Arundel *et al. Journal of Big Data* (2023) 10:146
https://doi.org/10.1186/s40537-023-00801-9

Journal of Big Data

RESEARCH Open Access

A guide to creating an effective big data management framework

S. T. Arundel[1*], K. G. McKeehan[1], B. B. Campbell[2], A. N. Bulen[2] and P. T. Thiem[1]

*Correspondence:
sarundel@usgs.gov

[1] U.S. Geological Survey, Center of Excellence for Geospatial Information Science, 1400 Independence Rd, Rolla, MO 65401, USA
[2] U.S. Geological Survey, National Geospatial Technical Operations Center, 1400 Independence Rd, Rolla, MO 65401, USA

Abstract

Many agencies and organizations, such as the U.S. Geological Survey, handle massive geospatial datasets and their auxiliary data and are thus faced with challenges in storing data and ingesting it, transferring it between internal programs, and egressing it to external entities. As a result, these agencies and organizations may inadvertently devote unnecessary time and money to convey data without existing or outdated standards. This research aims to evaluate the components of data conveyance systems, such as transfer methods, tracking, and automation, to guide their improved performance. Specifically, organizations face the challenges of slow dispatch time and manual intervention when conveying data into, within, and from their systems. Conveyance often requires skilled workers when the system depends on physical media such as hard drives, particularly when terabyte transfers are required. In addition, incomplete or inconsistent metadata may necessitate manual intervention, process changes, or both. A proposed solution is organization-wide guidance for efficient data conveyance. That guidance involves systems analysis to outline a data management framework, which may include understanding the minimum requirements of data manifests, specification of transport mechanisms, and improving automation capabilities.

Keywords: ADOM, Data movement, Ingress, Egress, Rclone

Introduction

In many organizations and governmental agencies, the processes involved in managing geospatial data are complex, expensive, time-consuming, and often difficult to understand [1–4]. For example, the U.S. Geological Survey (USGS) is mandated to collect, maintain, disseminate, and preserve data about terrestrial elevation, surface waters, orthoimagery, and other themes for the United States [5]. These data "themes" are administered by the National Geospatial Program (NGP). In addition, many other data themes, along with diverse scientific research datasets, are managed by other programs and mission areas within the USGS.

This work recommends a minimal standard for administering large (> several petabytes) data programs, emphasizing data's movement, or conveyance, through the managing system or systems, which we call a data management framework (DMF). A use case study of USGS NGP's 3D Elevation Program (3DEP) illustrates existing solutions' shortcomings, details of existing solutions' limitations and motivations, and our

recommendations. Research has focused chiefly on solutions for data management relative to storage, processing, analytics, and quality control (for example, [6, 7]), sometimes in support of specific decision support systems [8, 9]. In addition, some research addresses solutions for data conveyance and, specifically, the ingression and egression of data within systems [10]. However, these topics are often abridged within the scope of workflow system studies [11, 12]. Thus, existing solutions fail to support the program requirements.

Our work addresses the relative brevity of knowledge regarding big data conveyance, with a current focus on geospatial data. Following Li et al. [13], we propose a "general-purpose scalable framework" to benefit agencies and organizations whose business is the ingression and egression of big data. We seek to identify data conveyance standards to augment existing organizational data management standards. We also recommend steps toward current system discovery to support guidance development, define possible specifications of proposed environments, and review alternate solutions to meet these requirements.

Related work

The "Big Data revolution" instigated "disruptive innovations" across many organizations and disciplines over the last quarter century [14], resulting in unprecedented analytical opportunities and challenges [2, 15–18]. In the geographic sciences, some workers have interpreted the big data revolution as a recommitment to the Quantitative Revolution that swept the discipline beginning in the 1950s and the Geographic Information Science (GIS) Revolution of the late 1980s [19–22]. As big data required "innovation" across the data and geospatial sciences, some researchers have even suggested constructing a new discipline to investigate pertinent scientific inquiries: "information geography" (for example, [23] or "geographic data science" (for example, [24]).

