WHITE PAPER

people. purpose. impact.

"Within a larger Data Enterprise Architecture, a well-planned and tiered Data Lake is the foundation for meaningful data–and even more importantly–meaningful business analytics."

–Jason Carter, President, UNCOMN

FROM DATA SWAMPS TO DATA LAKES

Making Data Manageable for Business Analytics

Joshua Leesmann
jleesmann@uncomn.com

UNCOMN White Paper Series | November 2020

Contents

Figures

Executive Summary

Introduction to the Challenge | Putting all your data in a single place will not, in and of itself, solve the data management challenge.

Disparate data and groups within the enterprise are often confronted with a need to modernize legacy systems and develop methods to easily analyze and identify trends, while leveraging investments to reduce redundant use of resources (both financial capital and human capital).

UNCOMN provides this white paper as a vehicle for discussing lessons learned in making one's data manageable for business analytics and leveraging investments in technology transformation across the enterprise.

Client's Big Data across a Big World | The United States Transportation Command (USTRANSCOM) had this challenge with their data.

Over any week, USTRANSCOM conducts more than 1,900 air missions, with 25 ships underway and 10,000 ground shipments operating across 75% of countries globally– all of which generates terabytes of data that need to be curated for business decisions.

In order to provide the best roadmap, TRANSCOM invested in a process that:

- Defined the data lake's boundaries and data roles/personas within the context of the larger enterprise
- Utilized a tiered approach to metadata management (in the context of the objective and goals of the enterprise)
- Met compliance objectives around data governance and quality control
- Included an automation-forward mindset to identify efficiency and effectiveness opportunities

To start, USTRANSCOM did not have one system, one process, or one data warehouse that maximized the efficiency of big data. Many data gaps and inabilities to achieve data visibility for quick business decisions across their enterprise data environment (EDE) stemmed from big data emerging from disparate legacy systems–systems that did not share similar data elements.

Because of the different data models using different data elements in each of these legacy systems, there was no way to easily analyze and identify trends and correlation between elements of large data sets. In short, USTRANSCOM's entire EDE resembled more of data swamp than a clear and curated data pool.

UNCOMN Solution | After analyzing and discussing the challenge at USTRANSCOM with their big data, UNCOMN and our partners recommended a wholistic solution that enabled an enhanced "data lake" from within the larger enterprise data environment.

We performed our enterprise data modernization solution for USSTRANCOM under two contracts: the 5-year the Enterprise Architecture, Data, and Engineering (EADE 2) contract, worth $175 million; and the shorter 2-year the Big Data Analytics using Enhanced Data Lake (BDAEDL) contract, worth $400 thousand.

Applying Data Lakes

Modernization

Enterprise Data Modernization | For our USTRANSCOM projects, UNCOMN applied our solution within a wholistic approach.

UNCOMN's solution hit upon the following points in our modernization process:

- We made sure to define the data lake's boundaries within the larger enterprise data environment (EDE)
- We worked with USTRANSCOM to clearly define the personas (data roles) downstream from the data lake
- We utilized tiered metadata
- We developed with USTRANSCOM a plan for data governance
- We included automated processes where it was most advantageous
- We defined and implemented quality control for the future

Enterprise Data Environment

Data Flows across the EDE | UNCOMN worked closely with the Government to make sure that any data lake would be properly positioned and defined within the larger EDE. As big data flow through an organization, any and all data that flow through a data lake must be curated properly (*Figure 1*).

Any data lake can be curated in different ways. For USTRANSCOM, UNCOMN explained that not all data have the same use. Just because it is now relatively easy to collect data, we still needed a clear goal in mind. A data lake can transform into a data swamp when Government agencies or commercial firms do not set parameters about the kinds of data they want to gather and why.

We worked to set limits on data amounts, so that USTRANSCOM could did not end up in the trap of establishing a well-organized data lake, only to have it turn back into a data swamp flooded with information they may never need. We also researched any subagency silos within USTRANSCOM, so as not to exacerbate the common problem of gathering data without rhyme or reason.

Because different offices had differing opinions about which kinds of data are most useful to USTRANSCOM at a given time, we orchestrated a plan that considered the final business analytics that USTRANSCOM required at various levels. Working back from those final requirements to clearly defined goals about data usage.

