DATA DISCOVERY and DISTILLATION SOLUTION

Summary and Informational Whitepaper

August 2019

Executive Overview

Throughout the federal government, agencies and organizations have accumulated hundreds to thousands of petabytes of unstructured data over the past several years. This large amount of data is currently creating a massive issue of disorganization and lost information. Regardless of the type, location, or classification of the data, without deep and accurate production of data discovery and subsequent distilling of information into "metadata" that describes the content, it becomes largely useless. Tagging the data would allow:

- More accurate and complete searches
- Digitization and cataloging of existing records
- Identify the type(s) of content included in the file stores
- Defensible deletion/expungement of data per disposition rules
- Provide a foundation for a "Data as a Service" platform
- More efficient and cost effective use of space and storage (both physical and digital)
- The easy adherence to new and upcoming compliance standards

For the most part, the content of these stores of data are geographically and logically dispersed with no associated metadata that describe the content of the file. There exists a requirement not to move the actual data, necessitating an architecture that builds the metadata and has pointers to the logical location of tagged data. Additionally, given the size of the collection(s), a largely unsupervised machine learning capable tagging tool, housed on a high-performance platform is recommended to allow as rapid processing and tagging as possible.

Over the course of many interviews with personnel from several federal agencies, requirements for tagging additional data types such as images and full motion video were identified. To address the wide variety of missions, and in order to "future proof" the platform, there is a requirement to design a flexible, modular, and easily deployed platform that can scale from processing hundreds of terabytes to hundreds of petabytes.

The purpose of this document is to outline the Data Discovery & Distillation platform, as well as various solution configuration options and highlight it's flexibility as both a purchase option and as an "as-a-service" solution that can suit the individual needs. There are several other optional solutions that can be purchased either a la cart or added in as part of the solution, though the pricing for those options is mostly dynamic and priced out at an enterprise-level based on need or size, amount of data being processed, and the depth of the project itself. This allows considerations to be made on a per project basis, where flat rate pricing would not account for the mission-specific needs of individual applications. Furthermore, the platform is infinitely customizable and can be configured to meet all goals, regardless of storage needs or services required.

To extend the flexibility to the budget side of the equation, it will be available as a "Data Tagging as a Service", with a monthly license as well as a perpetual license. Additionally, it can be delivered in a shielded, self-contained shipping container as well as a data center ready solution.

Regardless of configuration or mission goal, the data discovery & distillation platform can meet the immediate needs of any agency, and help with all of your future records management, digitization, and e-discovery requirements.

Market Analysis

The future of the government technology and data processing lies in the development and adoption of artificial intelligence (AI) and machine learning (ML) technologies, such as the classification and indexing outlined in this report. The large amounts of unstructured data and the value being lost is becoming wildly apparent to all government organizations, and now they are building out a strategy to harness the power of this information and better understand the data that they are sitting on. The process to making sense of this data is a long one, and the fundamental obstacles need to be understood before the solution is reached. But the good news is that the government is currently a leader in this front and will be adopting many new technologies, such as this data tagging solution, in order to harness the power of your information.

Currently, data quality is among the biggest challenges faced by AI projects, and what many don't realize is that this data requires more than reformatting to be useable, it needs to be labeled and indexed to be able to provide an explanation for later ML-based decisions. According to a senior IBM executive, data-related challenges are a top reason why clients have halted or cancelled AI projects. Citing that typically, 80% of the work with an AI project is collecting and preparing data and many companies are unprepared for the associated cost and work required. Our data tagging, digitization, and unsupervised indexing solution provides the perfect stepping stone to make sense of these vast stores of data and provide a strong foundation for later AI development.

The International Data Corporation (IDC) estimates that roughly 80% of all data is unstructured. This represents a massive waste of resources, both in the value of the data itself, as well as the money and space used to store the unstructured data in the first place. As such, the IDC notes that organizations could harness an additional $430 billion in productivity gains by as soon as next year if AI technology is embraced and valuable insights are gained from this unstructured data.

