ERDC/ITL TR-21-2

US Army Corps
of Engineers®
Engineer Research and
Development Center

ERDC
ENGINEER RESEARCH & DEVELOPMENT CENTER

Engineered Resilient Systems

Data Lake Ecosystem Workflow

R. Cody Salter, Quyen T. Dong, Cody A. Coleman, Maria A. Seale,
Alicia I. Ruvinsky, LaKenya K. Walker, and W. Glenn Bond

April 2021

Information Technology Laboratory

The U.S. Army Engineer Research and Development Center (ERDC) solves the nation's toughest engineering and environmental challenges. ERDC develops innovative solutions in civil and military engineering, geospatial sciences, water resources, and environmental sciences for the Army, the Department of Defense, civilian agencies, and our nation's public good. Find out more at www.erdc.usace.army.mil.

To search for other technical reports published by ERDC, visit the ERDC online library at https://erdclibrary.on.worldcat.org/discovery.

ERDC/ITL TR-21-2
April 2021

Data Lake Ecosystem Workflow

R. Cody Salter, Quyen T. Dong, Cody A. Coleman, Maria A. Seale, Alicia I. Ruvinsky, LaKenya K. Walker, and W. Glenn Bond

U.S. Army Engineer Research and Development Center (ERDC)
Information Technology Laboratory (ITL)
3909 Halls Ferry Road
Vicksburg, MS 39180-6199

Final Technical Report (TR)

Prepared for Headquarters, U.S. Army Corps of Engineers
Washington, DC 20314-1000

Under ERDC FADs 0642033036, “FY20 ERDC ERS Program Support” and 0603833D8Z, “Engineering Science and Technology”

Abstract

The Engineer Research and Development Center, Information Technology Laboratory's (ERDC-ITL's) Big Data Analytics team specializes in the analysis of large-scale datasets with capabilities across four research areas that require vast amounts of data to inform and drive analysis: large-scale data governance, deep learning and machine learning, natural language processing, and automated data labeling. Unfortunately, data transfer between government organizations is a complex and time-consuming process requiring coordination of multiple parties across multiple offices and organizations. Past successes in large-scale data analytics have placed a significant demand on ERDC-ITL researchers, highlighting that few individuals fully understand how to successfully transfer data between government organizations; future project success therefore depends on a small group of individuals to efficiently execute a complicated process. The Big Data Analytics team set out to develop a standardized workflow for the transfer of large-scale datasets to ERDC-ITL, in part to educate peers and future collaborators on the process required to transfer datasets between government organizations. Researchers also aim to increase workflow efficiency while protecting data integrity. This report provides an overview of the created Data Lake Ecosystem Workflow by focusing on the six phases required to efficiently transfer large datasets to supercomputing resources located at ERDC-ITL.

Contents

Figures and Tables

Figures

Tables

Preface

This study was conducted by the U.S. Army Engineer Research and Development Center (ERDC) for the Army 6.3 Big Data Analytics program and the Office of the Secretary of Defense-funded Engineered Resilient Systems program under Funding Authorization Document (FAD) 0603734A, "Military Engineering Advanced Technology" and 0603833D8Z, "Engineering Science and Technology," respectively. The technical monitor for this work was Dr. Owen Eslinger.

The work was performed by the Institute for Systems Engineering Research, the Computational Analysis Branch, and the Scientific Software Branch, all of the Computational Science and Engineering Division, U.S. Army Engineer Research and Development Center, Information Technology Laboratory (ERDC-ITL). At the time of publication, Dr. Simon Goerger was Director of the Institute for Systems Engineering Research, Dr. Jeff Hensley was Chief of the Computational Analysis Branch, and Mr. Tim Dunaway was Chief of the Scientific Software Branch; Dr. Jerry Ballard was Chief of the Computational Science and Engineering Division; and Dr. Robert Wallace was the Technical Director for Engineered Resilient Systems. The Deputy Director of ERDC-ITL was Ms. Patti Duett and the Director was Dr. David Horner.

COL Teresa Schlosser was the Commander of ERDC, and Dr. David W. Pittman was the Director.

THIS PAGE INTENTIONALLY LEFT BLANK

1 Introduction

1.1 Background

In today's digital world, the role and growth of data is ever increasing. According to the International Data Corporation, the global datasphere will grow from approximately 33 zetabytes in 2018 to over 175 zetabytes by the year 2025 (Reinsel et al. 2018). This increase in data represents a 430% growth in data over a mere 7 years. As the world becomes more and more digital, businesses and government and military organizations are looking to data to gain insights that will better inform their decisions. For example, ERDC-ITL was recently given 23 terabytes of sensor and logbook data from the U.S. military's rotorcraft fleet. Through the consolidated storage and analysis of this data, computer scientists and engineers gained insights on rotorcraft performance resulting in a potential life increase of 700 flight-hours for 75% of a rotorcraft fleet (Seale et al. 2018). These findings could result in significant cost savings for the U.S. military, all through the analysis of big data.

As a result of this work, ERDC-ITL is receiving multiple requests to perform similar analyses on large datasets from across the Department of Defense. To increase efficiency and to ensure customer satisfaction, ITL's Big Data Analytics team set out to document and implement a big data transfer workflow. This workflow includes phases, processes, tasks, and responsible stakeholders. This documentation was created to support this effort by providing a detailed description of the Data Lake Ecosystem Workflow.

1.2 Objective

The objective of this documentation is to provide a detailed description of the Data Lake Ecosystem Workflow.

1.3 Approach

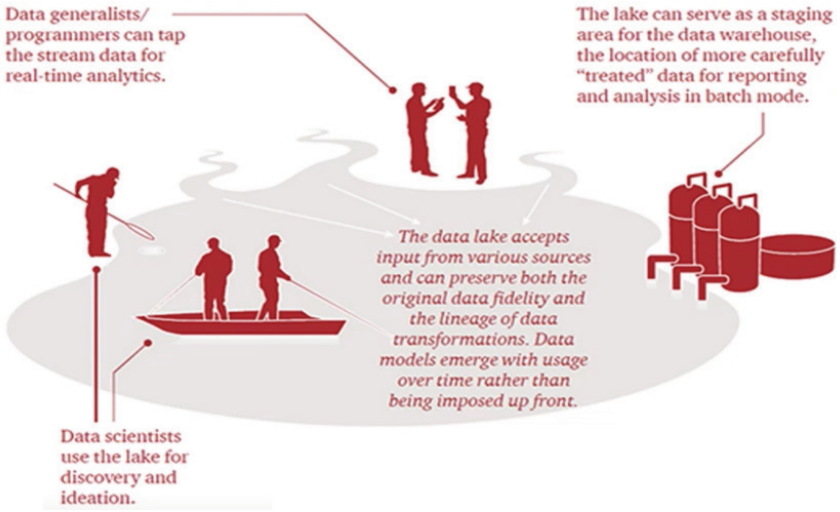
This documentation begins with a review of techniques used to elicit workflow information from stakeholders. This is followed by a description of the various stakeholders and workflow processes (Chapter 3). The work concludes with a description of each step in the workflow (Chapter 4) followed by outlines of the detailed process workflows (Appendix A).

2 Data Lake

The use of big data to solve problems often requires the use of complex analytics and transformations to gain useful insights. However, certain standards, policies, and procedures must be followed to enable the safe, reproducible, and accurate analysis of big datasets. One such data construct forms the basis of most big data research performed by ERDC-ITL's Big Data Analytics team; that data construct is the "data lake." According to Amazon Web Services, a data lake is "a centralized repository that allows you to store all your structured and unstructured data at any scale" (Amazon Web Services 2019). The IBM Corporation defines data lakes similarly, as "next-generation hybrid data management solutions … their highly scalable environment can support extremely large data volumes and accept data in its native format from a wide variety of data sources" (IBM Corp. 2018).

These definitions accurately capture most traits of a data lake but omit one critical quality. Once created, data lakes are immutable. This means that the raw data stored in a data lake cannot be modified. The original fidelity of the data is preserved, and all transformations and analytics are performed on copies of the original dataset. These copies, transformations, and analytics, along with the original dataset, are collectively known as the data lake ecosystem. Figure 1 provides a visual explanation of data lakes and how they are used (PwC 2015).

Figure 1. A graphic illustration of a data lake and its users.



Source: PwC (2015).

