Modern WhitePaper

DataOS®

The Fastest Path from Data to Decisions

Table of Contents

TheModern
DataCompany

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

DataOS®: Fastest Path from Data to Decisions

Current Challenges in Leveraging Big Data for Operational Excellence

There has been a proliferation of new enterprise data management solutions over the last decade. While there are a lot of solutions in market, they have not met the objectives for those who manage data within organizations. Implementing these solutions has been a challenge resulting in poor access and quality of data. Broadly, the challenges they face in operationalizing their data reserves fall under these three pillars:

- Poor visibility of data and insufficient metadata
- Time and effort needed to ensure data's access and quality
- Vendor imposed constraints

Poor visibility of data and insufficient metadata

It is clearly established that there is no dearth of actionable data in the age of Industrial IOT and 5G. However, consumers within organizations have low visibility of all valuable data. This is due to data silos created by different departments utilizing different systems that don't communicate well with each other. Moreover, large chunks of data exist in organizations 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.

Consumers of data also struggle with insufficient metadata leading to challenges like absent lineage and/or provenance. This prevents a user from leveraging the complete value of a dataset.

Time and effort needed to ensure data's access and quality

Typically, due to the "non-self-serve" nature of these pipelines, consumption and management of data are done by different individuals and machines in organizations. The time to deliver anything from a simple dashboard to a predictive operational system depends on the way data pipelines are structured. Ad hoc data requests go to the IT department where they are added to a queue behind many such requests. Building an ad hoc pipeline and ensuring compliance with the organization's governance policy takes significant time and effort and weighs down on data's value.

Additionally, consumers often retrieve data with suspicious quality. In 2018, Kaggle found that data scientists spend 40% of their time cleaning and organizing the data they receive. The reason is that data pipelines are seldom designed with principles first.

Vendor imposed constraints

Rather than rely on standard, open data formats, many vendors choose specialized formats that cannot be used or understood by other systems. Vendors naturally want to retain their customers and they do this, in part, by making it hard to move customer data off of their platforms. But what this implies is that data users are often unable to leverage the very data they own (data lock-in).

A Paradigm Shift is Needed

Organizations today need a transformational product that solves the complexity of data infrastructure while also delivering the agility and performance needed to help meet the challenges of today and tomorrow. Such a product should embrace the following key tenets:

Future proof investments

Organizations are investing 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 primarily built to cater to a specific use case or persona and hence are only “point solutions.” Organizations stitch these point solutions together in order to create their data infrastructure. These data infrastructures require numerous integrations and lots of services to patch everything together, and keep updated and maintained for daily operations.

Increasingly, as a result of incorporating point solutions, data architectures have become far too rigid. Organizations often struggle to integrate, govern, process, and syndicate data to external entities in order to generate the value they desperately need. Most non-technology organizations are not getting value out of their data investments.

A new school of thought is emerging. This new paradigm postulates a future where loosely coupled but tightly integrated building blocks enable organizations to compose the data architectures they need. This would serve to deploy a data fabric, 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. Such a composable platform would look like 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’s bricks in which the same components can be used to construct either a toy car or a bus.

A governance framework that truly democratizes data

Operationalizing and scaling data, while still governing it, has become a complex riddle. Getting value out of data is challenging with this sort of overhead. Organizations in the business of core technology have simply chosen to create their own proprietary tools to make this possible. Data’s true democratization is instrumental — only then will data be available for use by all authorized consumers. A scalable data governance framework is quintessential to deliver this objective so that hundreds of thousands of humans and machines seeking data get automatic access.

Active metadata management to bolster data discovery and governance

Each dataset also has rich quantities of associated metadata. In order to leverage this data, especially for machine learning use cases, it is important to incorporate knowledge graphs to model the organizational domain. Knowledge graphs are abstractions for organizing data from multiple sources, capture information about entities in a certain domain, and understand relations between them. This can be done by cataloging metadata and connecting structured and unstructured, internal and external data sources, and tracking their ownership over time. By managing data as a powerful network of information, enterprises can move away from point-to-point integrations via ETL (extract, transform, load) and APIs and pivot towards a more modern way of powering all their data needs from one place. This provides enterprises the agility to try any data application, ML model or 3rd party data product, without the need for extensive integrations. All of the investments that organizations are placing in data products can now yield the ROI they expect and need, by powering them with trusted high-quality data.

