/hours instead of quarters/years.
- With on-demand provisioning to support any level of scale, the ability to create and manage reusable artifacts, and built-in capabilities such as data storybooks to aid in delivery of insights, DataOS has you covered.

- **Embedded Machine Learning Capabilities**

- Tap into the machine learning capabilities native to DataOS
- The integration of preferred machine learning tools with DataOS is fast and seamless
- Apply machine learning to any type of data, in any type of repository, whether in the cloud or in a hybrid environment

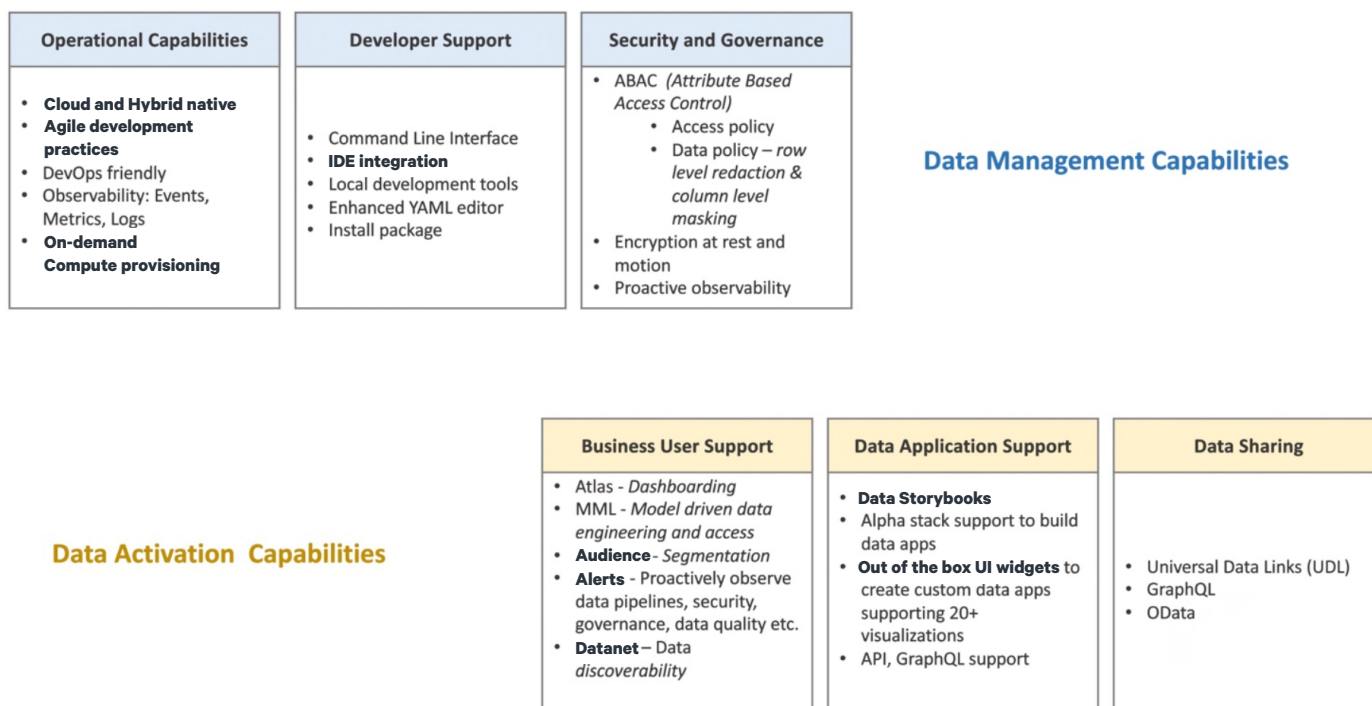


With these advantages, DataOS delivers tangible value across many use cases:

- Build analytical products faster with superior data experience—in hours or weeks instead of quarters or years.
- Leverage embedded machine learning tools in DataOS, or make use of your preferred machine learning toolset
- Focus on value creation from data instead of integration and process work.
- Accelerate analytics with automated data cataloging and profiling capabilities.
- Build out solutions easily without extensive training in the DataOS environment.
- Drive new revenue models by sharing or collaborating with secure, governed data.
- Modernize your infrastructure to better enable automated, scalable analytical processes.