"Government agencies are awash in unstructured and difficult to interpret data.
– Bill Eggers, Exec. Dir. Deloitte's Center for Government Insights

The good news is that the Department of Defense in particular has long been a pioneer of language processing and machine learning, so the government is in a wonderful position to be able to benefit from this technology and take advantage of analyzing, tagging, and indexing their massive and growing stores of unstructured data to continue forging the path for artificial intelligence adoption beyond the government and throughout the technology sector.

Expanding on the DoD's involvement in AI, according to government financial reports for the FY2020 fiscal year, the Department of Defense's cyber spending jumped to a reported $9.6 billion under the current budget proposal, which marks an increase of $1 billion over FY2019. Of that, a staggering $927 million – nearly 10% of the entire proposed cyber budget – will be allocated to artificial intelligence/machine learning technology. This includes $208 million which will be allocated to building out the Pentagon's newly established artificial intelligence program. This shows a staggering amount of support for AI/ML technology and despite the fact that these figures are based on a proposed budget and are subject to modification, it illustrates a significant commitment to developing technologies in this space to help bolster technology modernization throughout the defense sector. And compared to last

year, this is a marked shift towards further adoption of AI/ML technology, giving us a glimpse into the future of technology priorities in the government space.

Solution Overview

The foundation of the modular platform is built around a highly capable selection of artificial intelligence and deep machine learning solutions to provide quick and accurate classification, indexing and cleansing of unstructured records. The selection of tools provides the ability to use different tools to allow for solution optimization depending on the type of data being processed and specific mission requirements, with a great range of flexibility for future adaptation. Regardless of the final design, there is emphasis placed on accelerating the ingest and unsupervised tagging/sorting stage to address the needs presented by large data set sizes seen throughout the government.

The primary goal of this solution is to provide client agencies the optimal platform for quick and accurate data tagging and digitization, whether it is several dozen petabytes of documents, emails, and other unstructured data; tagging image or FMV surveillance data; or even tag cyber security logs and data that are in various sources such as cloud storage and file shares. This will be completed by utilizing proven deep machine learning tools to rapidly identify content, intent, sentiment, and sensitive information producing an index of metadata that allows for federated searching across the entire data store, without the need to move files from their original location. The unsupervised tagging and digitization tools are in turn supported by an in-memory data grid and hardware choices from high performance compute/high performance storage platforms with high integrity. The entire solution is in turn protected by a selection of state-of-the-art cyber security tools.

The combination of components is dependent on mission, space, and packaging requirements. All solutions are available from a HUBZone certified contractor with deep experience with federal government contracts, systems integration expertise, and Authority to Operate (ATO) certifications. In addition, the contractor also specializes in delivering portable solutions in self-contained shipping containers, adding to the flexibility of the product package for those that need additional facility space to house the hardware.

This platform is conveniently offered either as a “Data Tagging as a Service”, with easy monthly or quarterly payments based on an agency’s preference removing any capital expenditures or as an outright purchase of the equipment and product licenses without only recurring maintenance costs (excluding expansion or upgrades). Additional services include conversion for analog files (paper and micrographic media), and access controls to set access permissions and credentialing are also available in conjunction with the other services as a part of this solution.

This solution provides several benefits to users that extend well past compliance with federal directives and enabling managing as a record. The solution will help with digitalization of files and documents in old and out dated formats, even damaged or difficult to read documents in preparation for compliance requirements. A well-tagged and sorted data repository is also enabled for inclusion in AI and analytics initiatives in support of the agency mission.

The overall utility and flexibility of this solution is a significant benefit, it can help agencies prepare for digitization as the government continues to modernize, or help better prepare, cleanse, and structure data to aid in the integration of artificial intelligence solutions. Large stores of data are extremely

valuable, but only when properly sorted and easily accessible via proper tagging and storage methods, both of which are provided by this solution package.