Certain scenarios exist that call for the modification of a data lake, but these scenarios are at the Data Owner's direction and discretion. One such scenario is a data update. Data updates occur when newly acquired data must be added to an existing repository for storage and analysis. Scenarios such as this may be accommodated if requested by the Data Owner.

ERDC-ITL adopted the data lake for multiple reasons. One reason is that data lakes allow for the storage of structured and unstructured data in a single repository. Many collaborators require storage solutions that can accommodate dissimilar data types. For example, rotorcraft collaborators are looking to gain insights through the combination of two to three dissimilar datasets into one large repository. Data lakes can meet this need by allowing for the creation of processed datasets called AI-Ready datasets for future study and research.

In data management, it is critical to implement storage solutions that preserve data fidelity. Data transformations and analyses can degrade the fidelity of original or master datasets. The gradual degradation of data through transformation and analysis could impact future research, resulting in misleading or erroneous insights. Data lakes can preserve data fidelity through forced data immutability. Enforcing this requirement not only preserves the fidelity of the data; it also ensures the integrity of all existing and future analytic workflows that use the data.

Creating a data lake, however, first calls for the transfer of large datasets to high performance computing (HPC) systems. Several tasks and agreements must first be completed before the transfer and ingestion of data. Because of the intricacies of the process, the U.S. Army Engineer Research and Development Center (ERDC) team set out to understand and document these tasks to ensure the success of future research projects involving big data transfers and ingestions. The primary method of soliciting this information was stakeholder interviews. The next chapter reviews the process of information solicitation.

3 Data Lake Ecosystem Workflow Process Elicitation

A number of ERDC-ITL employees have facilitated the transfer and ingestion of large datasets onto HPC assets located at the ERDC. The processes that enable these data transfers are typically group efforts involving contributions by representatives from ITL and across ERDC. To document this workflow, researchers interviewed and questioned seven individuals familiar with the data transfer and ingestion process. These individuals represented various offices and branches within the ITL and throughout ERDC. Interviews were held in-person or via teleconference and typically lasted approximately 1 hour in length.

3.1 Stakeholder interview

The interviews conducted during the workflow elicitation process used a semistructured interview style (Laplante 2013). Like most structured interviews, each interview was scheduled in advance with the interviewee, and a list of questions was produced to lead the conversation. These questions, however, were asked informally and in a conversational style, similar to an unstructured interview. Furthermore, the interviewee was allowed to direct the conversation through his or her responses to preselected questions. Combining structure with a conversational style resulted in productive, semistructured interviews that effectively elicited information from subject matter experts. Additionally, the interviewers typically paired two interviewers to one interviewee. The first interviewer moderated the interaction while the second interviewer carried the responsibility of notetaking to record all pertinent insights gained during the interview. The moderator was responsible for guiding the conversation using the prepared questions, but was also allowed to deviate from the script, as needed, to gather additional information. Allowing such deviation from the prepared questions provided the interviewers with opportunities to capture previously unknown insights.

Researchers held interviews in two rounds. During the first round, researchers collected high-level information about the workflow and gained general knowledge about the process. This general knowledge allowed researchers to construct an overview of the full Data Lake Ecosystem

Workflow. While constructing the workflow overview, researchers noticed that the workflow could be naturally divided into nine distinct processes:

1. Initial Customer Interaction
2. HPC Access
3. Data Transfer Agreements
4. Data Access Rules
5. Data Transfer
6. Data Ingest
7. Raw Data Storage and Update
8. Data Transformation
9. Data Analytics.

With these processes in mind, researchers used the second round of interviews to ask targeted questions about each workflow process. Researchers continued the activity of information elicitation until a complete and thorough workflow was attained.

The information gathered during interviews was then aggregated and transformed into a set of tasks. These tasks were arranged sequentially and assigned to responsible parties, or stakeholders. Researchers identified nine unique stakeholders for the Data Lake Ecosystem Workflow:

1. Data Owner
2. Collaborator
3. Collaborator Support
4. ERDC Director
5. ITL Director
6. ITL Support
7. HPC Support
8. Technology Transfer Officer
9. The Office of Counsel.

3.2 Activity diagram

Collectively, the process, task, and stakeholder information was consolidated into process-specific activity diagrams. Activity diagrams "represent behavior in terms of the order in which actions execute" (Friedenthal et al. 2012). One unique feature of activity diagrams is their ability to attribute

task responsibility to groups or individuals. Activity diagrams achieve this by displaying tasks in rows while displaying the responsible party in columns. By doing so, one can easily understand those individuals responsible for a given task. Multiple tasks may be displayed in a single row. In such cases, tasks may execute in parallel.

Table 1 lists the elements of an activity diagram that describes the interaction between a driver and his or her vehicle. The driver initiates the interaction by starting the vehicle. The Control Accelerator Position and Control Gear Select activities begin once the ignition is placed in the ON position. The driver then controls the accelerator position and gear selection simultaneously, at which point the vehicle provides power. These three activities occur in parallel, as indicated by their location in the same row. The driver is responsible for controlling accelerator position and gear selection, while the car is responsible for providing power. This information is represented through the tasks' column positions. These two activities end when the driver turns the vehicle off. All pertinent notes are collected in the Notes column of the activity diagram.

Table 1. An example of an activity diagram.

#	Driver		Vehicle	Notes
1	Start vehicle			
2			Ignition ON	
3	Control Accelerator Position	Control Gear Select	Provide Power	
4	Turn off vehicle			
5			Ignition OFF	

By interviewing numerous individuals familiar with the Data Lake Ecosystem Workflow, researchers discovered the necessary tasks and documented their findings using activity diagrams. Appendix A includes these diagrams.

Collectively, the Data Lake Ecosystem Workflow was broken down into nine distinct processes. Tasks within these processes were allocated to nine different stakeholders across the ERDC and ITL. The following chapter provides detailed overviews of the processes, tasks, and stakeholders that make up the Date Lake Ecosystem Workflow.

4 Data Lake Ecosystem Workflow

Before data can be transferred to supercomputing assets at ERDC, many preliminary tasks must first be accomplished. These tasks are distributed across nine processes, and multiple parties must collaborate to ensure their successful completion. At a higher level of abstraction, these nine processes and their associated tasks may be organized into six phases (Figure 2):

1. *Educate & Interact*. During the Educate & Interact phase, collaborators are educated on ITL's capabilities and past successes in the area of data analytics. Information is also collected on the dataset(s) to be transferred.
2. *Access & Agreements*. In the Access & Agreements phase, three activities occur simultaneously.
    a. First, our collaborators work to gain access to HPC assets located at ERDC.
    b. Next, ITL staff and collaborators author, review, and approve a Data Transfer Agreement.
    c. Finally, both parties create and approve Data Access Rules (DARs).
3. *Transfer & Ingest*. The Transfer & Ingest phase includes two major activities:
    a. The transfer of data to the HPC.
    b. The ingestion of data onto the HPC.
4. *Store & Update*. During the Store & Update phase, ITL provides safe and secure storage of collaborator data. ITL also helps facilitate data updates as needed.
5. *Transform & Analyze*. In the Transform & Analyze phase, ITL and our collaborators perform data transformations and analyses to gain meaningful insights from the data.
6. *End of Life*. In this last phase of the Data Lake Ecosystem Workflow, ITL helps facilitate the transfer of collaborator data to a new storage location. ITL also removes all collaborator data from HPC resources located at ERDC.

Figure 2. Data Lake Ecosystem Workflow phases.

To successfully step through the six phases outlined above, stakeholders from across the ITL, ERDC, and our partner organizations must collaborate. In total, at least nine different stakeholders and stakeholder groups must work together to safely transfer data between government organizations. The following section provides a detailed description of these stakeholders and stakeholder groups.

4.1 Stakeholders

Nine different stakeholders must contribute their time and expertise to complete the necessary tasks within the Data Lake Ecosystem Workflow: (1) Data Owner, (2) Collaborator, (3) Collaborator Support, (4) ERDC Director, (5) ITL Director, (6) ITL Support, (7) HPC Support, (8) Technology Transfer Officer, and (9) the Office of Counsel. Table 2 lists and describes the nine stakeholders.

Table 2. Stakeholders.