By attaching a semantic layer (built on ontologies) to the data, data can be moved to any user or system in the format that they need. Users would then have a representation of data that uniquely identifies and connects data in common business terms. This layer helps end users access data autonomously, securely, and confidently.

Abstract out complexity of data engineering

Data consumers with unequal technical abilities should be able to leverage the platform for their purposes. While non technical consumers should be able to self serve their authorized purposes via GUIs, technical staff should be able to perform their tasks declaratively (that is, by declaring what must be accomplished without describing an explicit control flow). This would dramatically reduce the data engineering resource overhead required to run such a platform.

DataOps friendliness

DataOps is a mix of practices and processes that utilizes automation to boost agility to enable improving the speed and efficacy of data analytics. A modern data platform must be built to adopt its principles. Hence a modern platform must have:

- Infrastructure observability to track and audit events, metrics, and logs — this is also critical with respect to the predictability of costs
- Developer-friendly CLI and local dev tools
- IDE integrations and enhanced YAML editor
- Multi-cloud and hybrid-cloud friendly

DataOS - The Fastest Path from Data to Decisions

DataOS creates the fastest path from data to decisions by delivering an agile and composable data operating system that delivers 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 lets business and technical users 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 into any data architecture such as a data fabric, data mesh, lake house, or many others, and delivers democratized access to trusted data right when you need it.

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

- **All things data:** DataOS lets you manage the entire lifecycle of your data workloads within one product. From ingesting data to metadata management, governance to observability, activating and sharing data — DataOS is a one-stop solution to go from data to decisions.

- **Composable, Open and Interoperable:** The first truly composable platform that lets users build the specific data architecture they require. It is built on open standards and formats so users always have complete control over their data and avoid vendor lock-in.

- **Enhance, Don't Replace your Existing Architecture:** Skip the hassle of having to rip your existing investments or move your data. DataOS weaves a modern connective tissue through your entire data stack. Users can augment and modernize current investments by leveraging the quality data and ontology built by DataOS.

- **Model-Driven Data Management:** Enable business-driven use and management of data. Streamline data pipelines. Democratize access to data with model query language instead of SQL

In order to understand DataOS' capabilities, it is useful to break them down into three layers.

Data Layer

Data OS allows the connection of data, whether its on-prem or cloud, without moving or copying it or storing data on an acid-compliant lake house, combining the best of warehouse and data lake technologies. It lets users optimize compute resources by providing the flexibility to isolate and assign them by teams or workload.

Knowledge Layer

Organizations can create an ontology of data that encodes relationships, metadata, and lineage, letting technical and business users discover, browse, and query fresh and high quality data and get trusted insights.

Activation Layer

DataOS enables users to activate and share data and insights across the organization. Users can create custom reports and dashboards on any dataset in minutes. It enables business users to drive their own insights with easily accessible and query-able data with model driven management. It is easy to build trigger alerts that can be sent to various destinations. Datasets can be shared in a secure and auditable environment.