Workflow Procedure and Visualization

The versatility and power of the complete data-tagging and digitization solution can be attributed to the many components that make up the moving parts of this tool. In order to improve the flexibility of the product and provide a truly unique solution that can be configured to meet the needs of any user, these components can be nearly endlessly customized, replaced, or removed. However, the basic functions of the solution remain the same regardless of the configuration.

In order to better understand the order of operations with which this data tagging and digitization solution operates under, the visual simulation pictured below was created to show the path the data takes from the ingest stages, all the way to the end user, along with the multiple stops it makes along the way, which are all designed to provide the fastest and most accurate experience possible for all users submitting search queries for the prepared data.

Figure 1 below illustrates the workflow path, from initial ingestion and indexing, through final document retrieval by end users.

Ingestion and Indexing

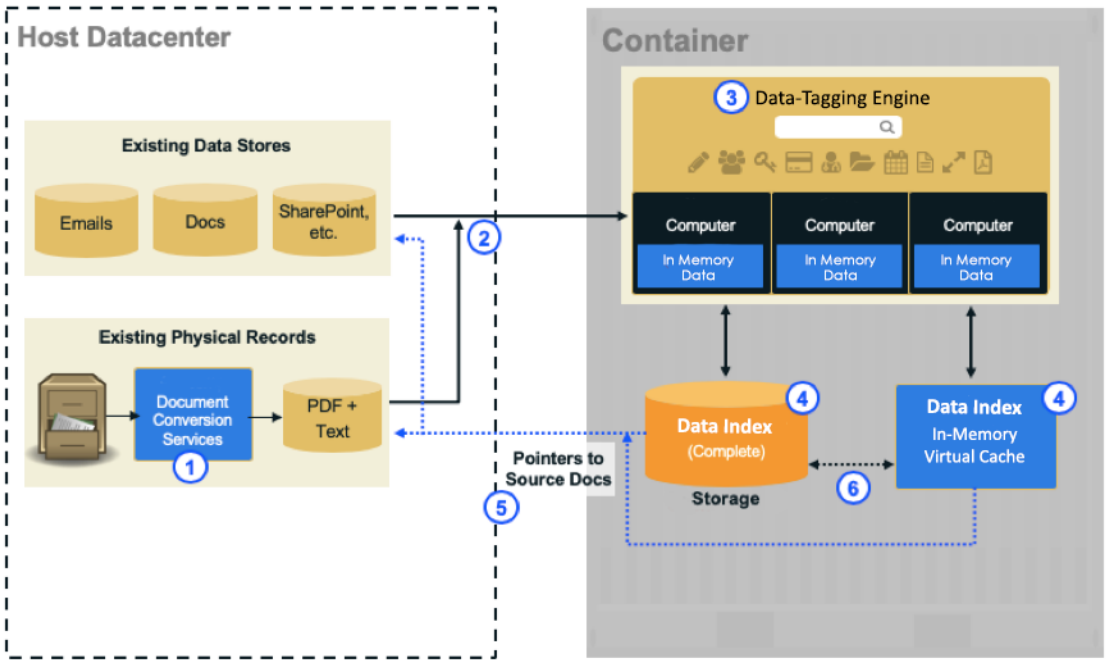

Figure 1: The Data Discovery workflow

The ingestion process begins in the host data center, with existing data stores and physical records already hosted by the client in an unstructured format. These potentially very large stores of data are

extremely difficult to navigate and lack the appropriate metadata tags or search capabilities to make finding specific pieces of information nearly impossible.

This unstructured data is then ingested into the Data Tagging Engine, which – with the help of computing platforms running high performance In-Memory Data solutions – unsupervised - sorts, tags, and indexes the information being processed. This data can then either be transferred to a Data Index on a storage platform for long term storage, and retrieval at a later date, or into a Data Index that remains in the In-Memory Data for immediate retrieval and use.

The process for physical documents is largely the same, however it undergoes a digitization process before the tagging and indexing can begin. This state of the art solution can read and digitalize even the most compromised and difficult to process physical documents and transform them into an easy-to-read, useable digital format that will not only aid with long term storage, but also make later retrieval much simpler.