Stakeholder	Description
Data Owner	The individual and/or organization that owns the data to be transferred
Collaborator	An individual or team outside of ERDC-ITL assisting with the transfer, transformation, and analysis of data
Collaborator Support	Individuals working under the collaborator during the research process
ERDC Director	The Director of the U.S. Army Engineer Research and Development Center
ITL Director	The Director of the Information Technology Laboratory
ITL Support	All ITL employees supporting the data research effort
HPC Support	A team of individuals responsible for managing and maintaining U.S. Department of Defense (DoD) HPC assets
Technology Transfer Officer	An individual responsible for the facilitation of organizational agreement review and approval
Office of Counsel	A team of individuals responsible for the legal review of organizational agreements

The Data Owner stakeholder is defined as the individual or organization that owns the data of interest. It is important to understand that the Data Owner and the Collaborator are not always the same individual or organization. The Collaborator is an individual or organization outside of ERDC-ITL that is assisting with data transfer, transformation, and/or analytics. Collaborators often approach ERDC-ITL to perform research activities using their data or data owned by partner organizations. Collaborator Support represents those people working directly with or under the Collaborator during the workflow process.

Rare instances exist where the ERDC Director and/or the ITL Director may be involved in the Data Lake Ecosystem Workflow. If involved, each individual's primary responsibility is to review and approve organizational agreements affecting the transfer of data to ERDC and/or ITL. The involvement of either individual is determined by the rank and grade of the signatory from the data-owning organization. Additional details may be found in the section titled Data Transfer Agreements.

The ITL Support stakeholder represents all ITL employees supporting a data lake research project. This includes the project investigator and all

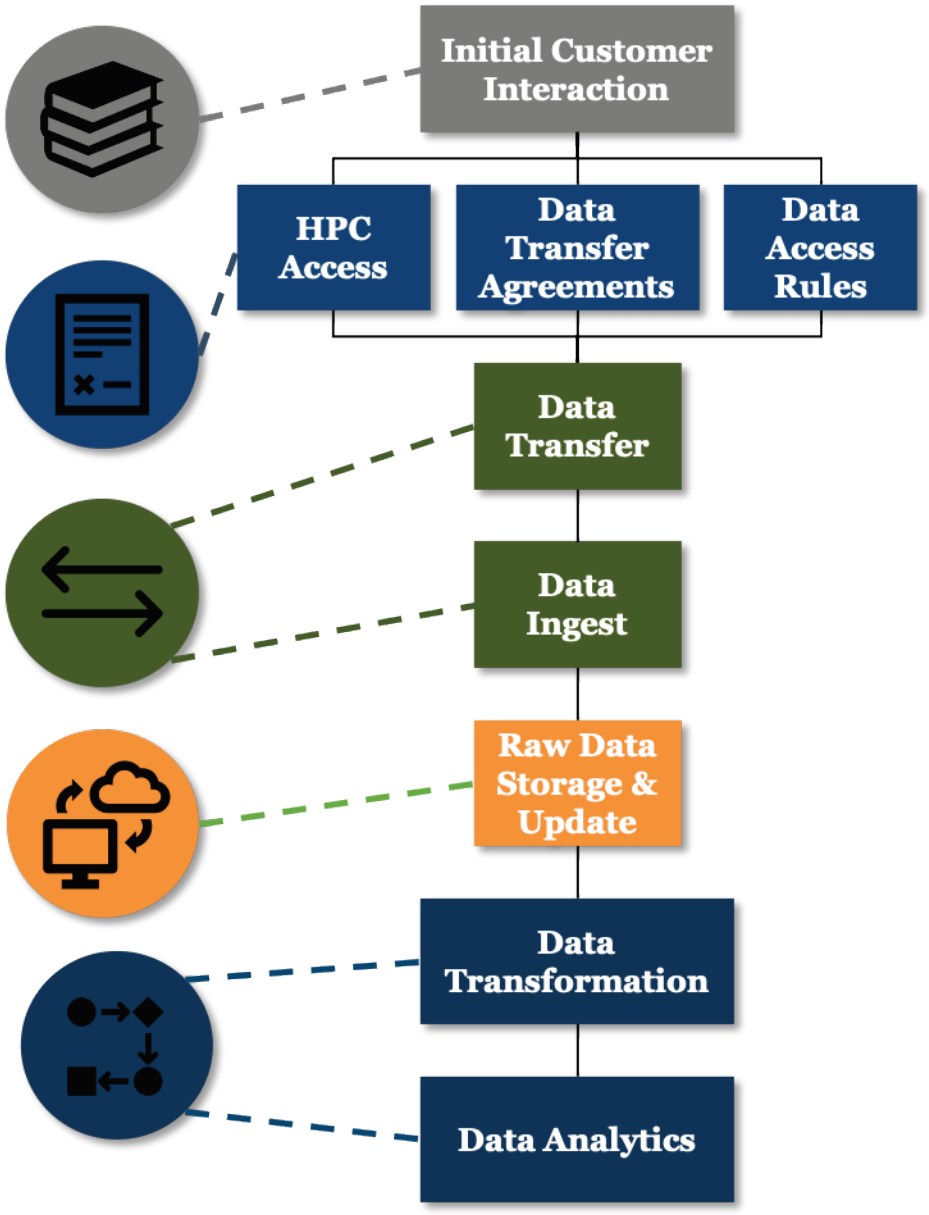
other supporting employees. HPC Support represents the team of individuals that manage and maintain the HPC platform as well as those individuals responsible for granting user access to the HPC environment.

In the case where organizational agreements are required, the laboratory Technology Transfer Officer (TTO) and the Office of Counsel must be involved. The TTO facilitates the review and approval of organizational agreements such as Memorandums of Understanding (MOU). The Office of Counsel participates in the review of organizational agreements to ensure that ERDC is legally protected when entering into agreements of this type.

4.2 Workflow processes

As stated previously, the Data Lake Ecosystem Workflow is composed of six phases. Each phase may be broken down into a collection of tasks and activities, known as a process. Each process is mapped to a single phase of the workflow as shown in Figure 3.

Figure 3. Phase to process map (End of Life phase is not included).



The first phase in the Data Lake Ecosystem Workflow is the Educate & Interact phase. This phase is composed of one process: Initial Customer Interaction. The Initial Customer Interaction process includes all preliminary interactions between the Data Owner, Collaborator, and ITL Support. This process includes tasks such as synergistic education and completion of a viability assessment.

After the Educate & Interact phase is complete, the Access & Agreements phase commences with three processes initiated in parallel. These processes are the HPC Access, Data Transfer Agreement, and Data Access Rules processes. The HPC Access process includes all activities required to gain access to DoD supercomputing assets. This process can be quite lengthy, so it is very important to initiate this process as soon as possible. The Data Transfer Agreement process outlines the steps needed to author, review, and approve a valid Data Transfer Agreement (DTA) or MOU. Like the HPC Access process, the Data Transfer Agreement process can take quite some time. This is due to the fact that multiple offices and approving officials must review organizational agreements in order to gain approval. Finally, the Data Access Rules process captures the steps needed to create a valid set of Data Access Rules.

The next phase in the workflow is the Transfer and Ingest phase made up of the consecutive execution of the Data Transfer and Data Ingest processes. The Data Transfer process may begin after the DTA and Data Access Rules processes are complete. Unique instances exist, however, where data transfer can take place before the establishment of transfer agreements and DARs. These instances occur only at the Data Owner's discretion. Regardless, the Data Transfer process documents the steps necessary to transfer data to sensitive HPC systems. The transfer of large datasets to HPC systems may be accomplished through multiple mechanisms, but this information will be reviewed later. The Data Ingest process includes those steps needed to upload data to the HPC. This process and its tasks are dependent upon the transfer mechanism used.

Once data is on the HPC, the Store & Update phase is begun at which time the Raw Data Storage & Update process is executed. Once all data is transferred and uploaded to the platform filesystem, it is then placed in an appropriately organized storage structure. While in storage, the data may sustain slight modifications and updates. These activities are captured in the Raw Data Storage & Update process.

The final phase in the Data Lake Ecosystem Workflow is the Transform & Analyze phase. This phase consists of two processes - Data Transformation and Data Analytics. These processes occur in sequence, and each captures the steps necessary for transforming and analyzing data, respectively.

At this time, the End of Life phase includes no processes since this phase has not yet been observed. Tasks and stakeholders will be documented once this activity takes place. Overall, the Data Lake Ecosystem Workflow is a lengthy process with multiple steps and opportunities for parallel effort. The complete Data Lake Workflow is provided in Figure 3. Please note that parallel processes are grouped in single rows. Descriptions of each workflow process are also provided in Table 3.