Regardless of the nomenclature, the practical reality for many agencies and organizations—as echoed by the calls for a new discipline—is that managing big (geospatial) data is complicated, as the inherent size, speed, and heterogeneity magnify management requirements [25]. Many workers have attempted to describe the challenges of this big data revolution, with several identifying the "3Vs" of big data—high volume, high velocity, and great variety [16, 25–27]. To this, some researchers added additional challenges. For Goodchild [26], data quality is the fourth big data management challenge. Similarly, Marr [27] adds the additional "Vs" of veracity and value, noting that if big data volume, velocity, variety, and veracity are all managed competently, the analytical value will follow. Finally, Kitchin [16] adds several additional management challenges and big data characteristics, including the need for flexibility and scalability and the necessity of data being "uniquely indexical."

Yet, regarding data management, the first "3Vs" form a baseline of potential obstacles, although Madden [28] suggests this construct is too simplistic. High data volume can congest data ingress procedures and overwhelm storage capabilities. In contrast, high data velocity from streaming services, real-time sensors, or other sources often renders even recently collected data obsolete before it can be processed or analyzed [10, 26, 29]. Highly variable big data sources and types slow ingress to an organization's workflow, potentially flooding management systems [2, 10, 26, 30]. By one measure, unstructured

data constitute 90 percent of all big data, some from mobile phones and similar devices [29]. Invariably, mismanagement of the 3Vs at ingress results in downstream issues at egress, rendering data analytics suspect and potentially offloading critical processing tasks to end-users [26, 27, 31, 32].

In response to these big data challenges, creating effective database management ecosystems, leveraging advanced tools, such as machine learning algorithms, and providing appropriate user training would be valuable [28]. Rather than adding complexity, any response by data managers would ideally seek to simplify by returning to the goals and requirements of an organization's information systems [33]. Such an exercise, which would ideally be undertaken before implementing the recommendations generated through this study, might incorporate some form of the Zachman Framework. This framework seeks to identify the "5Ws1H" (what, where, who, when, why, and how) of an organization's information systems ontology at each data model level (conceptual, logical, physical) [34], even if the enterprise architecture and data workflow are thought to be already understood. The Zachman Framework has proven beneficial in designing and understanding geospatial system data models [35]. Many big data geospatial models involve an ontological framework to solve specific information systems problems [30, 36].

Importantly, as with the Zachman Framework, our DMF does not proscribe a single set of rules that can be applied without analytical consideration and contextual investigations. In general, our approach involves an overall understanding of the system, creating a data manifest, specifying transport mechanisms, and improving automation capabilities, all with an emphasis on advancing data conveyance. Thus, for this work, and in the spirit of other research, a DMF ideally includes:

1. A conceptual framework designed to frame data management and movement problems.
2. An analytical process to identify 'tools' (software, procedures, and methods that can be used to move data).
3. The identification of the minimal metadata required and how to organize it.
4. Specifications to meet data movement and automation requirements.
5. High-level rules designed to analyze and improve data workflows.

Within the data conveyance workflow, ingress and egress nodes constitute an information system's input and output points [37, 38]. Data ingression occurs when data are conveyed into the workflow from any input resource, a difficult task when data are integrated from several disparate sources [6]. For Varadharajan et al. [6], the concept of ingress was part of an extensive DMF that sought to take field data, ingress the data, and apply the appropriate metadata to the data, therefore allowing discoverability and access by a community of users by prevailing standards. Upon that foundation, the datasets' quality assurance and quality control could be applied, allowing for integration and new product generation, which would then be egressed into the broader scientific community [6].

In a case of a mineral management system, data ingression sources were diverse, including both structured and unstructured data [7]. Structured data resources ingested

by the system at ingress included GIS and other geological and business data, whereas internet, mobile, and video sources contributed unstructured data. According to this system design, conveyance initiation throughout the system was at the impetus of users and administrators, including data ingress and egress [7].

Other DMF examples offer more complexity regarding ingress and egress. For example, Arenas et al. [11] created a data management framework to establish a "secure data research" environment that specifically addressed the ingression and egression of data with different security sensitivities. In their solution, data conveyance procedures are governed by security tiers, specifically defined administrator roles, and an initial dataset classification at ingress that determines workflow path, among other features [11].

