PUBLIC

DataCTRL: Enterprise data catalog and data sharing system

Your trusted data solution – for when security isn't optional

Table of contents

NS2 Mission is pleased to present an overview of DataCTRL. DataCTRL provides organizations with a secure data cataloging, data transformation, analytics, and data sharing solution. To construct DataCTRL, NS2 Mission leveraged our extensive cloud, security, and data management expertise, ultimately creating a solution that fulfills the most complex customer needs. DataCTRL is scalable, redundant, and provides superior performance to other solutions within the market.

The Problem: Managing Mass Amounts of Data

As commercial and government organizations grow in size and technical complexity, it has become increasingly harder to manage and analyze massive volumes of data. Data organizations must satisfy their stakeholders and data consumers with timely, trusted, and appropriately formatted data, at scale, all while governing data access and data use policy. This increase in data consumption has generated the following challenges:

- It is more difficult to leverage enterprise data available to users, developers, and data scientists.
- Organizational data, content, and structure becomes unmanageable over time.
- Data teams need to comply with oversight compliance standards and best practices for data retention, deletion, and auditing.
- Sensitive data needs protection from attack and exfiltration.
- The sheer variety of data compounds create challenges in controlling and monitoring who can see what data.

How to Begin: Solving Data Management Issues

There are some basic steps that organizations can take in order solve these challenging data management and analytic problems. Organizations can take advantage of Open-Source tools for data processing and establish a team to assemble these tools into working pipelines. An organization can also decide to custom develop a framework that solves these issues. Unfortunately, these solutions will expose sustainability and maintenance issues as projects grow in complexity. Organizations will need dedicated development teams to support either solution. There is also no corporate accountability unless organizations leverage support agreements with the Open Source providers.

So, what options do we have?

The Solution: DataCTRL

Fortunately, a solution exists that allows an organization to quickly stand up a data cataloging and sharing system that supports today's real-world data challenges: **DataCTRL by NS2 Mission.** DataCTRL meets the challenges by:

- Connecting to existing datastores, data bases, cloud storage, and many more data sources to illuminate data discovery capabilities across your data landscape.
- Providing peace-of-mind through government-grade auditing with oversight of data use and access.

- Centrally implementing native policy enforcement.
- Securing your data with data encryption for data in-transit and at-rest.
- At-scale data discovery, analytics, and insights through a robust data modeling capability which postures data for a multitude of development, analysis, and machine learning use cases.

DataCTRL is NS2 Mission's core data management solution, populating a data catalog engineered to satisfy the stringent security and data access policies for the most secure government organizations. DataCTRL enables programs with data challenges to succeed in today's evolving data landscapes. It's designed for ease-of-use at every step – from installation, to development, to monitoring. It allows a development team to quickly stand up a data cataloging and sharing system to ingest, track, process, and route datasets intelligently. DataCTRL ensures the system will be secure, fault-tolerant, and scalable. The framework allows developers to build projects consisting of custom cloud-ready workflows, easily deploying, monitoring, and managing them in real time as data flows through the system.

DataCTRL supports flow and dataset isolation, ensuring that each project or dataset can have its own access controls, configuration, processes, and reporting. Users can be granted access to multiple projects and datasets. Administrators can delegate appropriate powers to the project and dataset owners, empowering them to manage access to their data.

Figure 1: What key functions does DataCTRL address?
DataCTRL offers all-encompassing support and functionality – from data modeling and data sharing to government-grade security protocols.

Features of DataCTRL

DataCTRL is a secure data platform designed to handle continuously updated datasets, facilitate data discovery, and accelerate data sharing and analysis. DataCTRL tackles data challenges while providing granular access controls and robust auditing to help execute your mission-critical operations.

DataCTRL Catalog

DataCTRL Catalog provides a framework for managing data deliveries, exposing storage repositories, tracking data pedigree & lineage, managing data structures, and inspecting data for consistency. The Catalog is designed

to handle continuously updating datasets, facilitating discovery of new datasets, and detecting changes or interruptions to existing datasets. It provides granular access controls for datasets without preventing authorized users from discovering new data.

The Catalog promotes the reuse of common processing capabilities, such as categorization, indexing, and text extraction. Downstream systems can use the Catalog API to discover and export data into external systems. Catalog is your hub for content profiling, discovery, and inventory of your enterprise's known data assets.