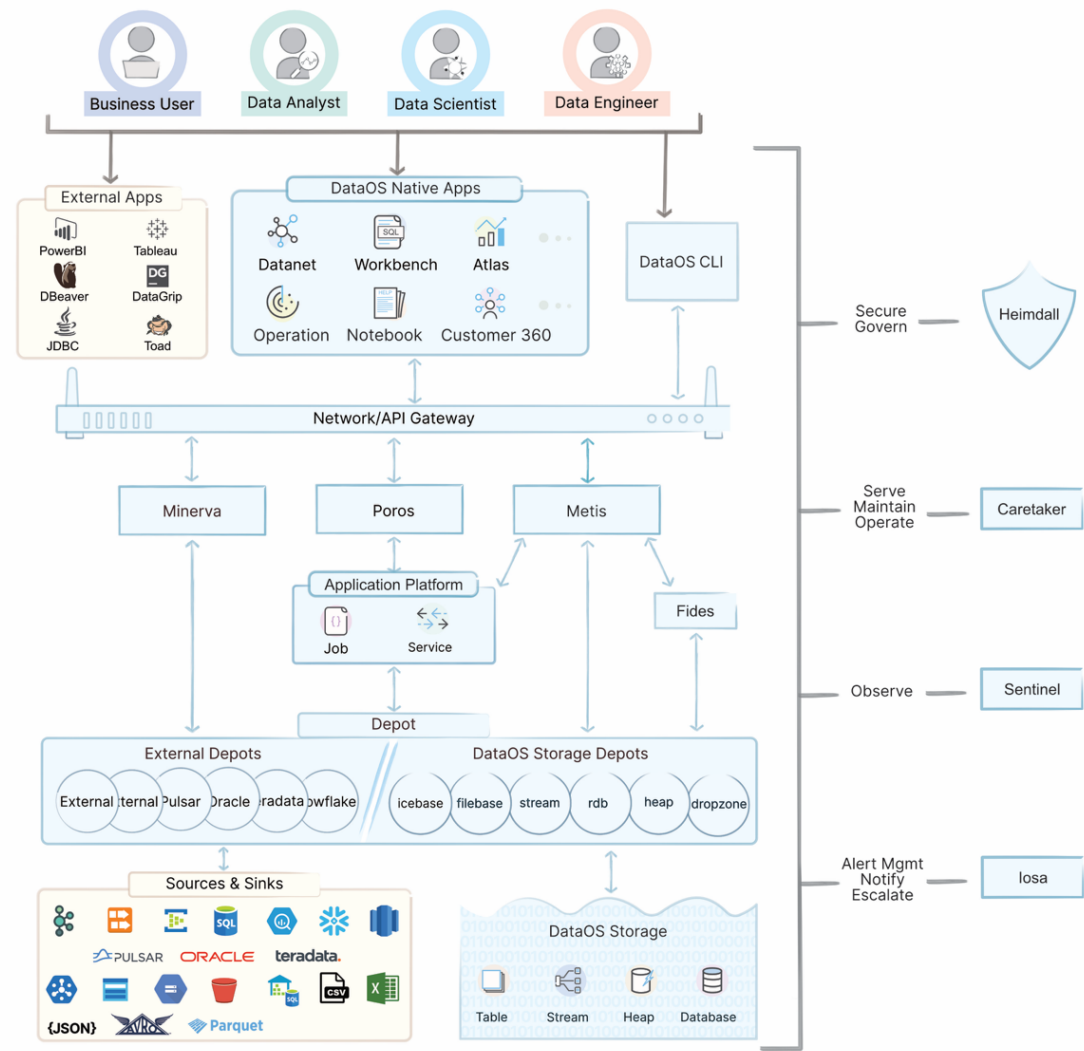
DataOS Capabilities Overview

There is a set of DataOS-specific abstractions and applications that are important to understand 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. This makes creating analytics and data science processes much simpler than in other common environments.