Record Retrieval

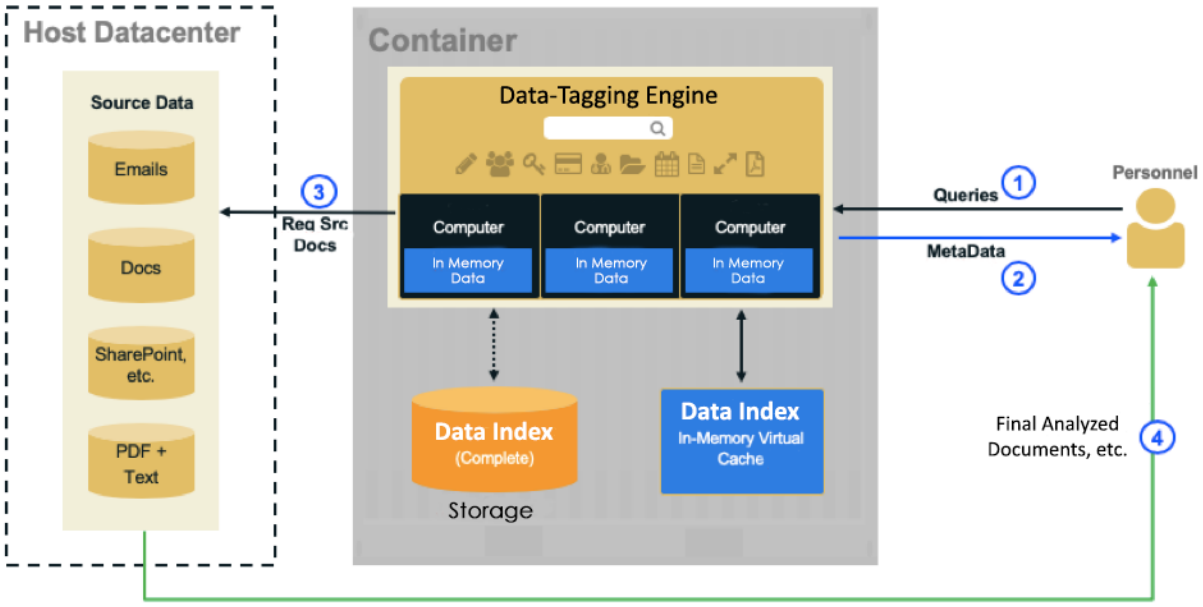

Figure 2: Record retrieval

Once ingested, tagged, and stored, the data is readily available at the fingertips of all records officers or other personnel that would need to interact with the metacard index of the stored data.

As previously mentioned, the tagged data itself will either live in a storage platform of the client's choosing – either a high performance solution, or a cost effective, yet reliable alternative – or for the data that is needed for immediate retrieval, information will be stored in the In-Memory Data Virtual Cache. While this space is limited and valuable, it will afford personnel with near-instantaneous recall of any data requested in a search query, so it is a sensible solution for frequently used or required pieces of data.

The sorting and retrieval process applies to all types of data, from traditional and simple files, like emails or text documents, to pdfs, videos, or digitally conversions of physical-format documents. This means the Data Tagging Platform is infinitely flexible, and can be utilized to sort, store, and recall data of all types, and can be used for a number of different applications.

Solution Components

The purpose of this solution is to provide a flexible, reliable platform that can be used to identify, tag, and store large amounts of unstructured data quickly and efficiently using AI and deep ML technology, for later recall by personnel, easing the time burden imposed by attempting to sort through large amounts of unstructured data. This platform also streamlines the storage process, providing significant cost savings by allowing data of all file types – including converted physical documents – to be stored in an easily navigated format.