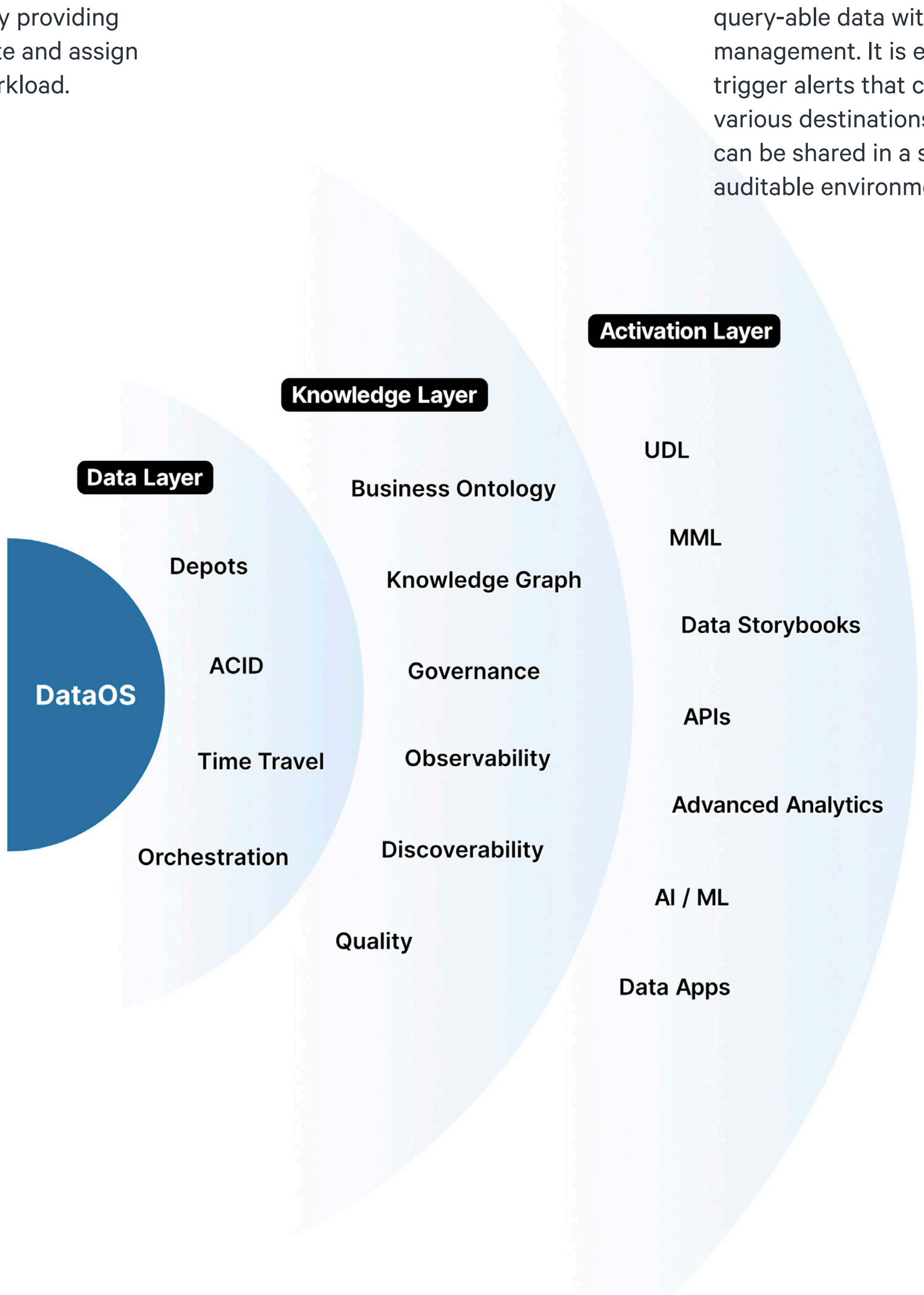

Table: DataOS Elevates Data Experience for All Personas

Persona		Landscape before DataOS	DataOS powered solution	Layer
Data Stewards and Custodians		Difficult to manage such systems at scale as it becomes complex to track and apply appropriate permissions to roles and to monitor usage	The DataOS tag-based governance gives teams an automated and scalable way to ensure governance and compliance for data across your ecosystem. Our attribute-based access control (ABAC) future-proofs your organization and lets it adapt to changing and new compliance regulations. With ABAC it is easier to apply policies dynamically at the time of querying based on who the user is and what they're attempting to do with data. Make granular decisions on what data authenticated users can access thru row-level filtering and column masking.	Knowledge
Data Product Consumers and Business Analysts		No standard metrics across the organization. Every data consumer creates their own copies after ascertaining lineage and provenance — there is no way to tell if similar tables already exist. Different teams may report different data for a business question and thus low overall confidence on data pervades. Valuable metadata gets lost between point solutions and non-collaborative systems.	Connect your data, whether its on-prem or in the cloud, without moving or copying it, or store your data on an ACID-compliant lake house, combining the best of warehouse and data lake technologies. DataOS lets you optimize your compute resources by giving you the flexibility to isolate and assign them by teams and workload.	Data
		Data access is often limited to a small number of credentialed users responsible for the data on a day to day basis. Access limitations are typically a result of concerns about protecting data from compromise, mis-use or mis-interpretation. Data cannot be leveraged if no one knows it exists. However, data is siloed across organizational structures. Discovery and access to data is cumbersome and routed through IT. Most often, ad hoc pipelines have to be built. Building them and ensuring compliance to organization's governance policy takes significant time and effort.	Create an ontology of your data that encodes relationships, metadata, lineage, and provenance and let your technical and business users discover, browse, and query fresh and high quality data and get trusted insights. The DataOS Datanet creates a connective tissue between all your data systems, continuously adding usage-based intelligence to your data. Easily discover, search, and find trusted and context-aware data. Understand the journey, relationships, and impact of your datasets and make decisions with confidence.	Knowledge