As mentioned previously, a primary data ingress challenge is storage capability, especially considering the characteristics of unstructured data, of which the schema is unknown or amorphous. Some organizations utilize a traditional data warehouse system to house their structured data, such as GIS files, while concurrently using extendable schema-on-read tools to store unstructured data across a distributed cluster, such as Apache Hadoop. Schema-on-read applies a schema to data as the data are extracted from a stored location rather than during storage upload (write). For example, Li et al. [7] used this bifurcated methodology to manage mineral data. Many other frameworks, including those in the geosciences, use Apache Hadoop programming models, such as MapReduce and Hadoop Spatial [30, 36, 38–44]. For example, Google engineers developed an early MapReduce analytics and management framework in 2004 [2, 45], although the company may have migrated its data processing away from the model in recent years [46]. Many of these frameworks leverage cloud computing resources to manage shared distributed services, including storage services [2, 29].

Other frameworks developed their own management tools [6, 47]. Hagedorn et al. [48] compared MapReduce and Apache Spark in processing and managing "Big Spatial Data" and found that their team's tool (STARK) recorded faster execution times in some instances. Other researchers developed a management and analytical tool known as DataSHIELD (Data Aggregation Through Anonymous Summary-statistics from Harmonised Individual levEL Databases), which manages data workflows "without physically transferring or sharing the data and without providing any direct access to individual-level data" [49]. DataSHIELD was designed to conform to privacy and confidentiality laws in the United Kingdom and address ethical concerns about sharing sensitive research data [49, 50].

Still, all big geospatial data management frameworks, methodologies notwithstanding, must confront the additional metadata challenge. Metadata provide documentation about the content, source, lineage, structure, accuracy, attributes, development, authorship, spatial and temporal resolution, authorized usage of data, and other potential data attributes [51, 52]. These wide-ranging concepts about metadata can be distilled into the common phrase "metadata are data about the data" [52]. Yet, the creation and maintenance of metadata are complex in any environment but are more so at the volume and velocity associated with big data [53].

Several relevant standards for geospatial metadata have been developed to help govern metadata management. For multiple disciplines and organizations, best metadata practices can be summarized in the Findable, Accessible, Interoperable, and Reusable

(FAIR) Guiding Principles [54]. FAIR aligns somewhat with the primary concerns of any data manager [55]. Yet, FAIR guidelines were designed to reduce end-user confusion and enhance "machine-actionability," which is the capacity to automate data management tasks, especially relative to data object identification and processing at ingress [54]. To this end, some have found that an emphasis on the FAIR principle of data reusability also allowed for the reusability of workflows [56].

In addition to FAIR best practice guidelines, geospatial data associated with most U.S. government agencies or projects must adhere to the Content Standard for Digital Geospatial Metadata (CSDGM) authorized by the Federal Geographic Data Committee (FGDC) [51, 52]. The CSDGM framework, in turn, is "harmonized" with the International Organization for Standardization (ISO) metadata guidelines (ISO 19115) [57]. Whether through the FGDC framework or the FAIR principles, establishing metadata conventions within a data management framework is shown to have organizational benefits, specifically—but not exclusively—related to the interoperability and quality of data [51–53].

As Goodchild noted [26], data quality is tied to big data management issues. Li et al. [7] underscored this idea when noting that "data is the basic strategic resource of national development" in the People's Republic of China. Yet, as the name suggests, big data do not lend themselves easily to quality assurance (QA) and quality control (QC). For example, Varadharajan et al. [6] semi-automated their QA/QC procedures for the U.S. Department of Energy's Watershed Function Scientific Focus Area (SFA) data using R Markdown and Jupyter Notebooks, in which automated algorithms identify anomalies in the large datasets obtained from field sensors. Regardless of the methodology, assuring data quality is necessary for a big data management framework, as the cost of carrying poor data is enormous, both in dollars—perhaps as much as $3.1 trillion per year to the United States economy [4]—and in poor decision-making [27]. Our solution conveys data to and from a data validation process, as illustrated through a use case—elevation data processing in the NGP.

Approach

Understanding the current environment

As mentioned in the Background section, beginning any investigation of an existing data management system with a discovery process is important. This process would ideally seek to understand an organization's information system's goals, requirements, and variables. In particular—and for our purposes here—this investigation will emphasize data conveyance. To abet this exploration, the approach begins by developing a system process workflow (Fig. 1). This workflow, developed by the program subject matter experts in many cases, advances a detailed understanding and visualization of data movement.

Whereas our focus is on data conveyance, the cornerstone of conveyance relies on the data themselves because most data movements are specific to data type and purpose. Thus, the next step is identifying and enumerating the data inhabiting the process workflow. After all, although the phrase "form follows function" is an architectural maxim, it can also be extrapolated to describe how data types and the "5Ws1H" shape the workflows in which data conveyances occur.