Table 3. Workflow process descriptions.

Workflow Process	Description
Initial Customer Interaction	This process includes all preliminary interactions with data owners and collaborators; tasks include collaborator education as well as completion of the viability assessment
HPC Access	This process includes all activities required for access to sensitive HPC systems
Data Transfer Agreements	This process outlines the steps required to author, review, and approve a valid Data Transfer Agreement (DTA) or Memorandum of Understanding (MOU)
Data Access Rules	This process outlines the steps for creating Data Access Rules
Data Transfer	This process documents the steps necessary to transfer and upload data to sensitive HPC systems
Data Ingest	This process includes the steps needed to upload data to the HPC system
Raw Data Storage & Update	This process outlines the steps taken to update the data lake
Data Transformation	This process shows the necessary steps for transforming the data
Data Analytics	This process shows the steps necessary for analyzing the data

The following subsections provide an in-depth look at each of the nine workflow processes that make up the Data Lake Ecosystem Workflow. Each process is composed of multiple tasks, but these tasks vary in complexity. Additionally, processes such as data transfer include many possible paths. The process facilitator must choose the appropriate path depending on defined criteria outlined in the following subsections.

4.2.1 Initial customer interaction

The Initial Customer Interaction process is composed of two primary activities: education and information collection. This process begins when a Data Owner or collaborator expresses interest in working with ERDC-ITL. Upon this expression of interest, ITL Support works to schedule and facilitate an initial meeting where ITL capabilities, successes, and processes are discussed. ITL Support also collects information on the dataset of interest during this meeting. To do so, a viability assessment (or questionnaire) is delivered to the Collaborator. This assessment is used to collect critical descriptive information about the dataset to be transferred. Once all information is collected, ITL Support and the Collaborator discuss tasking and future work and agree on the level of effort required to successfully accomplish the outlined tasks.

In support of the Initial Customer Interaction process, ITL developed two supporting documents. The first of these documents is a briefing designed to inform potential collaborators of ITL's capabilities and successes in the area of big data analytics. Future versions of this document will offer more general education on data science and high-performance computing.

The second document is a viability assessment. This fillable form was designed to collect information on the dataset(s) of interest in order to better inform the Data Transfer, Ingest, Transformation, and Analytics processes. Additionally, the viability assessment is used as the starting point for data documentation and metadata creation.

4.2.2 HPC access

During the HPC Access process ITL Support has the primary duty to facilitate access to HPC resources by pointing new users to their Service/Agency Approval Authority (S/AAA) and other HPC Support personnel. No documentation was created in support of this workflow process, as all required information is located on the Department of Defense High Performance Computing Modernization Program website (https://www.hpc.mil/).

4.2.3 Data transfer agreements

Before transferring data between government organizations, many data owners require that a DTA be in place. DTAs come in many different forms. ITL's Big Data Analytics team has experience with three primary agreement types: MOU, Memorandum of Agreement (MOA), and data sharing agreement (DSA). According to the Army regulation *Preparing and Managing Correspondence,* an MOU is used to "describe broad concepts of mutual understanding, goals, and plans shared by parties when no transfer of funds for services is anticipated" (HQDA 2013). An MOA, however, is used to "establish and document common legal terms that establish a 'conditional agreement' where transfer of funds for services is anticipated" (HQDA 2013). Both the MOU and MOA tend to be general organizational agreements. The DSA is tailored specifically for the transfer of data between government organizations. DSA's cover ideas such as intellectual property, data/derivative data ownership, data sharing, and data destruction.

The primary objective of the Data Transfer Agreement process is to establish a DTA (e.g., MOU, MOA, or DSA) between ERDC-ITL and the Data Owner. The first step of this process is to determine the appropriate author of the DTA. Data owners with a high level of concern regarding the storage and security of their data will likely choose to author the agreement. Additionally, owners with data transfer experience generally prefer to author transfer agreements. More relaxed owners or owners with less experience, however, may allow ITL to author the agreement. Regardless of organization, the DTA author is usually the principal investigator or technical point of contact (POC) for the project under consideration. Once the author is identified, the appropriate signatories must be selected. Typically, a DTA is signed by two officials, one from each organization. These individuals must have the same grade. For example, if the partner organization selects a signatory at grade GS-15, ITL must select a signatory of equivalent rank.

After the author and signatories are selected, the author may write the DTA. Upon completion, the author should submit the DTA to their organization for internal and legal reviews. At ITL, a DTA is first reviewed by a branch and division chief. After reviewing the DTA, the division chief passes the DTA to the laboratory TTO. The TTO is responsible for

facilitating the review and signing of all DTAs. Upon receipt of the agreement, the TTO will likely review the document. Once reviewed, the TTO will send the DTA to the Office of Counsel for legal review. The DTA will then be reviewed by the partner organization (if not already reviewed) and signed by each party's designated signatory.

The Data Transfer Agreement process is not complicated, but this process can take a substantial amount of time due to the number of people and offices that must review the DTA. In summary, a DTA must be reviewed and agreed upon by both organizations. The review and approval process may change on a case-by-case basis, but both parties must reach an agreement before signing the document. Additionally, officials of equivalent grade from each organization must approve and sign the DTA. Once signed, data may be transferred between the two organizations. If needed, DTA templates and examples exist for the purpose of aiding authors with DTA creation. DTA templates are provided through the Army Knowledge Online (AKO) hub (https://www.us.army.mil/content/armyako/en.html).

4.2.4 Data access rules

To ensure that collaborator data is properly stored and managed, ERDC-ITL often requires all data users to sign DARs. DARs are an individual agreement outlining allowable data use activities for users of a specific dataset. All individuals with access to a dataset must sign DARs. Users must also strictly adhere to the guidelines presented in the DARs. In rare cases, some data owners may not require that DARs be in place. It is highly suggested that DARs be established, however, to help ensure data security.

Collectively, the Data Access Rules (DARs) process is somewhat similar to the Data Transfer Agreement process previously outlined. The first step in this process is to determine the author of the DAR. ERDC-ITL has a DAR template and multiple DAR examples to aid in the creation of a DARs document. Once created, the DAR should be reviewed and agreed upon by both parties. At the writing of this report, DARs did not require branch chief, division chief, TTO, or legal review. Leadership review is allowed and encouraged as it provides an opportunity to obtain constructive feedback, and it promotes leadership situational awareness.

DARs lose their effectiveness if not enforced. Thus, ERDC-ITL suggests that all partner organizations establish a data access controller to help facilitate data access permissions. This individual may be selected before the DAR are accepted by all parties. The primary responsibilities of this individual include the following: (1) ensure that all initial data users sign the DAR, (2) create and maintain a list of allowed data users throughout the life of the project, (3) provide a list of data users to the Data Owner as needed/requested, (4) ensure that new users agree to the DAR, and (5) facilitate the removal of delinquent or non-compliant users from the data working group.

In closing, DARs are an individual-level agreement that outlines data use best practices. The DAR process may be consolidated into four basic steps: author, agree, enforce, maintain. The DAR must be authored at project inception. Once authored, all parties must agree on the practices provided. Individuals at each partner organization must then enforce the DAR by facilitating document signing and data permissions. These individuals must also maintain data permissions as users enter and exit the working group or organization.

4.2.5 Data transfer

The next process in the Data Lake Ecosystem Workflow is the Data Transfer process. The first step of the Data Transfer process is to ensure that a DTA is in place between ERDC-ITL and the Data Owner. Data transfer may only take place after ERDC-ITL and the Data Owner sign a DTA. Data transfer may not occur before this moment unless allowed by the Data Owner. If data transfer is permitted without the establishment of an organizational DTA, consider obtaining a written statement from the Data Owner documenting this fact. In the future, ITL leadership may require DTAs be in place before the transfer of any datasets may take place.

Once an agreement is in place, ITL Support and the Data Owner must determine the appropriate data transfer mechanism. To date, five transfer mechanisms exist: Mail Delivery, Hand Trade, HPC Upload, Online File Transfer, and Encrypted Email. The Mail Delivery transfer mechanism requires the Data Owner to download their data onto a storage device, such as an external hard drive, and to mail the hardware to ERDC-ITL. If the device is password protected, the password should be emailed to ERDC-

ITL using encrypted email. Some data owners are not comfortable with entrusting their data with a mail courier. Thus, they prefer that data and hardware be exchanged in-person. Situations such as this are best facilitated using the Hand Delivery transfer mechanism. When hand delivered, all data must be placed on a storage device and delivered, in-person, to ITL Support. In such cases, one of three exchanges may occur: (1) the Data Owner may travel to ITL to deliver the data, (2) ITL Support may travel to the Data Owner to retrieve the data, and (3) the two parties may meet in a mutually agreeable location to exchange the data.

