DATA DISCOVERY and DISTILLATION SOLUTION

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 "metadata" describing the data, it becomes largely useless. Tagging the data would allow:

- More accurate and complete searches
- Digitization & cataloging of records
- Identify content included in the files
- More efficient use of space & storage
- The easy adherence to new & 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.

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

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 cost and work associated with that. Our data tagging, digitization, and 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 Data Discovery Workflow

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 room for flexibility for future adaptation. Regardless of the final design, there is emphasis placed on accelerating the ingest and 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 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 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, all supported by state-of-the-art security tools.

For more information, contact us at info@gssfedsales.com or call us at 240-482-4720.