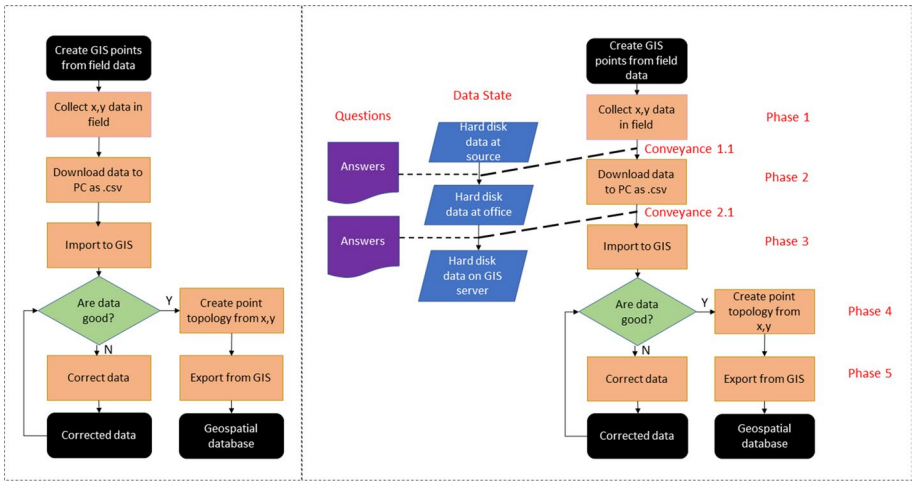

Fig. 1 Approach example of a process workflow (left) that captures the creation of geographic information system (GIS) points from *x, y* data collected in the field, and the resulting data management framework workflow (right), which identifies phases, conveyance, and data states to answer pertinent questions

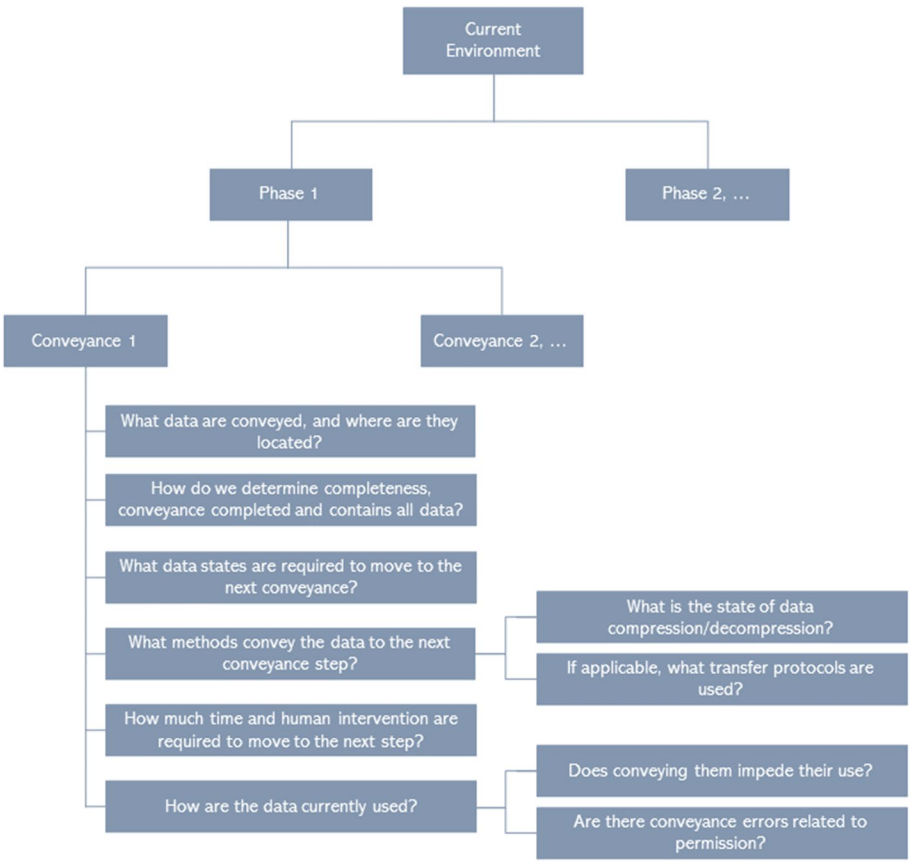

Fig. 2 As an exercise of the DMF, phases and their conveyance steps from the process workflow are identified to discover answers to questions