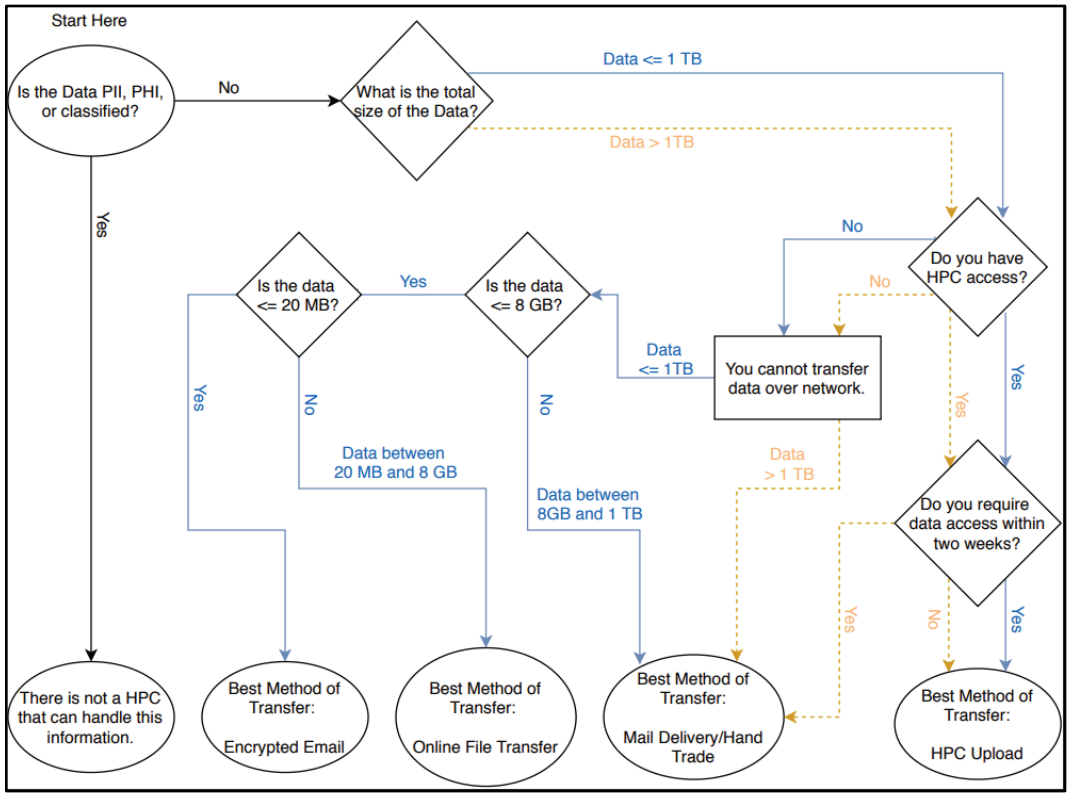
Users with access to HPC assets at ERDC-ITL may choose to upload their data directly to the HPC. This mechanism, known as HPC Upload, is best for collaborators with large datasets and no immediate time constraints. The transfer of data directly to the HPC can be a slow, frustrating process because of its dependency on government internet networks. Network constraints, such as internet upload speed, download speed, and network reliability, between the Data Owner's organization and the data destination should be considered before choosing this transfer mechanism.

The transfer of small datasets is best facilitated by one of the two following mechanisms: Online File Transfer and Encrypted Email. The Defense Information Systems Agency (DISA) hosts a safe and secure online file transfer service known as DoD SAFE (https://safe.apps.mil/). This service is best suited for data ranging in size from 20MB up to 8 GB. The final transfer mechanism is encrypted email. This transfer mechanism is best for datasets below 20MB. Please note that the Online File Transfer and Encrypted Email transfer mechanisms are frequently used to update larger datasets. If the complete dataset of interest is small enough to be transferred using one of these mechanisms, all parties should reevaluate the use of HPC assets to analyze the dataset of interest.

The information discussed above is shown graphically in Figure 4. This decision tree may be used by ITL Support and the Collaborator to determine the appropriate data transfer mechanism based on three parameters: data size, HPC Access, and time constraints. Please note that Personally Identifiable Information (PII), Personal Health Information (PHI), or classified data are not permitted on HPC assets at ERDC-ITL. However, the potential to store and analyze information of this nature is high if funding were to be provided to procure resources for this purpose.

Overall, the Data Transfer process is a critical step in the Data Lake Ecosystem Workflow. Improper data transfer could affect all downstream steps. Additionally, this process represents a critical data security moment. ITL Support should work to ensure that the optimal transfer mechanism is selected while also working to satisfy all Data Owner and collaborator transfer requirements.

Figure 4. Data transfer decision tree.



4.2.6 Data ingest

The sixth step in the Data Lake Ecosystem Workflow is the Data Ingest process. The activities that make up the Data Ingest process are dependent upon the chosen data transfer mechanism, but the overall ingest process may be summarized in a few basic steps. The primary objective of the Data Ingest process is to upload transferred data onto the HPC. Once uploaded, the raw data must be converted to a format that is compatible with the selected database management system. These two tasks represent the overall ingest process, but slight modifications are required to accommodate each unique transfer mechanism.

When data transfer is executed using the HPC Upload mechanism, data ingestion is a simple process. Once the data is placed on the HPC using the HPC Upload mechanism, ERDC-ITL must ensure that the data is placed in the correct directory. If the data was not placed in the correct location, ITL Support must move the data to the correct directory. At that point, the data must be converted into a format that is compatible with the database management system to be used.

The Mail Delivery and Hand Trade mechanisms share the same ingestion tasks. To begin, all necessary hardware must be collected, and the data must be uploaded from the storage device to the HPC workgroup directory using an intermediate machine. Please note that the intermediate machine must have access to the HPC. Additionally, the data should not be placed on the intermediate machine. The intermediate machine simply facilitates the transfer of data from the storage device to the HPC. If data is mistakenly placed on the intermediate machine, the personnel involved should ensure that all data is permanently deleted from the intermediate machine. This may require that the intermediate machine's recycle bin or trash be emptied to ensure thorough removal of the deleted data. When transfer is complete and the data is placed in the appropriate directory, the raw data is converted into a format that is compatible with the database management system to be used. Afterward, the data storage hardware should be returned to the Data Owner, if necessary.

The Online File Transfer and Encrypted Email transfer mechanisms share similar ingestion tasks, but slight differences exist between the two. In general, the first step is to download the data from the encrypted email or the online file transfer service to the intermediate machine. Once downloaded, ITL Support may upload the data to the HPC from the intermediate machine. If the data was received through an encrypted email, however, ERDC-ITL must first decrypt the data before transferring the data to the HPC. After the data is transferred to the HPC, ITL Support must then convert the data to a format that is compatible with the selected database management system. Finally, ERDC-ITL must then permanently delete all data from the intermediate machine.

Overall, the Data Ingest process may be consolidated into two distinct tasks: data upload and data formatting. Mild process variations exist and are dependent upon the selected transfer mechanism, however. The

conclusion of the Data Ingest process marks the end of the Transfer & Ingest phase and the beginning of the Store & Update phase. The first, and only, process in this phase is the Raw Data Storage & Update process.

4.2.7 Raw data storage and update

Within the Data Lake Ecosystem Workflow, the Raw Data Storage & Update process is the simplest process. This process has two fundamental objectives, to include the provision of safe data storage and the facilitation of data updates as required. Supercomputing resources located at ERDC-ITL have the capacity to store approximately 13 petabytes of data (ERDC DoD Supercomputing Resource Center 2012). To date, ERDC-ITL has successfully stored datasets as large as 23 terabytes while also managing data access and ensuring data security.