One of the key benefits of this solution is the makeup of the final product. This solution is very much a compilation of its parts, and as such, takes a very modular form. It can be configured nearly endlessly to meet the needs of every application; if a client needs faster storage, it can be configured as such, if a client needs greater access control to meet security protocols, that can be arranged, as well. No matter what the mission needs are, this data tagging solution can be suited to those needs.

The following is a brief summary of our recommended organizations that could supply a product to help construct all the pieces that make up the full functionality of the solution. This includes many of the key components, such as storage, compute, tagging, scanning, and a few other major parts that make up the greater whole that is the entire comprehensive solution. This is not an exhaustive list, due to the flexibility of the solution, but it is intended to cover the core functionality. For more application-specific builds, please contact us with desired configuration options.

Data Tagging

The data-tagging and security component that makes up such a major part of this solution provides the functionality that enables organizations to better organize their information while removing risk and making their data more productive and secure. The data tagging component of this solution is optimized to work with a multitude of different file formats, from standard text files such as emails or word processing documents, to pdfs, and even videos. All of these formats can be ingested, analyzed, tagged, and sorted into a metadata index, allowing for quick search and recall of any type of file format to ensure that no data remains unstructured and unusable.

This solution searches and discovers all your organization’s unstructured information such as the aforementioned emails, documents, and other files, then it understands what is important using proven deep machine learning and artificial intelligence technologies that can extrapolate the key points of any piece of data. After this, the unsupervised data tagging solution automatically classifies the data, tells you where it is, and sets predetermined access permissions based off the content of the data.

All types of data are then tagged into a metacard index, allowing it to be easily searchable and well organized. Furthermore, this tool has the ability to index and search for files in any format or type,

reducing any restrictions or complications due to file type. Once tagged, an extensive audit system is maintained, documenting everything that is created, moved, or deleted, regardless of where it's stored or its file format.

Storage

When working with such large amounts of data, the storage component of the solution is crucially important. However, each application will have a different need for storage. Many programs will want near-instantaneous recall of their newly sorted and organized data and will want to tier one flash storage solution to not only reliably house their data but provide quick recall for increased usability and convenience.

On the other hand, some programs will call for a mostly stationary storage solution; one that will house the data for long periods of time and won't necessitate a very quick recall or constant search queries. And thus, for this application a reliable yet cost-effective storage solution can be employed in order to make the pricing model more attractive for users that are willing to compromise slightly on speed while retaining reliability and longevity.

Provisions have been made for both of these necessities as well as any combination thereof in between, allowing for both solutions to be combined in an endless array of combinations to meet both the performance and cost needs of every client.

Option 1: Performance-Oriented Storage

The high speed, performance oriented storage solution, aimed at those that need rapid, regular access to indexed data leverages the speed and efficiency of all-flash storage to achieve your agency's mission while still meeting budgetary requirements. The tier one, all-flash storage solution offers 10x the performance of standard disk storage, along with greater resiliency, and simplicity, all while costing less than a standard disk in many cases. This solution capitalizes on available time and space by offering industry leading speed, along with the smaller datacenter footprint afforded by the more agile and compact flash storage.

This high-performance storage component can help organizations increase performance and reliability of their stored data, while helping them save a significant amount of money in time-related expenditures. Additionally, this virtualized infrastructure solution is completely scalable, and promises no planned downtime for upgrades or maintenance. Government agencies can rely on the all-flash performance, security, resiliency, and simplicity to help accomplish their mission at speed.

Option 2: Cost Effective Storage

For data that is mostly at rest and kept for archival purposes or are stored for later retrieval where speed and performance are not the top priority, we offer a more traditional storage solution, much like what many datacenters are employing today. This budget-friendly, reliable solution is still mindful of performance, but utilizes traditional spinning disk hard drive storage for large tagging and storage projects that do not require top tier speed for retrieval of tagged and indexed data.

This solution uses a highly reliable, cost-effective storage designed to serve specific use cases and demands. While the solution may use a more traditional format and comes in a more cost-effective package, it does not sacrifice any necessary functionality or any key traits such as reliability. Additionally, the storage solution utilizes features such as sync and share for all users and real time file sync across locations with data archiving with unmatched security and standards compliance.