Next, we analyze the process workflow to discover pertinent phases, defined as seemingly discrete stages (Fig. 2). The discretization of phases is subjective and dependent on the system, data, and architectures. More specifically, system phases are typically composed of the locations of, mechanisms to, and reasons for repose within a data management workflow, including any minor data movements within a phase and data at repose, such as in the QA/QC process. These phases usually include ingress, internal data transfers, and egress and can be visualized as nodes connecting these larger movements.

Within the phases, conveyances are identified (Figs. 1, 2). Ideally, for each conveyance, the investigation determines what data exist, where the data are currently stored, and where the data are moving, in addition to basic metadata details, such as data providence. The cataloging of conveyances also notes the method of indexing/discovery and processes associated with metadata and management. Ultimately, this process leads to the construction of a data manifest, which is a listing and description of a dataset.

Regarding data conveyance, this data manifest specifically details and maps ingress/egress procedures related to the datasets and defines how data move through a workflow. For our purposes, a fully documented data conveyance has (1) preconditions to trigger, (2) a trigger, (3) a transportation method, and (4) postconditions to declare completion. Such procedures within any system involve a list of conditions that must be attained before conveyance to the next phase. What, for example, triggers data movement from data storage to QA/QC? Does that trigger involve a determination of completeness or the fulfillment of certain processing tasks? These conditions and triggers are delineated in the investigatory process.

Regarding completeness, discovery participants review how data are known to be complete as a precondition of any conveyance, either at ingression or egression or internally within phases or between phases. Completeness, a key component of data accuracy [52], should be defined clearly, along with the process by which completeness is determined. Other workflow process characteristics of importance include the state of data compression/decompression and whether and what transfer protocols are used. The answers to these and other questions reveal the duration of phases and conveyance steps, elucidating movement speed within the workflow through human intervention and network/disk speeds. Finally, it is important to understand whether and where network transfer issues are related to problems with file/system access and whether those transfers impede the use or processing of data.

Our DMF realizes the necessary tasks identified through this approach. The answers to questions at each framework step inform the subsequent solution. Any solution to the myriad of conveyance challenges includes a conceptual framework that informs data management and conveyance issues, specifically by identifying potential tools (software, procedures, methods) to move data. This solution identifies the ideal metadata, data movement, and automation requirements. The resulting high-level rules developed from this discovery procedure assist the analysis and improvement of data workflows.

Define requirements of desired environment

As a supportable data management framework is constructed, defining the ideal technical environment in which the data workflow could be conducted is important. In addition, the management and infrastructure systems necessary to support the identified

data conveyances must be enforced for the desired environment to standardize framework invariants. For instance, the information uncovered by the investigation may demand a data management environment in which the movement rates are generally greater than the current rate of operation; in fact, this deficit of velocity is often a primary reason to conduct the data management analysis in the first place [29]. In crafting the desired data management environment, several elements can be considered to determine the required and possible conveyance rates, including file compression/decompression methods and rates; data sharing protocols; enumeration methods, data manifests, and metadata; and orchestration protocols and data state changes.

The data compression and decompression methods currently used in a workflow also help dictate the ideal environment, as these methods foster efficient use of data pipelines. Compression and decompression methodologies are often data-specific, especially in geospatial workflows [58]. For example, lidar.las files require compressed LAS file (LAZ) compression, whereas raster files may be subject to a choice of many compression techniques. Likewise, simpler, two-dimensional data types may necessitate quadtree compression or something similar [58]. In contrast, other more complex data types and architectures may require more complicated compression algorithms such as Huffman encoding (for example, [59]).

In addition, file-sharing protocols ideally standardize the communication between data servers (senders) and data clients (receivers) and ultimately aid in this conveyance. The desired standardization would ideally tend toward a single method by data type. Identifying an organization's desired data management environment would also be based partly on the enumeration of the files shared between senders and receivers under guiding data protocols, internally or externally. Finally, file collections need to be discoverable (through search), even when data are unstructured. For example, collections on disk not appearing in any database could be scanned for manifests, rendering them discoverable. Consequently, discoverability can be considered the process of building the file index. Thus, the data manifest is required to conceptually transform dataset bits into valuable resources. Without it, data are unusable in processing or analysis, leaving them no value to offset their cost.

