Data wrangling for effective migrations and acquisitions in oil and gas

Industries have changed vastly with mergers and acquisitions becoming the norm, ranging from asset/facility acquisitions to company mergers. In today's environment, organizations face the challenge of interrogating systems and data to determine what should and should not be included within the migration.

Identification and migration of large data lakes may seem daunting, especially when the information is unstructured and has been subject to decades of modifications or has been dormant.

Migration of copious quantities of files and data can have organizational impact when performing activities such as:

- Increased man hours spent locating documentation and data
- Interrogation and identification of acquired data
- Technology constraints: Inability to identify and retrieve metadata with ease
- Identification of true source information when files are duplicated/copied within systems and drives

Organizations must collaborate with IT partners that can offer global best practices, and leverage domain expertise paired with innovative automated AI solutions to release the value of information while also retaining and protecting critical data from unwarranted use.

IT partners can interrogate structured and unstructured data (including hard drives, USBs, etc.) to enable the identification of data as 'must haves vs. nice to haves vs. don't needs' during either migration or acquisition of data from and to companies and systems. Data wrangling services are ideally situated to assist with the identification, deduplication, transformation and migration of files and data during acquisitions, mergers and divestments.

Data wrangling should be tailored to each organization's requirement to ensure that the integrity and sensitivity of information is uncompromised, while releasing and transferring information en masse in a timely manner. Vast migrations should be managed by performing the 5 major activities as described below, as the basis for identification and management of migrations, which are then tailored and configured to organizational requirements.

Step 1: Discovery

Analyses of information management numbering and coding procedures and specifications is integral to understand and identify files and metadata within structured and unstructured environments for migration.

The purpose of the discovery identification process is to ensure that highly sensitive information is quarantined and not shared with other parties. Organizations need to identify critical information for migration from disparate locations while restricting unwarranted access and retaining the integrity of all information and related data.

Organizations should also consider and manage the following during the discovery stage:

1. Identify disparate information sources
2. File deduplication ensuring that one true source file is identified and the latest revision is maintained and migrated
3. Identification of redundant, obsolete and trivial information to ensure ROT is quarantined for archive and then removed
4. Compliance with personal and sensitive information requirements to detect and quarantine information that must not be shared, such as commercials, finance, passport, DOB, etc.
5. Identification of intellectual property information as detection of highly sensitive organizational IP is critical to safeguard the company against misuse of information

By utilising ML and automation enabling technologies, the risks associated with manual user intervention are mitigated while also reducing the TCO associated with divestment and procurement

Step 2: Metadata extraction

Extracting document metadata can be challenging, especially if there are hundreds of thousands of unstructured files. Traditional systems are typically set up to extract individual documents one at a time, which isn't viable in today's environment.

Files need to become fully searchable by utilizing OCR techniques, which enable metadata extraction. Extracted metadata should be utilised to create a taxonomy to allow the classification and identification of information.

During migrations of vast files and data, organizations typically need to identify and perform the following:

- Extract metadata from documents and images (document numbers, revisions, authors, well names, well identifiers, dates, etc.)
- Extract complex metadata using client defined taxonomies
- Extract technical data from documents (tags, line numbers, bore hole data, etc.)
- Automate depth registration of scanned log images
- Identify sensitive and legal files so they can be located to avoid data leakage
- Verify content to ensure information is correct and as expected
- Utilize multi-level de-duplication. Our data wrangling service utilizes multiple techniques to enable identification of duplicate files. Our wrangling process identifies exact duplicates and near match candidates by utilizing fuzzy matching techniques to isolate files for further inspection
- ROT (Redundant, Obsolete and Trivial) identification by configuring using client defined rules to enable identification and quarantine of ROT information

Step 3: Metadata and file transformation

Different document types have different attribution requirements and some content may need to be converted or repurposed from one format to another before migration, such as:

- Foreign language text converted into an English word file (note images/pictures are not converted and transferred to word)
- Identification and removal of illegal characters (such as *%$£"!@&)
- Format standardisation (dates, text cases, formats)
- Title description acronyms removed and replaced by whole words

- Transformation to new taxonomies/numbering schema of historical taxonomy numbering

Content needs to be accurately identified, extracted, classified, attributed and exported into an acceptable format to allow for target system population, or to align with the agreed organizational recipient system/migration requirements.

Step 4: Quality assurance and tracking

Common databases should be utilized for metric tracking and live dashboard reporting. An end to end audit trail should also be maintained throughout the conversion process, which can help achieve successful migration while monitoring the integrity and sensitivity of information throughout the entire process. During quality assurance, organizations need to ensure that the following are correctly and accurately identified and quarantined:

- Personal and sensitive information such as commercials, finance, passport, DOB, etc., to ensure compliance throughout

- Intellectual property information (highly sensitive)

Step 5: Transfer

All files and data should be retained within a cloud environment to ensure secure methods of migration and transfer of files and data. Once all extracted and transferrable content has been sufficiently classified and attributed, the content data should be transformed into export indexes such as system load sheets, client requirements, etc. The files themselves should be transferred by utilizing a secure cloud environment and files should be categorized in a structured directory such as folder structures.

Post migration

Post successful migration, acceptance of files by the receiver should be fully documented by way of transfer of liability/ownership. The originating organization should also retain a full built-in backup, which includes all files directories and automated metadata outputs, a migration dashboard that tracks migration progress, duplications, ROT and highly sensitive data.

Business benefits:

Data identification: Data wrangling techniques enable the identification and transformation of large volumes of files and data, which allows for accurate system attribute population

Improved information access: The identification and metadata extraction of true source information ensures end users access and utilize the most current information, thus reducing the probability of incidents and near miss accidents

Information integrity retention: Software and processes need to be tailored to safeguard the integrity of the files and metadata, thus ensuring the accuracy, consistency and reliability of files and data

Information security: Utilization of secure and access restricted file exchanges and working environments is imperative to ensure the authorized use of information

Data protection: Utilize the cloud environment to safeguard information from corruption, compromise or loss.

About the author

Janine Murray
Principal Consultant,
Wipro Limited.

Janine Murray is an IM Consultant with over 15 years of experience in the O&G industry. She has extensive FE/Operations and Major Capital Project (MCP) Information Management experience. She has deep experience with IM brownfield modifications, greenfield enhancements, MCP joint ventures, closeout, and MCP handover to Operations. Additionally, she is experienced with document cleansing and data extraction techniques for digitizing O&G legacy assets. She can be reached at: **janine.murray@wipro.com**

Wipro Limited
Doddakannelli, Sarjapur Road,
Bangalore-560 035,
India

Tel: +91 (80) 2844 0011
Fax: +91 (80) 2844 0256
wipro.com

Wipro Limited (NYSE: WIT, BSE: 507685, NSE: WIPRO) is a leading global information technology, consulting and business process services company. We harness the power of cognitive computing, hyper-automation, robotics, cloud, analytics and emerging technologies to help our clients adapt to the digital world and make them successful. A company recognized globally for its comprehensive portfolio of services, strong commitment to sustainability and good corporate citizenship, we have over 160,000 dedicated employees serving clients across six continents. Together, we discover ideas and connect the dots to build a better and a bold new future.

For more information, please write to us at **info@wipro.com**