Data Transformation and Pipelining

DataCTRL works with and through your existing data managements capabilities. You are empowered to leverage existing data transformation solutions (Spark, Hadoop, commercial, etc.), but can also utilize the powerful and scalable ETL solution provided by DataCTRL. DataCTRL enables user-designed data processing pipelines to run in a scalable, fault-tolerant, distributed cloud environment. This allows engineers to deploy both scheduled and continuously processing data transformation tasks. Data processing flows support runtime configuration, custom error handling, and 3rd party software integration. These pipelines scale to meet demand, with the ability to run thousands of tasks in parallel. It is designed to be fault tolerant by handling software or network errors, rebalancing the work, and allocating new resources to replace lost ones. The solution exposes a simple UI for deploying and managing user flows. The UI is used to monitors all processing, providing users with real-time status updates, processing failures, and performance bottlenecks as data moves through user flows.

API

DataCTRL exposes three core data management Application Programming Interfaces (API) that allow external developers and systems to interact with the platform. The Catalog API exposes hundreds of RESTful webservice endpoints for all data catalog operations. It allows users to publish data, manage data flows, monitor data movement, manage P&L, export data, and hundreds of other catalog related operations. The Transformation API allows users to deploy and monitor the ETL pipeline. Systems and users can submit ETL tasks directly into the system and directly interact with running transformation tasks. The Authentication API is used to make access control decisions within the system and can be leveraged by external applications to make similar decisions.

Intelligent “Floating Data Warehouse

DataCTRL not only processes your data at scale, while generating a robust metadata catalog, but it also postures your content in a way that provides granular data sharing services and promotes advanced analytics from your data. We describe this capability as the Intelligent Floating Data Warehouse (iFDW). The iFDW capability enables the rapid ingest of data with virtually infinite scale, like a data lake. iFDW also utilizes semantic context to meaningfully query across data, from many source systems, like a data warehouse – with intelligence, not a “Data Swamp”. This allows DataCTRL to interact with your data, with the elastic and scalable qualities of today’s data lake, while utilizing the semantic qualities of a data warehouse. This capability unites diverse data sets for aggregated analysis, ensures increased data quality, enables provisioning of trusted data to consumers, while adding enhanced governance and access control safeguards.