Ideally, a specified data state signals that they are ready for transfer, validating what data will be shared to some degree. Ready data can be defined as those that have been validated to a certain extent, compressed, needed, or other important elements. Data would not be transferred before the specified ready state is reached, at which point the orchestration protocol—an agreed interaction to start the transfer—is triggered. Data states and orchestration protocols enable automation. Through the labor of all these steps, the desired data management environment becomes apparent. These steps guide a specification of minimal metadata and data conveyance mechanisms.

Results and use case

Overview

Our resulting solution, which follows from the DMF discovery process, with its emphasis on data conveyance, has some advantages. Other solutions detailed in our literature review are helpful conceptually [7]. However, many are situation-specific [47] or tool-specific (for example, [48]). By contrast, we present a flexible framework and an

accompanying open-source tool. Our framework can be used to identify challenges such as data bottlenecks, bad transfer methods, or poor accounting. Additionally, this framework is expected to pinpoint gaps in data management, improve data transfer methods, and provide a minimum metadata guideline necessary for cataloging, mindful of leveraging automation. Below we illustrate an implementation of our framework to address challenges in a USGS use case.

Use case discussion

USGS 3D elevation program

USGS NGP products are most commonly used in transportation and navigation, natural resources and mining, infrastructure and utilities, and urban planning and management [60] (Table 1). The USGS 3DEP (https://apps.nationalmap.gov/downloader/#/) products are the second-most accessed NGP product after US Topo. 3DEP, which was developed from the National Elevation Dataset, facilitates the planning, quality control, and delivery of high-resolution elevation data, such as light detection and ranging (lidar)

Table 1 Some relevant USGS NGP products and end-user egress statistics

		Files	Size
Topographic maps	US Topographic Maps		
	Current	65,240	3.0 TB
	Historical	162,067	4.0 TB
	Total	227,307	7.0 TB
	Historical Topographic Map Collection (HTMC)		
	PDF	372,515	2.0 TB
	GeoTIFF	183,113	1.5 TB
3D elevation program (3DEP)	Lidar Point Cloud (LPC)		
	Total	–	259 TB
	Digital elevation model (DEM)—current		
	2 arc-second	1,486	5.0 GB
	1 arc-second	11,536	121.6 GB
	1/3 arc-second	4,414	457.3 GB
	1 m	160,362	15 TB
	Digital elevation model (DEM)—historical		
	2 arc-second	2,250	5.1 GB
	1 arc-second	17,483	145.6 GB
	1/3 arc-second	8,050	641.6 GB
	1/9 arc-second	8,345	626 GB
	Original Product Resolution (OPR)	–	27 TB
	Alaska Interferometric Synthetic Aperture Radar (ifsar)		
	Digital Elevation Model (DEM)	–	415 GB
	Digital Surface Model (DSM)	–	249 GB
	Orthorectified Radar Image (ORI)	–	3.96 TB
	Base Maps		
	US Topo with Imagery Tile Cache	–	1.57 TB
	USGS Hydro Cache	–	815 MB
	USGS Imagery Only Cache	–	1.21 TB
	USGS Shaded Relief Only Cache	–	276 GB
	USGS US Topo Only Cache	–	854 GB

and interferometric synthetic aperture radar (ifsar) data and their derivatives [60]. The program handles very big data, on the order of 400 to 700 terabytes (TB) per year. Most incoming 3DEP data are generated by consulting partners contracted to obtain the various 3DEP products and must adhere to more than 600 data integrity requirements.

3DEP, however, consists not only of the incoming elevation data but also data that are contributed outside the contracting structure and, hence, all related infrastructure. This includes all pertinent USGS and contract personnel and the lidar collection project tracking software; the VALID8 software, which aids in the manual internal quality control of incoming data; and the automated ingestion of new data into the existing elevations layers, such as the 1-m and 1/3-arc-second datasets using software called LEV8 (pronounced *elevate*).