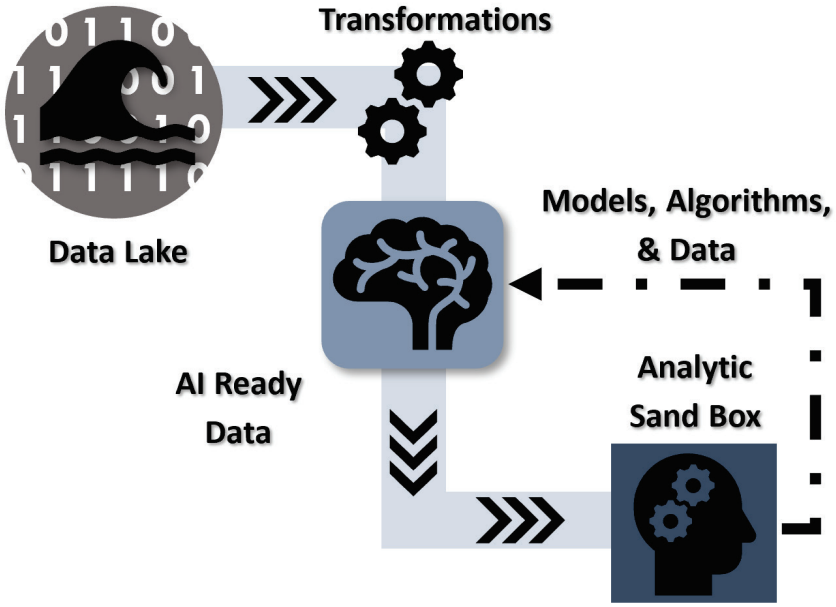
In some cases, data owners continue to collect data after their data is transferred to ERDC-ITL, thus requiring updates to their initial dataset. ERDC-ITL can facilitate this by using one of the five transfer mechanisms discussed previously. If an update is required, new data must be ingested and joined with the master dataset using the ingestion process previously outlined. Multiple, frequent updates may be facilitated as well. Once all data is stored on the HPC, users can harness the power of HPC to perform transformations and analyses across large datasets.

4.2.8 Data transformation and data analytics

The final processes in the Data Lake Ecosystem Workflow are the Data Transformation and Data Analytics processes. These processes are a part of the Transfer & Ingest phase, and they mark the end of the Data Lake Ecosystem Workflow. From a workflow perspective, the transformation and analytics processes are simple and are composed of few tasks. In reality, these processes require large amounts of time and effort and mark the true beginning of research. Three basic activities make up the Data Transformation and Data Analytics processes: confirm, perform, and store. ERDC-ITL should confirm all required transformations and analyses with the Data Owner before initiating the work. When confirmed, necessary transformations and analytics should be performed. Any resulting datasets, algorithms, models, and insights should then be stored for later use and reference.

The Data Transformation and Data Analytics processes are at the core of the Data Lake Ecosystem. As data is transferred and ingested, it is stored in a data lake. Recall that the primary characteristic of a data lake is its immutability. This means that the contents of a data lake cannot be changed unless directed by the Data Owner. In many cases, however, raw data received from the Data Owner requires extensive transformation before it can be analyzed. Thus, ITL Support must perform the appropriate transformations in preparation for analysis. This transformed data is no longer a part of the original data lake and must be stored in a new repository. This repository is known as the AI-Ready data repository. As the title indicates, the AI-Ready data repository is home to transformed data that is suitable for use in artificial intelligence algorithms. Users with the required permissions may access this repository to perform analyses as needed. These analyses are performed in the Analytic Sand Box, a secure environment in which users may create models and algorithms to gain data-informed insights. In some cases, these models and algorithms may inform future analyses. For this reason, some models, algorithms, and derivative data are submitted for storage to the AI-Ready data repository. Figure 5 graphically shows the described ecosystem.

Figure 5. The flow of data from the data lake to the AI-Ready data repository upon transformation. The AI-Ready data is used in the Analytic Sand Box to inform analyses, thus resulting in models, algorithms, and derived data.



Overall, the subject matter expert selects the activities that make up the data transformation and analytics processes. In general, these processes follow the *confirm*, *perform*, and *store* framework. More importantly, however, it is critical to understand how the data transformation and analytics processes fit within the context of the data lake ecosystem, as the infrastructure and practices composing the ecosystem enable large-scale data transformation and analysis.

5 Conclusion

Overall, the Data Lake Ecosystem Workflow represents a series of tasks designed to safely and securely transfer large datasets to supercomputing resources for storage and analysis. The Data Lake Ecosystem Workflow is composed of six phases: (1) Educate & Interact, (2) Access & Agreements, (3) Transfer & Ingest, (4) Store & Update, (5) Transform & Analyze, and (6) End of Life. Each of these six phases may be further decomposed into a number of processes. These processes provide the detailed information needed to transfer, ingest, transform, and analyze large-scale datasets.

The Data Lake Ecosystem Workflow begins with stakeholder education and information collection. When all parties agree on a satisfactory path forward, parties must then work together to author and approve all required organizational and individual agreements while also seeking access to the high-performance computer. Once complete, the Data Owner may transfer all data to ERDC-ITL. Once transferred, the data is ingested, stored, and updated as need. Researchers may then perform all needed transformations and analyses on the datasets in hopes of gaining valuable insights.

The workflow described in this document provides current and future ITL employees with a standardized data transfer approach. The creation and adoption of this workflow provides multiple benefits to ERDC-ITL and to our collaborators. Firstly, standardized, documented workflows and processes facilitate collaborator and employee education. Understanding the complete process allows ERDC-ITL to provide collaborators with a view of their full participation in this process. It also provides ERDC-ITL with tools for educating new employees, thus spreading knowledge across multiple parties. Furthermore, workflow standardization coupled with education results in process efficiency and completeness. Increased efficiency reduces time spent on non-research tasks, which increases productivity and project impact. Also, operating across a standard, complete process protects all parties involved while also protecting data integrity.

References

Amazon Web Services. 2019. *What is a data lake?* https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/.

HQDA (Headquarters, Department of the Army). 2013. *Preparing and Managing Correspondence*. Army Regulation (AR) 25-50. Washington, DCL HQDA. https://armypubs.army.mil/epubs/DR_pubs/DR_a/pdf/web/r25_50.pdf

ERDC DoD Supercomputing Resource Center. 2012. *High Performance Computing Systems - Onyx*. https://www.erdc.hpc.mil/hardware/index.html.

Friedenthal, S., A. Moore, and R. Steiner. 2012. *A Practical Guide to SysML: The Systems Modeling Language*. 2d ed. Elsevier Inc.

IBM Corp. 2018. *Build A Better Data Lake*. https://www.ibm.com/analytics/data-lake.

Laplante, P. A. 2013. *Requirements Engineering for Software and Systems*. 3d ed. Boca Raton, FL: Auerbach Publications. https://doi.org/10.1201/b15939.

PwC (PwC Digital Services). 2015. *Data Lakes and the Promise of Unsiloed Data*. http://usblogs.pwc.com/emerging-technology/data-lakes-and-the-promise-of-unsiloed-data/.

Reinsel, D., J. Gantz, and J. Rydning. 2018. "The Digitization of the World From Edge to Core." *Data Age 2025*. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

Seale, M., G. Nabholz, A. Ruvinsky, L. Walker, S. Abdullah, A. Strelzoff, D. Martinez, A. Hines, W. Bond, G. George, J. Church, O. Eslinger, D. Wade, A. Wilson, and N. Rigoni. 2018. "Unlocking Insights to Black Hawk Maintenance Data Using Innovative Big Data Analytics and Management Techniques." *ERS Journal* Fall 2018, pp. 32-46.

Acronyms and Abbreviations

Abbreviation	Term
AKO	Army Knowledge Online
DAR	Data Access Rule
DISA	Defense Information Systems Agency
DSA	Data Sharing Agreement
DTA	Data Transfer Agreement
ERDC	Engineer Research and Development Center
ERDC-ITL	Engineer Research and Development Center, Information Technology Laboratory
MOA	Memorandum of Agreement
MOU	Memorandums of Understanding
NACI	National Agency Check with Inquiries
PHI	Personal Health Information
PI	Principal Investigator
PII	Personally identifiable information
TTO	Technology Transfer Officer
DoD	U.S. Department of Defense
FSO	Facility Security Office
HPC	High Performance Computing
HPCMP	High Performance Computing Modernization Program
POC	Point of Contact
SAM	Systems for Award Management

Appendix A: Process Workflows Activity Diagram

Table A-1 summarizes the Initial Customer Interaction process. This process includes the following stakeholders: Data Owner, Collaborator, Collaborator Support, and ITL Support.

Table A-1. Initial Customer Interaction process.