This budget-friendly (yet still fast and reliable) solution is a purpose built product that is tailored to the needs of the client. That means product design starts by understanding the diverse workloads in customers' environments. And because workloads vary, the cost-effective storage solutions are offered with a broad line of specific products where each variation in the line has a clear intent and exceptional value.

Option 3: Hybrid Storage

Realistically, very few applications are going to fall so strongly on one side of the spectrum or the other as far as storage performance demands go. For every other mission that requires the "just-right" combination of performance and budget, a hybrid solution is offered. This balances out the need for high recall times for data that is used frequently, or for data that is undergoing the ingest and tagging process, and a more budget-conscious, traditional storage solution for the data-at-rest that is not needed as frequently or as quickly.

The specifics of the balance between high-performance storage and cost-effective storage solutions will be determined during the initial evaluation, and project level pricing and storage solution scaling will be recommended at that time.

Almost every real-life application of this data tagging solution will end up employing a hybrid storage solution at utilizes both of the above components to their full potential to create a truly unique and personally tailored product that can successfully meet whatever specific needs the mission calls for.

In-Memory Data

A leading open source in-memory data grid with tens of thousands of installed clusters and over 23 million server starts per month will be used to expedite the tagging process. The operational in-memory computing platform helps leading organizations all over the world, across several diverse industries manage their data and distribute processing using in-memory storage and parallel execution for breakthrough application speed and scale.

Data is distributed evenly among the nodes of a computer cluster, allowing for horizontal scaling of processing and available storage. Backups are also distributed to protect against the failure of any single node. The in memory data platform provides central, predictable scaling of applications through in-memory access to frequently used data and across an elastically scalable data grid.

This component provides unrivaled speed during the ingest and tagging process, utilizing the available space on existing memory stores and leveraging a data storage platform with unmatched speed providing the fastest and most efficient methods for processing and tagging very large stores of unstructured data.

Digitization

An important component that makes this solution a truly comprehensive tool is the digitization tool that can analyze, digitize, and categorize all formats of records, including old and outdated records, even records in poor condition. Many government organizations still rely on physical records, despite mandates that push federal organizations towards modernization.

Many of these physical records are aging quickly and deteriorating to the point where they will soon be unusable, and even in their current form are difficult to decipher. When that is combined with new and pending regulations that call for digitization of records, the window in which organizations can make sure their records are up to date and viable is closing quickly. Not only that, but digitization of records also greatly reduces the footprint required to house the information.

The digitization tool specializes in providing enterprise information and document services where quality, accuracy, and security of the information are vital to the success of the program.

Additional Services

Integration

Integration services are handled via a HUBZone certified small business with extensive experience in providing containerized solutions to various government organizations. This company will perform all integration work and will secure any ATOs, further simplifying the process.

The platform is available either as a data center ready solution or configured in an ISO shipping container which will be prefabbed and outfitted with power supply, air conditioning, and the selected hardware/software components.

Additionally, program management will be provided via the integration provider, ensuring sustainment and training for the lifecycle of the technical equipment, in order to ensure safe and proper operation by all personnel. This support will guarantee stability and growth of the storage solution, and the support of the solution will expand with the needs and progress of the project throughout the entire lifecycle.

Data Store Housing

The data needs of each client will vary based on the data stores available to them. While this platform is incredibly efficient with both speed, pricing, and footprint, there still needs to be an allowance for space needed to house the compute components of this solution. Provisions are made for a number of needs, a selection of which are listed below:

On Prem – Existing Data Center – The entire solution can be housed within a small portion of footprint within an organization's existing data center. This would reduce costs and is an ideal configuration for

those that have the space available. However, it is not a requirement, there are many other potential options for storage.

On Prem – Externally Housed Shipping Container – If there is not enough storage space available on premises in an existing data center, the solution integrator is equipped to provide a shipping container purpose built to house the data tagging solution. It will come complete with power, air-conditioned suited for a data center, and the compute necessary to run the tool.