Given the number of partners and staff involved, the complexity of the process through which data are moved and tracked, especially at ingression, and the sheer quantity, the 3DEP program presents an opportunity to implement our approach and study the results. For example, some 3DEP data conveyances involve the manual movement of files, creating shortcomings with the ingestion, use, and management of these data. More specifically, the manual manipulation of the file space during some processes results in at least one dataset deletion per quarter, one accidental data movement per month, and quarterly Amazon Simple Storage Service (S3) data sync failures. These shortcomings are compounded by the sheer size of the datasets and the volume of data being ingested. When data are compromised, recovering datasets that are already difficult to back up is very challenging.

However, changes in data management strategies, data storage technologies, data distribution methodologies, data types, the relative size of the average dataset download, and the growing volumes of data have further affected how the USGS provides access to and distributes its large datasets. Consequently, acquiring those data as an end user becomes a challenge, with some users experiencing long wait times (days to weeks) to acquire an elevation dataset or failing to acquire it at all. Thus, the 3DEP environment presents an optimal use case to apply our DMF.

Existing use case workflow and movement analysis

The current workflow reflects the complexity of the 3DEP environment (Fig. 3), beginning from the point where data are available for ingress to the USGS National Geospatial Technical Operations Center (NGTOC). Typically, a data delivery includes a single 3DEP contracted project, which is large enough to make data download difficult in the current environment. Thus, most data are ingested from hard drives manually connected to servers and copied using Windows file explorer or a Robocopy command. NGTOC staff also manually create a drive manifest and update the workflow status in the internal tracking system and the contracting office database. This indicates to the QC staff that the data are ready for analysis.

On the other hand, most downloaded data are pilot data or data corrected after an earlier rejection from the QC process. Thus, they are smaller than the original project deliveries. However, even these projects often fail to download after many attempts. Consequently, staff manually request a hard drive from the 3DEP contractor through email. Data failing QC may need to be moved and tracked through this system several