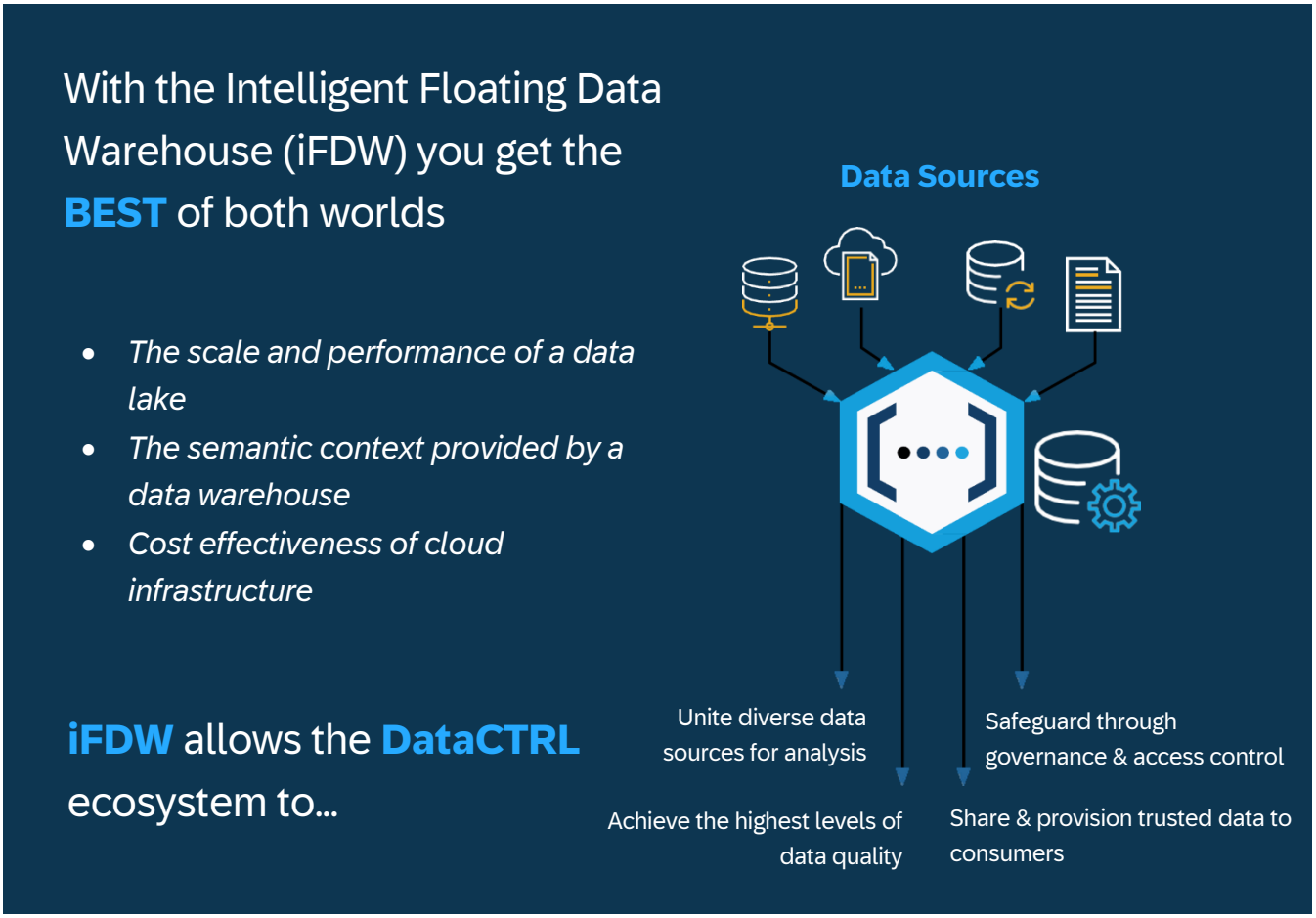


Figure 2: What are the benefits of iFDW?
iFDW combines the capabilities of a data lake and a data warehouse to scale at speed, save on costs, and ingest data effectively.

Data Modeling Levels

Leveraging their substantial data modeling experience, the NS2 Mission team implemented data processing levels within DataCTRL. Data processing levels are used to describe the amount and type of processing applied

to a data set. They classify how the data is stored throughout its life cycle. The processing levels allow data-driven organizations to satisfy multiple stakeholders and downstream applications with the right data, at the right time, in the format they need to best execute their mission. These processing levels correspond with various use cases that promote on-demand data transport, analysis, or data science workflows. This flexible ability to condition and posture data for efficient end-user analytics, at scale, is foundational to successful implementation of the iFDW. There are five primary data levels, but the concept can be extended to fit specific customer environments.

- **Level Zero (L0) -** Data is stored as received from the data provider. The file format will depend on the source. Typically, only data owners would have access. Data is in its raw format and no processing has been performed to reformat the data. This level is retained as a reference in the event data needs to be reprocessed or there is a question about the original value of a data element.
- **Level One (L1): Data Lake** - Data is stored as records using its *native data model*. The schema will depend on the source. Data is stored using a consistent file format. It eliminates the laborious task of converting files from one format to another and provides a versioned history of all schemas for data modeling. Data has been repackaged but no processing has been applied to the data. This level is used to stage the data in a structure optimized for processing.
- **Level Two (L2): Data Warehouse** - Data is stored as records using a *Common data model* (CDM). Records and fields are standardized to make it easier to use data from different sources repackaged and processed to facilitate analysis. Examples include creating de-normalized records and data enhancement like geocoding.
- **Level Three (L3): Data Mart** - Data is stored as records using an *Analytical data model*. They are optimized for analyzing data and Data Scientist tasks. The file format will depend on the data mart that will consume the file. For example, relational databases would typically consume delimited text files.
- **Level Four (L4)** - Data is stored as records using a *data dissemination model*. The file format will depend on the consumer. The schema will depend on the memorandum of understanding (MOU) between the data owner and the consumer. They are both optimized for transport.