Cloud Hosting – If neither of these solutions suit the needs of an organization, the tool can be placed into a cloud environment. This can be any variation of public, private, or hybrid cloud.

Access Control

Adaptive identity access management and cross platform information sharing for identities you don't or can't manage. Resilient helps organizations increase productivity by enabling real time access to sensitive data while maintaining privacy and security. This solution helps simplify IT by extending the power of existing IAM tools to manage access for an entire ecosystem of users.

Provides network-based, distributed architecture that is both highly flexible and scalable, allowing customers to safeguard their data and applications while enabling secure collaboration and information sharing. They make it easy to leverage existing identity and security products, or add external identity sources and services with predefined, custom policies.

Search and Analytics

The analytics tool provides instant insight engine is powered by thousands of advanced GPU cores that bring unparalleled speed, streaming data analysis, visual foresight, streamlined machine learning, and best-in-class partner innovation ecosystem to break through the old bottlenecks.

This tool was built from the ground up to address a critical need of national security, enabling real-time decisions in areas including cyber network traffic monitoring, smart city event tracking, and dynamic workforce management. The platform brings together all key elements of dynamic analytics in a unified platform: historical data analytics, streaming data analytics, location intelligence, and artificial intelligence. It includes a distributed, in-memory, GPU-accelerated database that utilizes a powerful combination of CPUs and NVIDIA GPUs to analyze massive, complex datasets with millisecond response times. A powerful tool that is perfectly matched to analyze and search through massive stores of data, capitalizing on the newly optimized and indexed datastores.

Pricing Information

To add to the flexibility of the data tagging platform, there are the various pricing options available for buyers to help best meet their needs for budgeting and expansion purposes. It is offered either as a complete purchase option or alternatively, it can be purchases as an innovative Data-Tagging-as-a-Service option, allowing the solution to be leased over time, having a lower impact on an organization's initial budget and extending the purchase price over a multi-year term, with a convenient maintenance and upgrade fee to eliminate any surprises that could arise in the pricing model and make the solution conveniently fit into every budget.

Due to the uniqueness of this solution, and how each iteration of the platform will differ from organization to organization in order to fit specific needs, the pricing will select the options, configuration, and scale of each order on an enterprise-level, rather than pricing by the petabyte which introduces several variables that could create difficulties for the customer as they attempt to expand and need every component to expand to meet those needs. Utilizing enterprise pricing, we are better able to cater to individual needs, and make pricing as efficient and accurate as possible one a case by case basis.

Due to this pricing model, this is not an exhaustive look at price scaling, but it will give a glimpse into the methodology that goes into the pricing process. For a tailored price quote based on mission objectives, please contact us to discuss your organization's specific needs.

Conclusion

The components of the solution outlined above are designed to work together to supply clients with a complete and comprehensive data tagging, digitization, and analytics solution to address a limitless amount of unstructured data in a scalable format that will be easy to maintain, and easy to implement as a long-term solution for incoming data.

The solution design is flexible to meet an endless variety of mission requirements using complimentary products that work together to meet and exceed all of the needs that accompany the sorting, tagging, logging, and storage of such large amounts of data. Additionally, the product is designed to meet these goals by the most cost-effective means possible, while not sacrificing the necessary performance needed to complete the mission objective.

With the option of "Data Tagging as a Service", on-going maintenance and project management along with deep-learning automation help ensure that the implementation, upgrades, and continued management of the solution are painless for the Air Force and are managed by the suppliers outlined above. Additionally, the housing of the compute and storage is also completely modular, it can either be palletized and housed in a vendor-supplied storage container or placed within the footprint of an existing data network.

Regardless of application or scope, this solution is designed to meet the mission requirements of a number of different platforms and providing a means to turn large amounts of unstructured, unknown, unavailable, and unusable data into a wealth of unparalleled knowledge and an invaluable agency tool.