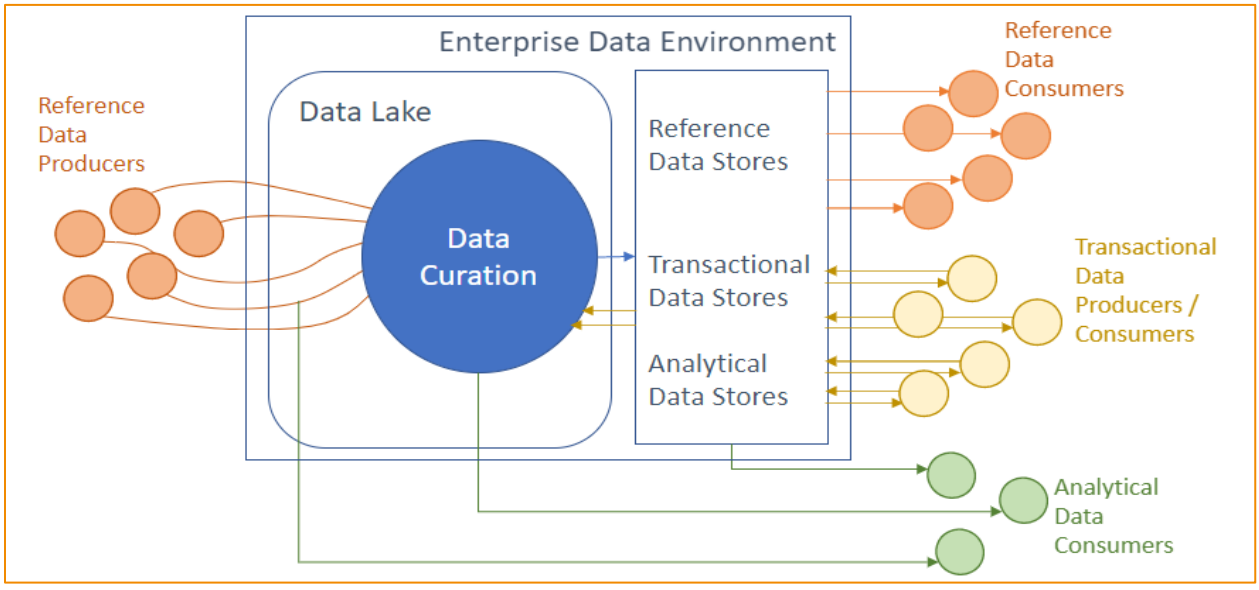

Figure 1. Data Flows

We established parameters to help prevent over eagerness when collecting data.

Personas

Not all Personas Touch the Data Lake Directly | In typical organizations, staff will interact with data in different ways, depending on their role or departmental function. A key component of highly successful organizations is to recognize that while there are different levels and perspectives, the source data should be the same. From this common dataset curated by the data lake, customized dashboards and model interactions can later be created for any number of personas.

Some personas, like data engineers or scientists, will require very little assistance as they interact with data as it flows into the organization and pools in the data lake. Other users, like key decisions makers, may require simplified persona(s) that only interact with the data lake via dashboards for specifically defined business analytics.

For the USTRANSCOM contracts, we worked with the Government through an iterative process to design data reports, dashboards, and analysis workflows to gain understanding of the "day in the life" of each identified persona and the types of information or modeling decision support needed for them to perform their job.

These designs were then incorporated into the user interface. Dashboards contained tabular data, visual graphs, tracking information, geographical and linear maps, and more.

After context-specific analysis of USTRANSCOM, we defined "persona" for their dataset as a stakeholder or operator, who has some need for information or decision support tools to perform his or her job.

Recommended Personas | Below are some of the common personas that UNCOMN often will recommend for our Government clients:

- **Data Engineers:** need ability to generate a data module, mature data through this process, and publish and maintain that dataset for the enterprise
- **Data Scientists:** need ability to create and train models to support Prescriptive and Predictive Analysis with Data Science Platform Capability
- **Governance Teams:** may be part of any Enterprise Data Management Team (EDMT) or Office (EDMO) and supporting governance bodies established across an agency to provide oversight of the data
- **Data Stewards:** assigned to provide stewardship over data maintained in the EDE, so they can collaborate with the EDMO or similar office within their organization and the consumers of data they are assigned to manage
- **Analysts:** need ability to consume data to conduct descriptive analytics to understand what happened based on received data
- **Process Owners:** need ability to consume data to conduct descriptive analytics to understand what happened with a process they manage
- **Operations Center Personnel:** need ability to consume data to conduct descriptive analytics to understand what happened based on current operational data
- **Key Decision Makers:** either Senior Decision Maker or are part of supporting staff that are incorporating data into their decision-making process
- **Cyber Analysts:** involved with the security and operations data being generated and managed

- **Systems Owners:** manage or support a system that produces, consumes, or does both within the EDE
- **Architects and System Engineers:** need to find data sources to support their projects throughout the design, build, deploy, and operate lifecycle

At the very least (and if the above seems too complicated), another way of thinking about data roles is simply to break them into three persona groups: Data Science, Functional Power Users, and Systems Dashboard Users. In this conception, how "deep" the user is in data defines the user grouping (*Figure 2*).