Data Managers	Data Scientists		Activate your data and share it across your enterprise. DataOS lets you create customized reports on any dataset in minutes, then share and collaborate on them securely. Build trigger alerts and notifications to 10+ destinations including ServiceNow, Jira, Pager Duty, Slack, MS teams, Pager Duty etc. Consume your shared data from SaaS systems like Salesforce, Zendesk, Google Ads etc.	Activation
	Infrastructure personnel	Systems do not adhere to DataOps principles. It is difficult to create and maintain pipelines. Engineers need to spend significant time performing repetitive data ingestion. Since data platforms change regularly, they also spend time building and maintaining, and then rebuilding, complex scalable infrastructure. Low latency data pipelines are required for real-time data, which are even more difficult to build and maintain. Data quality validation and monitoring is fairly time consuming as well.	Complete data observability into all your data activities from production to consumption of data to access and sharing of data including events, metrics, and logs. Create robust alerts across the data lifecycle by leveraging the DataOS observability of data pipelines, security and governance, and data management.	Knowledge
	Data Engineer		DataOS' foundational MML (Model-Map-Load) architecture brings a declarative model-first approach to pipeline creation. It abstracts away the complexity of pipeline building. MML makes it easy to build and manage data pipelines, helping data engineering teams to streamline and simplify their ETL processes. Out-of-the-box data quality and profiling functionality lets you get quality and trusted data without additional processing. DataOS comes with out-of-the-box data validation assertions including first-class support for custom assertions. Build trust and strengthen relations between engineers and stakeholders with SLAs for datasets, jobs, and workflows.	Activation
	Developers	Data access is often limited to a small number of credentialed users responsible for the data on a day to day basis. Access limitations are typically a result of concerns about protecting data from compromise, mis-use, or mis-interpretation.	With Universal Data Links (UDL) and GraphQL APIs interface, share data easily across your business ecosystem without copying or moving data. DataOS' best in class governance enables you to control access in a secure and auditable environment. Share data and collaborate with customers, internal teams, and external partners to unlock newer and richer insights.	Activation
	External System Call users			

Source: TABLE 6: AIR FORCE MAJCOM/FUNCTIONAL DATA PLATFORM 2.0 SYSTEM ACTORS
https://www.af.mil/Portals/1/documents/2019%20SAF%20story%20attachments/Tab%203%20SAF_CO_DSRA_Formatted.pdf?ver=2019-05-08-151010-237&timestamp=1557342896170