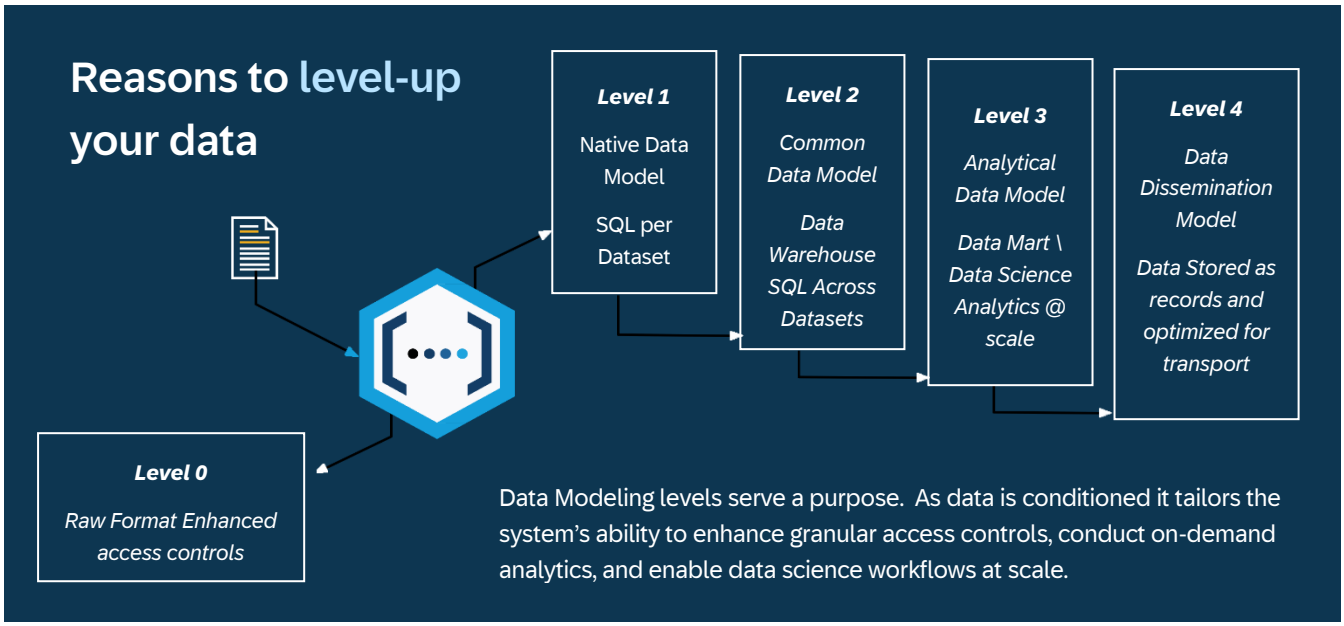


Figure 3: Reasons to level-up your data
How does data modeling impact the way your data is stored and handled?

Security

DataCTRL encrypts all data in transit, including communication with web applications, communication with third party systems, and internal communication. Network resources are configured using security best practices and a defense-in-depth approach to minimize attack vectors and restrict access to internal resources. All data stored in external storage is encrypted, i.e. S3, EC2, and RDS instances. Auditing is performed across the entire system, and audit logs can be delivered into enterprise audit systems. The solution also supports third party auditing for its application and API layers.

User Management

DataCTRL uses enterprise caliber Role Based Access Control (RBAC) for authorization and access control throughout its system. Users can have different roles for different parts of the data layer, minimizing unnecessary access. DataCTRL leverages enterprise user management systems, if available, allowing existing groups and permissions to map to system functionality. In addition, user management service is provided to third-party applications, acting as a single point of user management for multiple applications. The framework can also operate in a multi-tenant system. Different groups may use the same data management instance without interfering with or accessing each other's content.

Architecture

DataCTRL is purposely built to enable programs with data challenges to succeed. DataCTRL allows developers to build projects consisting of custom cloud-ready workflows, easily deploying, monitoring, and managing them in real time as data flows through the system. The DataCTRL architecture is designed to provide secure data cataloging while conditioning your data to support a wide variety of workflows, ranging from content discovery to secure data delivery, and machine learning capabilities. This leverages modern and scalable hyperscaler infrastructure to handle the most challenging data landscapes, operating at mission speed, with data sources ranging in volume and variety. In short, DataCTRL allows data driven organizations to tame chaotic data environments, while providing data cataloging and sharing services, through secure intelligent access for users, developers, and data scientists.

Tame Chaos

DataCTRL provides scalable enterprise proven Extract Transform & Load (ETL) framework codenamed DataCTRL Streams.