Metadata

Tiered Metadata | Metadata is key to understanding a data lake. It is information that describes other data. When appropriately used within a data lake, it acts as a tagging system that enables people to search for different kinds of data.

UNCOMN worked with USTRANSCOM to understand metadata can also create a tiered storage structure that stops a data lake from turning into a data swamp.

We organized their data with metadata tags denoting the source of the data or how it relates to some business analytic. We established a process to tag data as it flowed through the EDE in multiple ways.

Governance

Data Governance | At UNCOMN, data governance is used to define how to treat data, who should handle it, where the data goes, how long an agency or office retain the information, and more. Excellent data governance is what equips our Government clients to maintain a high level of data quality throughout the entire data lifecycle. Data Lakes require intelligent data governance.

Data Management Association Framework | For the USTRANSCOM contracts, UNCOMN planned and supported the development of a governed approach to ensure that data was fit for purpose. We made sure to align our processes with both Data Management Association (DAMA) Framework and the needs of our Government client.

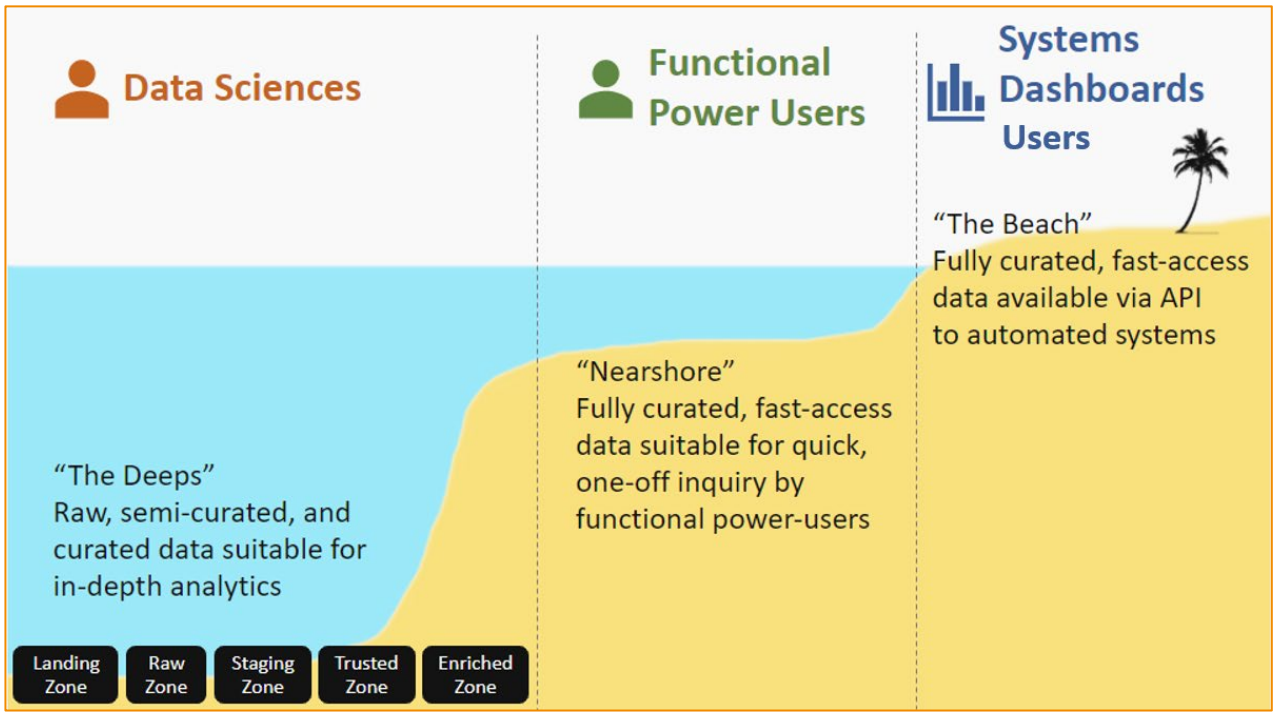

Figure 2. Personas Grouped by Depth of Data

We worked with both key decision makers and the specific data stewards to support definition of standards, requirements, and specifications for data-quality controls. The UNCOMN team collaborated them to refine and identify a data-quality assessment framework. Results of the assessment supported data stewards in exercising authority, control, and shared decision making (to include planning, monitoring and enforcement) over the management of data assets.