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 system encompasses the people, process, and technology required to ensure that data is fit for its intended purpose. The data governance features of most interest to analytics and data science professionals include:

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.
- This enables processes to be built while maintaining ownership and preventing access to resulting output until the process is complete and fully tested

Data Quality

- Users can use Metis to learn about their own data and its landscape, such as quality, profile, lineage, dictionary, and much more.
- This capability reduces the time spent wrangling, cleaning, and preparing data for analytics

Data Access

- Metadata of all datasets will be available for anyone in the organization to consume.

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. 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. What this means is that analytics and data science professionals can go about their work with confidence that the right security is in place even as they apply new logic to data.

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 analytical or machine learning tool in the 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 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

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 support any type of analytical processing from simple queries to complex machine learning algorithms.

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 that 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 building and delivering data products is faster and simpler for all users. It is the data operating system delivering scalable access to quality data from the analytical tools that you prefer.

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

Datnet - Search

DATAOS_datanet

SearchFabricPoliciesDepotsTag Manager

showcase

Relevance

Datasets

Runnables

Dashboards

3 results found in 9ms

Reset

Depots

icebase3

Collection

Quality

Popularity

LOW2

MEDIUM1

Users

rakeshishvakarma3

nabeelqureshi2

Tags

TABLE3

showcase3

Connect2

Customer1

Offline Sales1

Product1

Rio1

Dataset / icebase / retail / product

Product Source Data

Customer data ingested from external csv

nabeelqureshi, rakeshishvakarma

ConnectProductshowcaseTABLE

Freshness

2 days ago

Quality

MEDIUM

Popularity

LOW

Dataset / icebase / retail / orders_enriched

Sales Data Enriched

Offline sales data enriched with stores, customer and products

rakeshishvakarma

Offline SalesRioshowcaseTABLE

Freshness

few minutes ago

Quality

HIGH

Popularity

MEDIUM

Dataset / icebase / retail / customer

Customer Source Data

Customer data ingested from bigquery

nabeelqureshi, rakeshishvakarma

ConnectCustomerpii-policyshowcaseTABLE

Freshness

a week ago

Quality

LOW

Popularity

LOW

Lineage

DATAOS_datanet

SearchFabricPoliciesDepotsTag Manager

Datasets / icebase / retail / orders_enriched

1 Warnings

Sales Data Enriched

TABLEOffline SalesRioshowcase

Offline sales data enriched with stores, customer and products

Overview

Address

dataos://icebase:retail/orders_enrich...

Type

Table

Updated

December 06, 2021 04:58 PM

Owner

- rakeshishvakarma

Freshness

NA

Quality

NA

Popularity

MEDIUM

Access

You have permission to query this dataset

DictionaryLineageImpactFingerprintsProfileQualityNotesUsage

Runnables

Dataset dataos://combq/demo/custom...

Job Customer Dimension Ingestor

Dataset dataos://icebase:retail/orders_enriched

Dataset dataos://thirdparty01:none/pt...

Job Product Dimension Ingestor

Dataset dataos://icebase:retail/orders_enriched

Dataset dataos://thirdparty01:none/st...

Job Store Dimension Ingestor

Dataset dataos://icebase:retail/orders_enriched

Dataset dataos://pos3:none/pos_tra...