#	Data Owner & Collaborator	Collaborator Support	ITL Support	Notes
1	Express interest in Data Lake collaboration		Express interest in Data Lake collaboration	
2	Participate in elementary synergistic education	Participate in elementary synergistic education	Participate in elementary synergistic education	Use provided Education presentation to help facilitate discussion
3	Express desired outcomes of Data Lake collaboration		Communicate potential solutions to Collaborator and Data Owner	
4	Agree to continue collaboration		Agree to continue collaboration	
5	Schedule Initial Customer Interaction meeting	Schedule Initial Customer Interaction meeting	Schedule Initial Customer Interaction meeting	
6	Participate in Initial Customer Interaction meeting	Participate in Initial Customer Interaction meeting	Participate in Initial Customer Interaction meeting	Complete and store the viability assessment once finished
7	Express desired outcomes of Data Lake collaboration		Communicate potential solutions to Collaborator and Data Owner	
8	Agree on deliverables and “formally” begin collaboration		Agree on deliverables and “formally” begin collaboration	
9	Continue synergistic education as needed	Continue synergistic education as needed	Continue synergistic education as needed	

Table A-2 outlines the process for obtaining an account on the high-performance computer. Process stakeholders include the Data Owner, Collaborator, Collaborator Support, ITL Support, and HPC Support.

Table A-2. HPC Access process.

#	Data Owner, Collaborator, & Collaborator Support	ITL Support	HPC Support	Notes
0		Help facilitate the HPC Access process as needed		Complete the steps listed below if you do not have an HPC account
1	Visit the HPC Modernization Program website for access instructions			
2	Ensure that your security clearance or National Agency Check with Inquiries (NACI) is valid			Contact your Facility Security Officer if necessary. A NACI or security clearance is required to access sensitive HPC resources.
3	Identify your S/AAA by emailing require@hpc.mil			
4			Acknowledge requestor email and initiate HPC Access process	
5			Send requestor New Account Request Form and High Performance Computing Modernization Program (HPCMP) User Agreement	
6	Complete the New Account Request Form and the HPCMP User Agreement			
7	Submit the completed New Account Request Form and the HPCMP User Agreement to your S/AAA			
8	Complete the DoD Cyber Awareness Challenge training course			
9	Once complete, sign the provided training completion certificate			
10	Submit the signed training certificate to your S/AAA			
11	Ask your Facility Security Office (FSO) or Security Manager to send a Visit Request for you to the HPCMP Security officer via the ERDC Security Office			Visit the HPC Modernization Program website for additional information on Visit Requests.
12			Send requestor "pIE Approved by S/AAA	All requestors may not receive every email outlined; always

#	Data Owner, Collaborator, & Collaborator Support	ITL Support	HPC Support	Notes
			Pending SAM Activation" email Systems for Award Management (SAM)	check with the S/AAA for official account status updates.
13			Send requestor “pIE Account User Information Modified” emails if necessary	
14			Send requestor “pIE Account rejected…” emails if necessary	
15			Send requestor “pIE Account Ready for Activation” email	
16			Send requestor “pIE Account Activated by SAM” when pIE account is fully active	
17			Work with requestor to setup HPC accounts	
18			Activate HPC accounts	
19	Access HPC accounts after receipt of welcome notification email			
20		Notify HPC Support of members who no longer need HPC Access to a workgroup		
21			Remove necessary members from the desired workgroup	

Table A-3 outlines the Data Transfer Agreement process from the perspective of ITL authorship. This process includes eight of the nine workflow stakeholders. Table A-4 includes this same process, but from the partner organization authorship perspective.

Table A-3. Data Transfer Agreement process (authored by ITL).

#	Data Owner	Collaborator	ERDC Director	ITL Director	ITL Support	Technology Transfer Officer	Office of Council	Notes
1					Identify the research partner or collaborator			
2	Determine the appropriate author of the MOU	Determine the appropriate author of the MOU			Determine the appropriate author of the MOU			
3	Determine the appropriate signatories for the MOU	Determine the appropriate signatories for the MOU			Determine the appropriate signatories for the MOU			
4					Outline the roles and responsibilities of each party within the MOU template			Refer to the MOU template provided by Dr. Phoebe Lenear
5					Send draft MOU to the laboratory TTO for review			TTO – Technology Transfer Officer
6						Review draft MOU		
7						Send MOU to Office of Counsel for review		
8							Review and edit MOU as needed	
9							Return MOU to TTO	
10						Send MOU to ITL Principal Investigator (PI) for partner organization review		The PI is the ITL project and interaction lead
11					Forward draft MOU to partner organization for review			
12	Review and edit MOU as needed	Review and edit MOU as needed						

#	Data Owner	Collaborator	ERDC Director	ITL Director	ITL Support	Technology Transfer Officer	Office of Council	Notes
13	Return MOU to ITL PI	Return MOU to ITL PI						
14					Forward MOU to TTO for review and signature distribution			
15						Send MOU to Office of Counsel for final review		
16							Review and edit MOU as needed	
17							Return MOU to TTO	
18						Send MOU to appropriate ERDC signatories for review and signature		
19			Sign/review MOU	Sign/review MOU				
20			Return MOU to TTO	Return MOU to TTO				
21						Return signed MOU to PI		
22					Forward signed MOU to Data Owner or Collaborator for final signatures			
23	Sign/review MOU	Sign/review MOU						
24	Return MOU to PI	Return MOU to PI						
25					Send signed MOU to TTO for storage and final distribution			
26					Store and distribute finalized MOU as needed	Store and distribute finalized MOU as needed		

Table A-4. Data Transfer Agreement process (authored by Collaborators).

#	Data Owner	Collaborator	ERDC Director	ITL Director	ITL Support	Technology Transfer Officer	Office of Council	Notes
1					Identify the research partner or collaborator			
2	Determine the appropriate author of the MOU	Determine the appropriate author of the MOU			Determine the appropriate author of the MOU			
3	Determine the appropriate signatories for the MOU	Determine the appropriate signatories for the MOU			Determine the appropriate signatories for the MOU			
4					Receive MOU from partner organization			
5					Review and edit MOU as needed			
6					Forward MOU to lab TTO for review and distribution			This distribution will be to ERDC offices
7						Review and edit MOU as needed		
8						Send MOU to Office of Counsel for review		
9							Review and edit MOU as needed	
10							Return MOU to TTO	
11						Send MOU to PI for partner organization review and signatures		
12					Send MOU to partner organization for review and signatures			
13	Review and sign MOU	Review and sign MOU						

#	Data Owner	Collaborator	ERDC Director	ITL Director	ITL Support	Technology Transfer Officer	Office of Council	Notes
14	Send MOU to ITL PI	Send MOU to ITL PI						
15					Send MOU to TTO			
16						Send MOU to ERDC signatories for review and signature		
17			Review and sign MOU	Review and sign MOU				
18			Return MOU to TTO	Return MOU to TTO				
19						Return signed MOU to PI		
20					Send signed MOU to partner organization			
21	Store and distribute MOU	Store and distribute MOU			Store and distribute MOU	Store and distribute MOU		

Table A-5 outlines the many tasks composing the DAR process. The Data Owner, Collaborator, Collaborator Support, and ITL Support stakeholders participate in this process.

Table A-5. Data Access Rules process.

