ENTERPRISE e-SUPPORT INC.

EeS White Papers

Are you theorizing without data?

April 01, 2020

Tax ID: 20-8128250; **SBA 8(a) Certified**; GSA IT Schedule 70: **47QTCA18D00KL**

6564 Loisdale Ct., Suite 600, Springfield, VA 22150

☎ (703) 459-9705 **|** 📠 (703) 768-0547

Email Contact to Register: info@enterpriseesupport.com

Website: www.enterpriseesupport.com

"It is a capital mistake to theorize before one has data" - Arthur Conan Doyle

The issue of data unlocking

Cutting-edge drones and robots with the ability to make decisions autonomously, algorithms that process enormous amounts of data to identify national security threats, and smarter defense systems are slowly manifesting in a new stage of modern warfare spearheaded by developments in Artificial Intelligence (AI). The Department of Defense allocated nearly $1 billion for AI research and development in 2020, writing that continued investment in AI is essential "[t]o prevent the erosion of the U.S. competitive military advantage." Amid the AI hype, officials have pushed a core question to the side: **where is the data?** AI and machine learning (ML) algorithms must be trained on massive amounts of data before they are ready for real-world use, yet this data is often hard to discover and disorganized. In cases where individual agencies organize and label their data in different ways, it also becomes hard to compile data across agencies. Additionally, over time, data standards and vocabulary within agencies may weaken, hindering their ability to efficiently unlock and utilize specific data assets for their needs. Data governance, a collection of practices and processes to ensure formal management of data assets in organizations, is necessary to structure data and subsequently extract insights from them. Data sharing, the ability to share the same data assets with multiple users, is necessary to build larger datasets within and across agencies that can then be used to train AI and ML algorithms.

Strong data management is important for many other reasons as well. Disaster response necessitates that agencies be able to quickly discover certain data assets to inform their decisions for action. Weak federal response to Hurricane Katrina in 2005 was in part due to the dispersed nature of FEMA and the DHS, their inability to efficiently function in tandem, and disconnect with local and state resources. With better data governance and sharing across agencies, a quicker response may have been actualized. Effective data sharing and management could also reduce unnecessary bureaucratic expenses by decreasing discrepancies within and across organizations.

Policymakers have recognized the necessity of strong data management. The Evidence Act of 2018 mandates certain data management processes agencies must follow and offers guidelines on data use to inform policy decisions. Additionally, the Federal Data Strategy 2020 Action Plan lays out goals and principles for data governance for the next decade. Both these documents are new and need time to be implemented as necessary tooling and operational support is lacking to follow best practices at scale. In the status quo, federal agencies continue to operate using non-ideal databases and governance techniques, a roadblock for the future.

What should agencies be doing?

Agencies should be embracing and utilizing data governance practices to build data asset catalogs and use a common vocabulary along with proper definitions that provides detailed context to their data assets. They need to use a commonly understood "language" and define data assets in the same way as others to ensure valuable data isn't "lost in translation." Currently, different organizations may use different vocabularies to define their data assets, making it hard for them to collaborate. Organizations do not have to start from scratch in building this shared vocabulary either. The National Information Exchange Model (NIEM) is a common vocabulary that "enables efficient information exchange across diverse public and private organizations, locally, nationally, and internationally." Implementing NIEM helps reduce redundancy and maximize efficiency by providing constant, reusable data terms that are already organized within hierarchical data types that make it easy to understand specific information exchanges.

Given that federal agencies may have hundreds of databases, each with thousands of data elements and millions of records in them, implementing NIEM is not a trivial task. Organizations need to map their data assets to taxonomies to build catalogs for data discovery. They also need to align their data elements to a common vocabulary to be understood. Doing all this may require professional help or tools and likely necessitates strong industry-government collaboration.

What benefits are there to organizing data?

Once agencies have created data asset catalogs, or organized their data within structured hierarchies, their internal operations will benefit as data can be unlocked swiftly. The process of optimizing one's data environment helps identify and remove redundancies, inconsistencies, and obsolete data. An optimized data environment affords users the ability to rationalize their data portfolio and provides a starting point to define a single source of truth for their data needs, allowing them to find, understand, and unlock their data more easily. An optimized data environment also helps improve real-time performance to satisfy customer expectations in a dynamic environment and increases return on investment from IT infrastructure. Additionally, data stewardship and defining structure around data access is hard when data is disorganized. Improved data stewardship can improve data access and establish responsible data sharing within and across organizations. Data may have sensitive elements that need to be protected. Proper data governance for privacy can also make it easier to flag sensitive information and mask data. Organizing one's data makes it easier to conduct data analytics and extract important business intelligence that could provide organizations with a competitive advantage. It also

opens the door for self-service: if data is already organized, agencies do not have to call IT data custodians for their analytics needs, rather they can smoothly traverse data assets on their own.