Job Offline Sales Data Ingestor

Dataset dataos://icebase:retail/orders_enriched

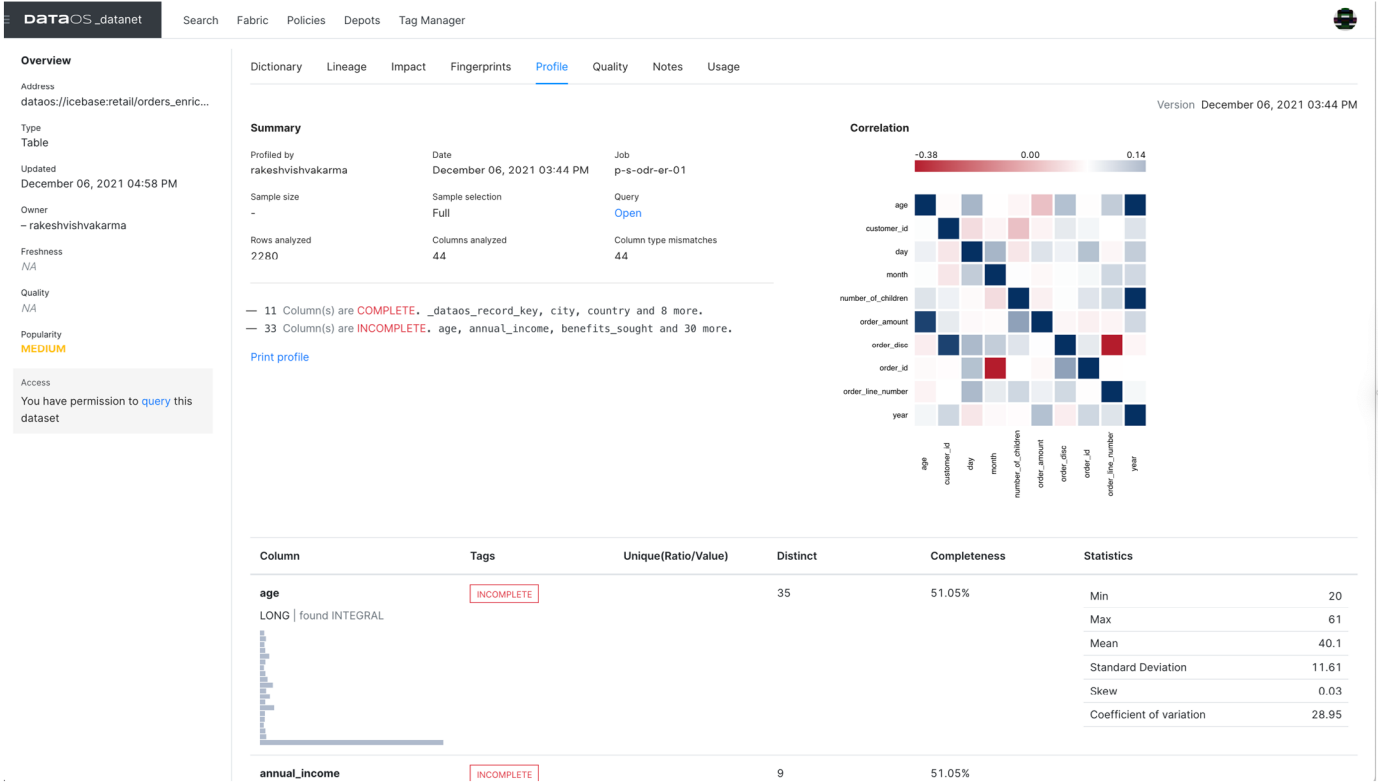
Job Offline Sales Data Enricher Jo...