We aligned data governance to a Raw-Managed-Published process, so that data stewards were responsible for governing data throughout the maturation process. This enabled governance to be decomposed into bite-sized chunks based on a physical data set. In so doing, we helped translate the DAMA governance framework into best practice daily operations.

Governance at the Source | As each source of data was loaded it began to be governed, curated, documented through artifacts, and then—once fully matured—published to the enterprise for consumption. As each data set was published, we advised and assist the Government in developing metrics to monitor timeliness, accuracy, completeness, validity, consistency, and uniqueness. Along with quality metrics, we assist in developing performance and usage metrics supporting enterprise data users.

As the EDE matured, UNCOMN provided monitoring services to ensure data quality and compliance were maintained, keeping governing bodies informed of EDE operational status and data quality issues.

No Data Governance Means No Data Lake | We learned that in the absence of rules stipulating how to handle data, everything gets dumped in one place with no thought of how that practice negatively affects the future use of the data. Failing to implement data governance also puts organizations at risk for landing in regulatory hot water, especially if they get audited.

Making data governance a priority as soon as Government agencies start the modernization process is crucial.

Automation

Automated Processes | UNCOMN has learned that if an organization has not entertained the idea of applying automation to help maintain a data lake, then their EDE can quickly become a data swamp before people realize what is happening.

Automation is becoming increasingly crucial for data lakes. It can do things such as standardize data usage practices across platforms and process all raw data in the same ways.

In the early phases of establishing a Data Lake within the EDE development for any organization, UNCOMN recommends leveraging semi-automated processes for creation, sourcing, storage, integration, and delivery of metadata—like we did with USTRANSCOM. Furthermore, as the EDE matures, machine learning can be used to support fully automated metadata tagging.

Automation Still Requires Planning | It is worth noting that bringing automation into the equation does not excuse key decision makers from ironing out a plan for how to use data.

UNCOMN recommends that they would need to settle that aspect first, then figure out how automation can help decision makers achieve their identified goals for how they intend to use business analytics.

Quality Control

Removing Errors from Data | Modernizing one's EDE means contending with numerous errors across the data pool. And even after a data lake has been constructed, it can deteriorate and become a data swamp unless

organizations make and stick to plans for regularly cleaning their data.

UNCOMN has learned that if data have errors—or there are duplicates in the database—it can be difficult for key decision makers or stakeholders to trust the information.

Quality Control | For our USTRANSCOM projects, we supported planning, implementation, and control of data quality metrics to monitor and control data flows from concept to operations.

Our approach provided proactive monitoring for data quality issues, including multiple software tools to resolve issues with non-compliant data. An example of how we impact sustainment quality operations is building scripts to monitor the ingest of Electronic Data Interchange (EDI) transitions that check the elements against the EDI standard. As EDI comes into the environment, the scripts log issues so that trends can be discovered, and a recommendation of a root-cause fix created.

Data quality was critical to transforming USTRANSCOM's application-focused data to enterprise-focused data assets supporting decision support. As new requirements for data interfaces or integrations entered our process, UNCOMN designed monitoring, evaluation, testing models, and use cases to support the consistent, certifiable, reconcilable activities used to validate capability.

We include a focus on timeliness, accuracy, completeness, validity, consistency, and uniqueness when building criteria for development and validation.

Data Lakes and Data Maturation | Furthermore, we learned that data lakes without a well-defined data maturation process can be problematic. Therefore, the data maturation process must be designed with built-in data quality metrics from the time raw data is published and through consumption.

We collaborated with previous contractors, leveraging the data maturation process to evaluate the data, compare to like data sources, and generate fact-based documents on the quality of the data compared to all sources across USTRANSCOM. We provided the findings to decision making bodies. We also provided team support to current program of record (POR) functional holders to ensure current interfaces and data exchanges were still operating and to support testing and verification of any new data exchanges.

In short, prioritizing data quality control avoids issues and makes the information maximally useful.

Conclusion

Conclusion | The establishment of a data lake within a larger enterprise data environment is an incredibly salient technology in making data manageable for business decisions.

But UNCOMN has also learned that having a data lake alone will not solve all of an organization's challenges when using big data for their business analytics.

Specifically, a data lake must be well defined, must be cleaned regularly (quality control), must geared for specific personas (data roles), must utilize tiered metadata (which can be automated), and must incorporate solid data governance with buy-in across the whole team.