Table: Summary of DataOS Capabilities Addressing DoD’s Goals and Objectives

		Why you can’t solve without DataOS	**How you can solve with DataOS**
DoD Goals	**DoD objectives**	**Landscape before DataOS**	**DataOS powered solution**
Make Data Accessible	2.1 Data is accessible through documented standard Application Programming Interfaces (APIs)	Data access is often limited to a small number of credentialed users responsible for the data on a day to day basis. Access limitations are typically a result of concerns about protecting data from compromise, misuse, or mis-interpretation. Ad hoc pipelines have to be built by IT teams upon request. Building an ad hoc pipeline and ensuring compliance with an organization’s governance policy takes significant time and effort.	With Universal Data Links (UDL) and GraphQL APIs, interface share data easily across your business ecosystem without copying or moving it. DataOS’ best in class governance enables you to control access in a secure and auditable environment. Share data and collaborate with customers, internal teams and external partners to unlock newer and richer insights.
	2.2 Common platforms and services create, retrieve, share, utilize, and manage data		
	2.3 Data access and sharing is controlled through reusable APIs		
Make Data Interoperable	6.3 Public data assets are machine-readable and available for consumption		
	6.4 Rapidly mediate differing data standards and formats without mission- critical loss of fidelity, precision, or accuracy		Connect your data, whether its on-prem or cloud, without moving or copying it, or store your data on an ACID-compliant lake house, combining the best of warehouse and data lake technologies. DataOS lets you optimize your compute resources by giving you the flexibility to isolate and assign them by teams/workload.
	6.5 Develop and promulgate a data-tagging strategy and subsequent implementation plan to enable data interoperability		
	6.1 Document and implement data exchange specifications for all systems, including those of coalition partners		The DataOS datanet creates a connective tissue between all data systems and continuously adds usage-based intelligence to the data. Easily Discover, search, and find trusted and context-aware data. Understand the journey, relationships, and impact of datasets and make decisions with confidence.
	6.2 Exchange specifications contain required metadata and convey standardized semantic meaning with the data set		
Make Data Visible	1.1 Data is advertised and available for authorized users when and where needed	Data cannot be leveraged if no one knows it exists — however much of it is siloed across organizational structures. Discovery and access to data is cumbersome and routed through IT.	
	1.3 All data sources are cataloged		
	1.4 Implement common services to publish, search, and discover data		
	1.2 Implement metadata standards including location and access methods for shared data	Valuable metadata gets lost among point solutions and non-collaborative systems.	
	1.5 Warfighting and business governance bodies make decisions based on live visualizations of near real-time data	Ad hoc pipelines have to be built by IT teams upon request. Building an ad hoc pipeline and ensuring compliance with an organization’s governance policy takes significant time and effort.	Activate your data and share it across your enterprise. DataOS lets you create customized reports on any dataset in minutes, then share and collaborate on them securely. Build trigger alerts and notifications to 10+ destinations including ServiceNow, Jira, Pager Duty, Slack, MS teams, Pager Duty etc. Consume your shared data from SaaS systems like Salesforce, Zendesk, Google Ads, etc. Create an ontology of your data that encodes relationships, metadata, lineage, and provenance which lets your technical and business users discover, browse, and query fresh and high quality data to get trusted insights.
Make Data Understandable	3.1 Data is presented in a way that preserves semantic meaning and is expressed in a standardized manner throughout organization	There are no standard metrics across the organization. Every data consumer creates their own copies after ascertaining lineage and provenance — there is no way to tell if similar tables already exist. Different teams may report different data for a business question, resulting in overall low confidence in data.	
	3.2 Utilize a common data syntax for the same data types and includes semantic metadata with data assets		
	3.3 Data elements are aligned into a comprehensive data dictionary with a controlled, yet flexible, vocabulary and taxonomy		The DataOS datanet creates a connective tissue between all data systems and continuously adds usage-based intelligence to the data. Easily Discover, search, and find trusted and context-aware data. Understand the journey, relationships, and impact of datasets and make decisions with confidence. The DataOS tag-based governance gives teams an automated and scalable way to ensure governance and compliance for data across the data ecosystem. Our attribute-based access control future-proofs your organization and lets it adapt to changing and new compliance regulations. Users make granular decisions about which data authenticated users can access thru row-level filtering and column masking.
	3.4 Data is base lined and inventoried in comprehensive data catalogs with relevant information on purpose, ownership, points of contact, security, standards, interfaces, limitations, and restrictions on use		
	3.5 Processes to create, align, implement, and manage business vocabularies, including enterprise standards		Create an ontology of all data that encodes relationships, metadata, and lineage which enables technical and business users to discover, browse, and query fresh and high quality data resulting in trusted insights.
	3.6 Adaptive, intelligent systems monitor data streams and identify opportunities to transform, combine, or derive new data providing increased insights		DataOS’ foundational MML (Model-Map-Load) architecture brings a declarative model-first approach to pipeline creation. It abstracts away the complexity of pipeline building. MML makes it easy to build and manage data pipelines, helping data engineering teams to streamline and simplify their ETL processes.
Make Data Linked	4.1 Implement globally unique identifiers so data can be easily discovered, linked, retrieved, and referenced		The DataOS datanet creates a connective tissue between all data systems and continuously adds usage-based intelligence to the data. Easily Discover, search, and find trusted and context-aware data. Understand the journey, relationships, and impact of datasets and make decisions with confidence.
	4.2 Utilize common metadata standards that allow data to be joined and integrated		
Make Data Trustworthy	5.1 Budget requests and the supporting budget process integrate data-focused evidence and Learning Agendas		Out-of-the-box data quality and profiling functionality lets you get quality and trusted data without additional processing. DataOS comes with out-of-the-box data validation assertions including first-class support for custom assertions. Build trust and strengthen relations between engineers and stakeholders with SLAs for datasets, jobs, and workflows.
	5.2 Data has protection, lineage, and pedigree metadata bound throughout its lifecycle		
	5.3 Execute data quality management techniques to assess and enhance data quality		
	5.4 Implement master data management for business, intelligence, and warfighting data		
	5.5 Properly tag and maintain all appropriate data and records in accordance with established processes and policies		The DataOS’ tag-based governance gives teams an automated and scalable way to ensure governance and compliance for data across the ecosystem. Our attribute-based access control (ABAC) future-proofs your organization and lets it adapt to changing and new compliance regulations. With ABAC it is easier to apply policies dynamically at the time of querying on the basis of who the user is and what they’re attempting to do with data. Make granular decisions on what data authenticated users can access through row-level filtering and column masking.
Make Data Secure	7.1 Granular privilege management (identity, attributes, permissions, etc.) are implemented to govern the access to, use of, and disposition of data	It’s role-based access control (RBAC) model renders simplicity for small scale organizations. However, it is difficult to manage such a system at scale as it becomes too complex to track and apply appropriate permissions to roles and to monitor usage.	
	7.2 Data stewards regularly assess classification criteria and test compliance to prevent security issues resulting from data aggregation		
	7.3 Implement approved standards for security markings, handling restrictions, and records management		
	7.4 Classification and control markings are defined and implemented; content and record retention rules are developed and implemented		Gain complete data observability into all your data activities — from production to consumption of data to access and sharing of data including events, metrics, and logs. Create robust alerts across the data lifecycle by leveraging the DataOS observability of data pipelines, security and governance, and data management.
	7.5 Implement data loss prevention technology to prevent unintended release and disclosure of data		
	7.6 Only authorized users are able to access and share data		
	7.7 Access and handling restriction metadata are bound to data in an immutable manner		
	7.8 Access, use, and disposition of data are fully audited		