Dataset Sales Data Enriched

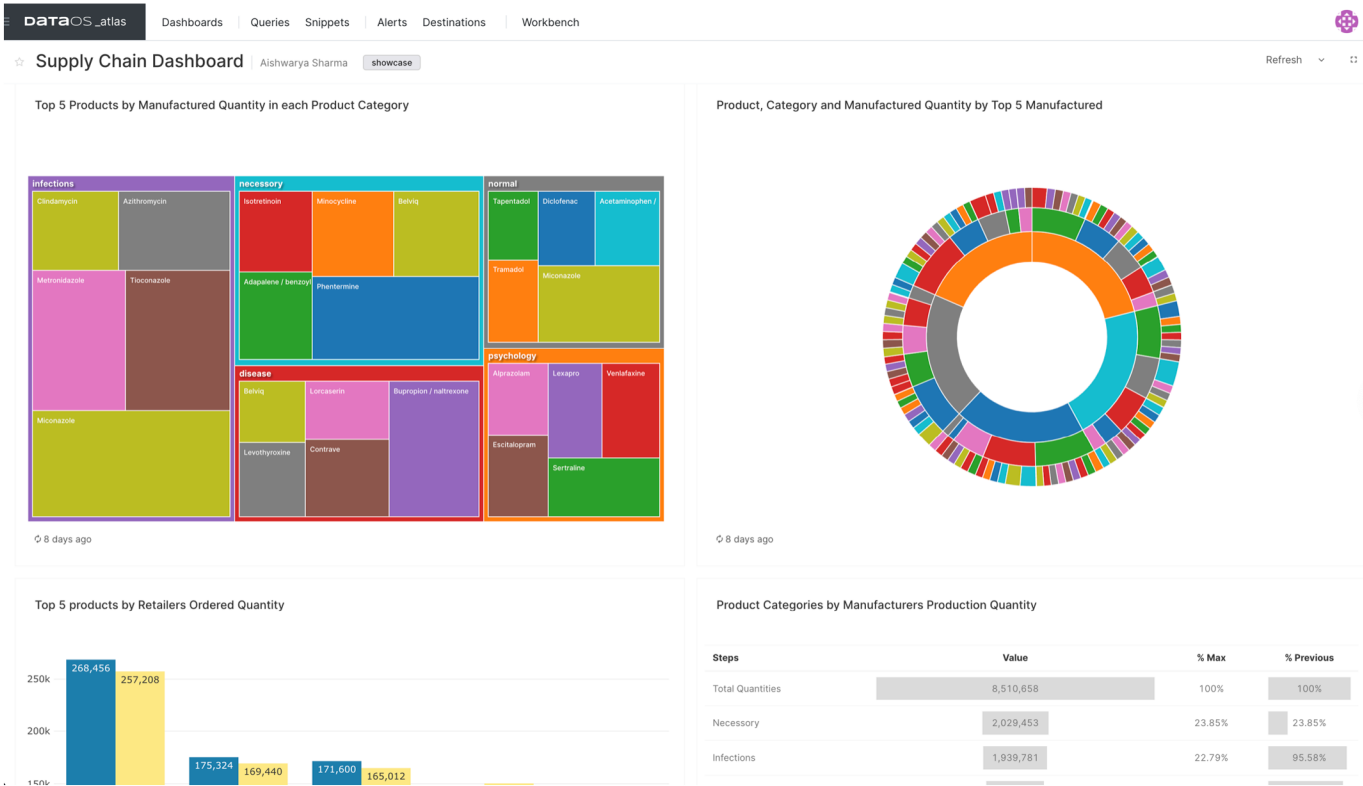
 DataOS®: A Paradigm Shift in Data Management – Creating Scalable Analytics
© 2022 The Modern Data Company. All trademarks are properties of their respective owners.

16

Data Profile



Atlas - Dashboard



dataOS®: A Paradigm Shift in Data Management – Creating Scalable Analytics

© 2022 The Modern Data Company. All trademarks are properties of their respective owners.

18

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.



DataOS®: A Paradigm Shift in Data Management – Creating Scalable Analytics
© 2022 The Modern Data Company. All trademarks are properties of their respective owners.

The Modern Data Company
306 Cambridge Ave
Palo Alto, CA 94306

TheModernDataCompany.com
info@TMDC.io