Operationalizing data environments using EeS Nishkarsh data products

Modernizing data architecture practices and improving data governance at the enterprise level is essential to achieve an organization's future goals, maximize efficiency in its current operations, and increase interoperability. It further provides visibility into data architecture and empowers organizations to mind map and control ongoing changes. Yet, in an age of ever-increasing data, data science and operationalizing data management and governance techniques is becoming tougher. Big data is a field that deals with ways to store and process large amounts of data that traditional data-processing platforms cannot handle. Learning about big data concepts such as HDFS and MapReduce, which splits input data into smaller chunks and queries them simultaneously in parallel, is key to helping organizations organize data assets on a larger scale. Having a centralized data source as such is preferable to expensive point-to-point data management because it affords the user a bird's eye view of their data. This efficient, low cost, fault tolerant data environment is the answer to the big data problem.
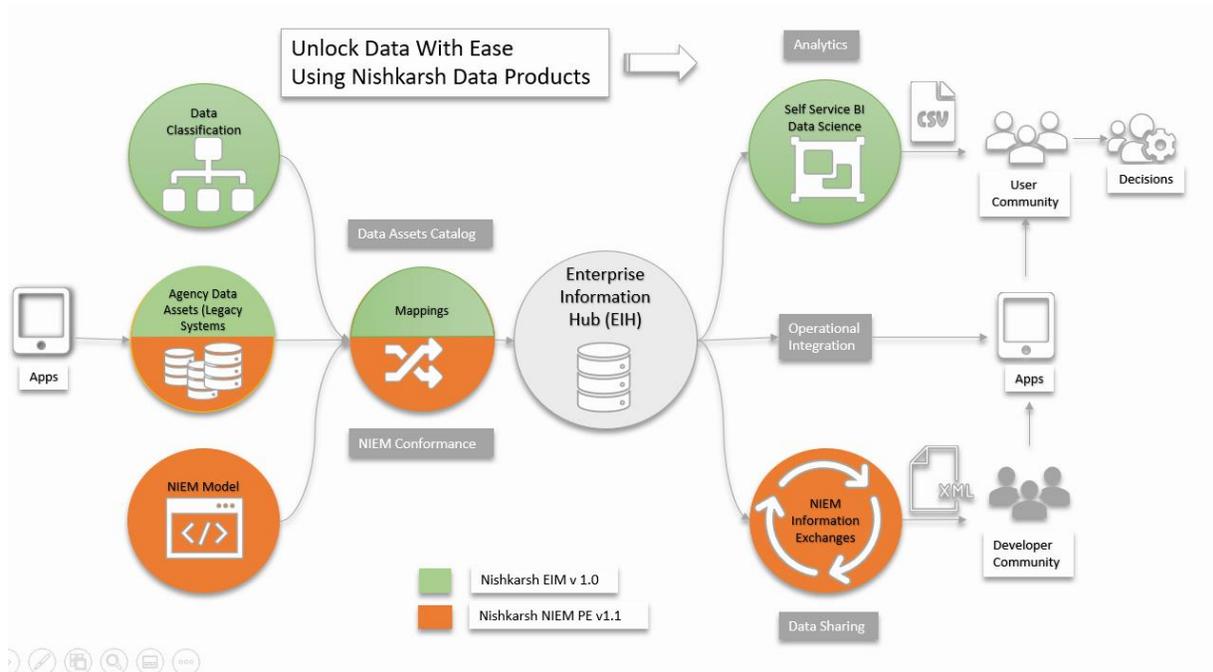

Figure 1: A conceptual view of a typical data environment and how to unlock data using EeS data products.

An **Enterprise Information Hub (EIH)** can serve as a single source of centralized data and rationalize master and reference data across an organization hosted in a big data environment. The Hub architecture and data movement is controlled by metadata, data about data. This could come from existing metadata repositories or be captured as data moves. Metadata can include everything from data element names to type definitions to the number of rows in a table to file formats and type hierarchies. Once metadata has been collected from an organization's existing data repositories, it needs to be processed and ingested to help form universal hierarchies that **data assets can be mapped to**. Based on business needs, Enterprise Data Architects may design staging, modeling, and rendering of data for consumption into the Hub for operational, analytical, and data sharing needs. The Hub will also use a variety of ingestion, metadata and access services to move data in, transform data, and send it out for consumption.

Many large IT organizations are seeking help to build and support such EIHs. Enterprise e-Support self-service products, the EeS Enterprise Information Management (EIM) tool and EeS NIEM PE tool help users locate and share data inside and across agencies.