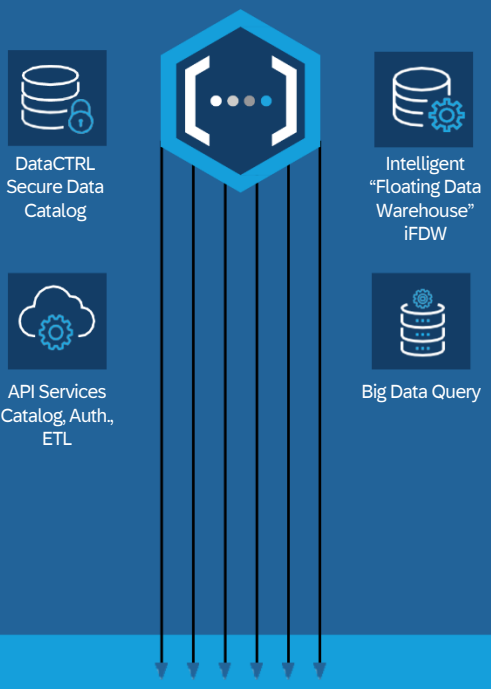
DataCTRL is capable of ingesting from multiple source systems, storage mediums, and can interpret a wide variety of data formats. Data can be pushed into the system or extracted from other systems.

DataCTRL can also leverage open-source ETL Tools and commercial data management capabilities.

Catalog & Share

All data is cataloged as it passes through the platform. The Data Catalog tracks data deliveries, transformations, pedigree, lineage, exports, and a multitude of other fundamental data management tasks.

DataCTRL exposes core API services for a cataloging, authentication, and transformations. External systems interact with these APIs to facilitate data movement, manage user's interaction with the data, and to author ETL operations.



Data can be transformed through various data modeling levels and stored in iFDW. The iFDW provides a virtual storage abstraction on top of physical data stores like S3, sFTP, Blob Storage, and other mediums. All data processed through DataCTRL can be cached within the iFDW.

DataCTRL includes multiple big data query solutions including SAP Hana, Presto, Elasticsearch, and Apache HBase. As data is transformed, data models can be stored within a query solution most appropriate for analyst's needs.

Secure Access

DataCTRL *tames chaotic data* environments, while providing *data cataloging and sharing* services, through *secure intelligent access* for users, developers, and data scientists.

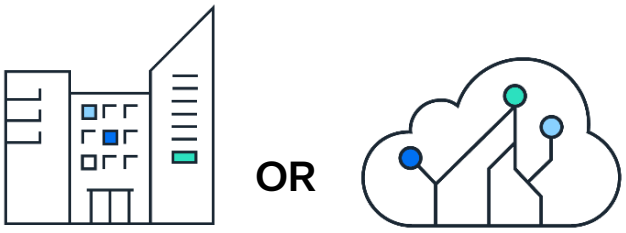
Figure 4: DataCTRL Architecture
DataCTRL is built to tame data chaos, create a platform to catalog and share data, and correctly secure entry points to your data sources.

Deployment Options

DATACTRL is cloud native, but also provides various deployment options tuned for your data landscape.

These options include on-premise systems and cloud based deployments in AWS and C2s.

DataCTRL is Accredited for use on the IC's Commercial Cloud Services (C2s) contract.



Summary

NS2 Mission provides DataCTRL so organizations with the most stringent security requirements can maximize their data experiences. DataCTRL securely catalogs and harvests metadata which empowers data managers to steward data to those who need it, all while obfuscating it from users without the appropriate credentials. DataCTRL can invoke the ability to condition your data at scale, promoting a wide variety of use cases. This allows data engineers, stakeholders, and analysts to conduct on-demand data discovery, analytics, and machine learning workflows.

About NS2 Mission

Volume Integration, LLC. doing business as **NS2 Mission** is a wholly owned subsidiary of National Security Services, Inc. (SAP NS2®), which is a wholly owned subsidiary of SAP® SE, the world's largest enterprise software company. NS2 Mission is a US corporation, staffed by US citizens with appropriate security credentials. NS2 Mission has facilities in the continental United States, with headquarters in Reston, Virginia.

Now, let us help you. Learn more at sapns2.com/DataCTRL.

For more information, see www. SAPNS2.com/Mission or call us at (703) 860-0865.

NS2 Mission | 703-860-0865 | www.sapns2.com/mission