Addendum Table:
Definition of Personas

Persona		Definition
Data Stewards and Custodians		Establish policies governing data access, use, protection, quality, and dissemination.
Data Custodians		Responsible for promoting the value of data and enforcing policies.
Data Product Consumer		Consume and view business intelligence dashboards (visualization and presentation) which include reports and metrics, and interact through simple selection and filtering or visualization tools.
Data Managers	Business Analyst	Consume QA, QCed, and enriched (metrics, enrichment, analytic outputs) data products to create additional enriched data products, reports, and business intelligence dashboards (visualization/ presentation); will combine and select data products (pre-defined); create or flag simple metrics; not responsible for data quality or implementing analytical functions.
	Data Scientists	Provide analytical support to the business analyst or SME by supporting development of queries, indices, analytical applications and approaches, new visualizations for business intelligence and automation of data enrichment; acts as a more technical extension of the business analyst and SME.
	Infrastructure Personnel	Technical personnel who maintain the core technology infrastructure, including hosting, base software component maintenance, implementation of coded data operations, and installation of packages and capabilities — comfortable with operating at the command line.
	Data Engineer	Responsible for defining, building, and managing the essential services which ingest, validate, remediate, transform, and store physical data assets required for analytics or other data management functions.
	Developers	Leverage the infrastructure to add capabilities; those who extend the basic infrastructure capabilities and add new base operations including support for moving data enrichment into the data ingestion topologies — comfortable operating at the command line and in an integrated development unit.
	External System Call Users	An action triggered by a call from an external system to a AF MAJCOM/ Functional Data Platform API.

Reference: DOD Data Strategy & Data Services Reference Architecture
Note: The numbering of goals and objectives is as per DoD's document

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®: Fastest Path from Data to Decisions