In order to modernize organizational systems, users must first find and organize data into a data asset catalog. Enterprise e-Support's Nishkarsh Enterprise Information Management (EIM) tool helps users do this by offering a simple interface that aids customers in managing their data assets by creating a standard-based hierarchical classification for their metadata. This creates lines of sight into the users' data that allows them to perform data analysis and unlock novel insights and relationships they may not have seen before. Furthermore, the tool optimizes a user's data environment, helps consolidate data, and redefines data stewardship boundaries. This newfound visibility and control of data helps organizations make accurate decisions efficiently.

When shared across organizational boundaries, data may lose its meaning due to the lack of a common vocabulary across organizations. The National Information Exchange Model (NIEM) is a common vocabulary that enables efficient information exchange across diverse public and private organizations. NIEM is already being used at places like the DoD, biometrics organizations, DHS, and many more to exchange data with partners. Enterprise e-Support's Nishkarsh NIEM tool helps organizations add the NIEM common vocabulary to their data elements before sending it out to their partners. EeS's NIEM tool has already been vetted by the NIEM program and is listed on the NIEM.gov tools catalog.

Nishkarsh Data Products – Summary Sheet

Nishkarsh EIM v 1.0
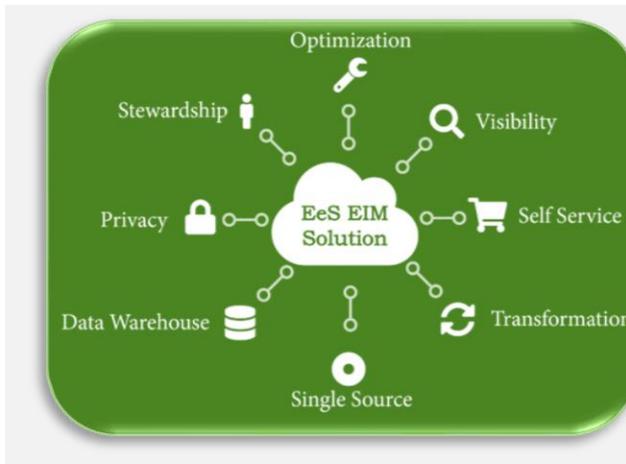
Nishkarsh NIEM PE v 1.1

Benefits	Features	Benefits	Features
• Optimization	• Classification of Data	• Standard Vocabulary	• Preloaded NIEM
• Visibility	• Data Assets Connection	• Common Context	• Search/Edit NIEM
• Self-Service	• Data Dictionary	• Persistent Mapping	• Import Data Assets
• Transformation	• Permissions Management	• Machine-Readable Data	• Browse Data Assets
• Single Source	• Reporting	• Validate Schema Subset	• Map Data Assets-NIEM
• Data Warehousing	• Search	• End-to-end Support	• Workspace for IEPDs
• Privacy	• Visualization	• Platform Agnostic (Windows, Linux, MacOS)	• Full IEPD Lifecycle
• Stewardship	• Data Collectors	• Out of the Box Start	• NIEM Conformance
• Records Management	• Import/Export	• Low Cost Desktop Edition	• Reporting
• NIEM Standardization	• Impact Reports		• Built-in XML editor
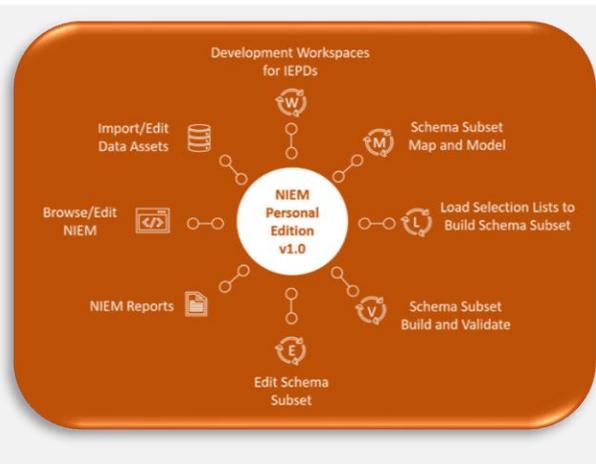
Conclusion

In a recent RAND report to the DoD, senior officials commented that "deploying enterprise AI at scale in an organization requires close cooperation with both the organizational entities in charge of data and data management and those in charge of storage and computing infrastructure ." Ineffective collaboration across agencies and inefficient practices within them need to be addressed. America's 21st century technical prowess relies on agencies better organizing their data and doing so requires

modernizing data management and data governance practices. Agencies need to better organize, manage, and automate the flow of their data before they can reap the benefits of futuristic AI and ML technologies.

Your organization is expected to respond to situations in diverse ways, either on its own or jointly with your partners, anywhere, at any time and under changing conditions. This requires a close examination of the inhibitors in your data delivery systems and how your data strategy will overcome them using data governance, augmented intelligence, and automation. Data must carry its meaning and flow uninterrupted in a structured manner via the man-machine ecosystem to benefit all and keep the organization's response healthy.