#	Data Owner	Collaborator	Collaborator Support	ITL Support	Notes
1	Determine who creates the Data Access Rules (DAR)	Determine who creates the DAR	Determine who creates the DAR	Determine who creates the DAR	Data Owners could be the Collaborator, Collaborator Support, or a completely separate entity. It is a case-by-case basis
2	If tasked to do so, create the DAR	If tasked to do so, create the DAR	If tasked to do so, create the DAR	If tasked to do so, create the DAR	
3	Complete DAR	Complete DAR	Complete DAR	Complete DAR	Only one entity should create the DAR
4	Designate Data Access Controllers for all organizations involved	Designate Data Access Controllers for all organizations involved	Designate Data Access Controllers for all organizations involved	Designate Data Access Controllers for all organizations involved	The Data Access Controller handles all access to the data working group for their particular organization; to gain access to a working group, employees must first consult the Data Access Controller
5	Agree on DAR	Agree on DAR	Agree on DAR	Agree on DAR	All organizations should review the DAR
6	Distribute DAR to all data users for review and signature	Distribute DAR to all data users for review and signature	Distribute DAR to all data users for review and signature	Distribute DAR to all data users for review and signature	
7	Sign DAR by all users	Sign DAR by all users	Sign DAR by all users	Sign DAR by all users	All data users must read and sign the DAR
8	Compile a list of data users who signed DAR	Compile a list of data users who signed DAR	Compile a list of data users who signed DAR	Compile a list of data users who signed DAR	
9	Send compiled list to appropriate party	Send compiled list to appropriate party	Send compiled list to appropriate party	Send compiled list to appropriate party	
10	Receive compiled list of users	Receive compiled list of users	Receive compiled list of users	Receive compiled list of users	
11	Maintain and update list of users as needed	Maintain and update list of users as needed	Maintain and update list of users as needed	Maintain and update list of users as needed	
12	Send updated list of users to appropriate party	Send updated list of users to appropriate party	Send updated list of users to appropriate party	Send updated list of users to appropriate party	
13	Receive updated list of users	Receive updated list of users	Receive updated list of users	Receive updated list of users	
14	Store list of users for safe keeping	Store list of users for safe keeping	Store list of users for safe keeping	Store list of users for safe keeping	

Table A-6 outlines the Data Transfer process. Multiple mechanisms exist for the transfer of data from owner repositories to supercomputing assets located at ERDC-ITL. These mechanisms are Mail Delivery, hand trade,

HPC Upload, encrypted email, and online file transfer. Six stakeholders participate in the Data Transfer process: Data Owner, Collaborator, Collaborator Support, ERDC Director, ITL Director, ITL Support.

Table A-6. Data Transfer process.

#	Data Owner	Collaborator	Collaborator Support	ERDC Director	ITL Director	ITL Support	Notes
1	Sign Data Transfer Agreement or verbally agree to transfer data without signed DTA	Sign Data Transfer Agreement or verbally agree to transfer data without signed DTA		Sign Data Transfer Agreement or verbally agree to transfer data without signed DTA	Sign Data Transfer Agreement or verbally agree to transfer data without signed DTA		We prefer to have a signed DTA; high priority scenarios may allow for data transfer before a Data Transfer Agreement is signed
2	Discuss possible methods of data transfer	Discuss possible methods of data transfer	Discuss possible methods of data transfer			Discuss possible methods of data transfer	Refer to Data Transfer Decision Tree
3	Agree on data transfer method	Agree on data transfer method	Agree on data transfer method			Agree on data transfer method	
4	Transfer data to ITL Support						
5						Receive data from data owners	
6						Notify data owners of receipt of data	

Tables A-7 to A-10 outline the various data ingestion processes. Determination of the proper ingestion process depends on the selected transfer process. As a result, the following tables outline the ingestion processes for each unique transfer mechanism.

Table A-7. Data Ingest - Mail Delivery or Hand Trade.

#	Data Owner	ITL Support	Notes
1	Send data to ITL Support		
2		Receive raw data from Data Owner via Mail Delivery or hand trade	Raw data is read-only access, except for the owner; this data is immutable.
3		Collect necessary storage devices, usernames, and passwords if needed	
4		Transfer raw data from Data Owner storage device to HPC using intermediate machine	Data Owner storage will likely be an external hard drive; the intermediate machine must have access to HPC; An intermediate machine will not be used if the data is too large
5		Place raw data in correct directory if needed	
6		Convert raw data into a format compatible with the Database Management System to be used	
7		Return Data Owner hardware to Data Owner	
8	Notify ITL Support of hardware receipt		

Table A-8. Data Ingest - HPC Upload.

#	Data Owner	ITL Support	Notes
1	Upload data to the HPC		
2		Receive raw data from Data Owner via HPC Upload	Raw data is read-only access, except for the owner; this data is immutable.
3		Place raw data in correct directory if needed	
4		Convert raw data into a format compatible with the Database Management System to be used	

Table A-9. Data Ingest - Encrypted Email.

#	Data Owner	ITL Support	Notes
1	Email data to ITL Support via encrypted email		
2		Receive encrypted email from Data Owner	
3		Download raw data from encrypted email to intermediate machine	Raw data is read-only access, except for the owner; this data is immutable.
4		Collect necessary usernames and passwords for decryption	
5		Decrypt raw data	
6		Upload raw data to HPC from intermediate machine	

#	Data Owner	ITL Support	Notes
7		Convert raw data into a format compatible with the Database Management System to be used	
8		Permanently delete data from intermediate machine	Be sure to permanently delete files by emptying the recycling bin or trash can.

Table A-10. Data Ingest - Online File Transfer.

#	Data Owner	ITL Support	Notes
1	Send data to ITL Support via online file transfer		
2		Receive raw data from online file transfer	Raw data is read-only access, except for the owner; this data is immutable
3		Download raw data to intermediate machine	
4		Upload raw data to HPC from intermediate machine	
5		Convert raw data into a format compatible with the Database Management System to be used	
6		Permanently delete data from intermediate machine	Be sure to permanently delete files and emails by emptying the recycling bin or trash can

The final three processes in the Data Lake Ecosystem Workflow are the Raw Data Storage and Update, Data Transformation, and Data Analytics processes. Tables A-11 and A-12 outline the many steps within these processes.

Table A-11. Raw Data Storage and Update process.

#	Data Owner	Collaborator	ITL Support	Notes
1	Discuss updating raw data if needed	Discuss updating raw data if needed	Discuss updating raw data if needed	
2	Send data updates to ITL Support as needed			Follow appropriate transfer methods as outlined through the Data Transfer Decision Tree
3			Receive data update from Data Owner	
4			Notify Data Owner of update receipt	
5			Begin Data Ingest process as described previously	
6			Maintain raw data as needed	

Table A-12. Data Transformation process.

#	Data Owner	ITL Support	Notes
1	Communicate required data transformations to ITL Support	Verify needed data transformations with Data Owner	
2		Transform data while preserving fidelity of original dataset	Potential data transformations: uncompressing, cleaning, engineering, or transforming data
3		Store transformed data in AI-Ready repositories	

Table A-13. Data Analytics process.

#	Collaborator	ITL Support	Notes
1	Communicate required data analyses to ITL Support	Verify needed analyses with Collaborator	
2		Create reproducible analytic products	
3		Store analysis results in appropriate location	

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
04/01/2021	Final Technical Report (TR)	

4. TITLE AND SUBTITLE
Data Lake Ecosystem Workflow

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT

6. AUTHOR(S)
R. Cody Salter, Quyen T. Dong, Cody A. Coleman, Maria A. Seale, Alicia I. Ruvinsky, LaKenya K. Walker, and W. Glenn Bond

5d. PROJECT NUMBER
s

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Engineer Research and Development Center (ERDC)
Construction Engineering Research Laboratory (CERL)
PO Box 9005,
Champaign, IL 61826-9005

8. PERFORMING ORGANIZATION REPORT NUMBER
ERDC/ITL TR-21-2

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Headquarters, U.S. Army Corps of Engineers (HQUSACE)
441 G St., NW
Washington, DC 20314-1000

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
This work was funded via ERDC FADs 0642033036, "FY20 ERDC ERS Program Support" and 0603833D8Z, "Engineering Science and Technology."

14. ABSTRACT
The Engineer Research and Development Center, Information Technology Laboratory's (ERDC-ITL's) Big Data Analytics team specializes in the analysis of large-scale datasets with capabilities across four research areas that require vast amounts of data to inform and drive analysis: large-scale data governance, deep learning and machine learning, natural language processing, and automated data labeling. Unfortunately, data transfer between government organizations is a complex and time-consuming process requiring coordination of multiple parties across multiple offices and organizations. Past successes in large-scale data analytics have placed a significant demand on ERDC-ITL researchers, highlighting that few individuals fully understand how to successfully transfer data between government organizations; future project success therefore depends on a small group of individuals to efficiently execute a complicated process. The Big Data Analytics team set out to develop a standardized workflow for the transfer of large-scale datasets to ERDC-ITL, in part to educate peers and future collaborators on the process required to transfer datasets between government organizations. Researchers also aim to increase workflow efficiency while protecting data integrity. This report provides an overview of the created Data Lake Ecosystem Workflow by focusing on the six phases required to efficiently transfer large datasets to supercomputing resources located at ERDC-ITL.

15. SUBJECT TERMS
Big data, Datasets, Electronic data processing—Workflow, Data curation--Workflow

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (include area code)
Unclassified	Unclassified	Unclassified	SAR	48	