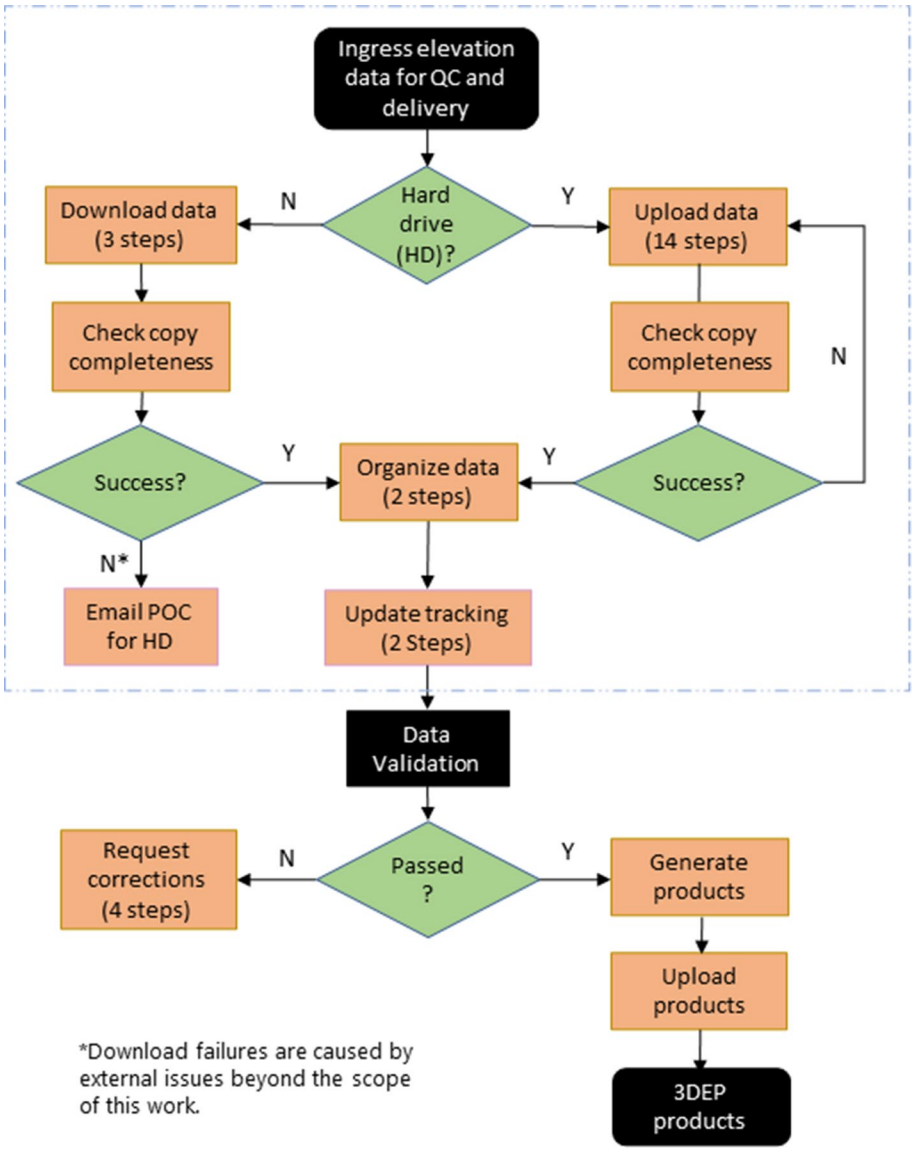

Fig. 3 Original 3DEP workflow. POC – point of contact

times before passing, often resulting in several sets of transfer processes, tracking, and copies of almost exact datasets. With the administration of so much data, an improved workflow would be beneficial.

Use case framework

Thus, we reorganized the process workflow based on our DMF to use phases to identify data states (stages) and conveyance needs between stages (Fig. 4). As a result, the following stages, conveyances, and facts were uncovered in our framework discourse and ultimately led to a final improved workflow (Fig. 5).

Phase 1 Stage: Hard drive data at source

How does one determine what data at ingestion exist and where?

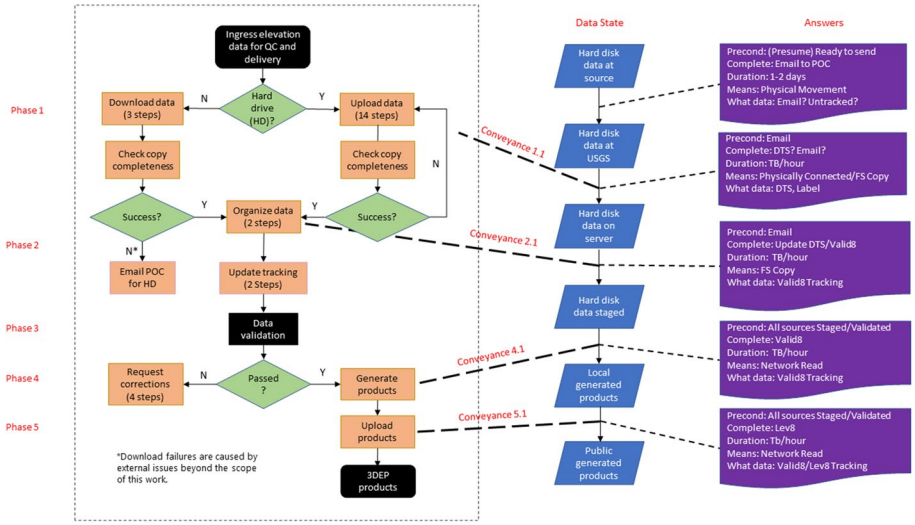

Fig. 4 3DEP process workflow transformed into a data management framework workflow, indicating conveyances, data states, and tracking information

This question, for 3DEP, is currently difficult to answer. For example, one may know where the data are with the contractor but not how they will be organized into work units within the contracted projects. However, the contracting point of contact (POC) can provide information about what *sets* of work units should eventually be part of a project, along with a low-precision area of interest. In addition, the contractor may provide a shipping tracking number via the contracting staff.

Conveyance: to USGS

1. Data move to the USGS once the contractor decides a work unit is ready, whether it is relative to 3DEP QC requirements.
2. The hard drive is shipped (often several at once).
3. The drive(s) arrive(s).
4. The Shipping and Receiving team sends an email to the USGS point of contact (POC) with metadata for the Drive Tracking workSheet (DTS), which the POC populates.

    - Minimum metadata requirements include but are not limited to Collection/Object Schema

        – Dataset ID
        – File Manifest

    - Hierarchy List
    - Possible Hashing –Required for verification of file contents if SHA-256 (Secure Hash Algorithm) is used
    - Hierarchy List
    - Transfer validation versus at-rest corruption

        – Spatial Class

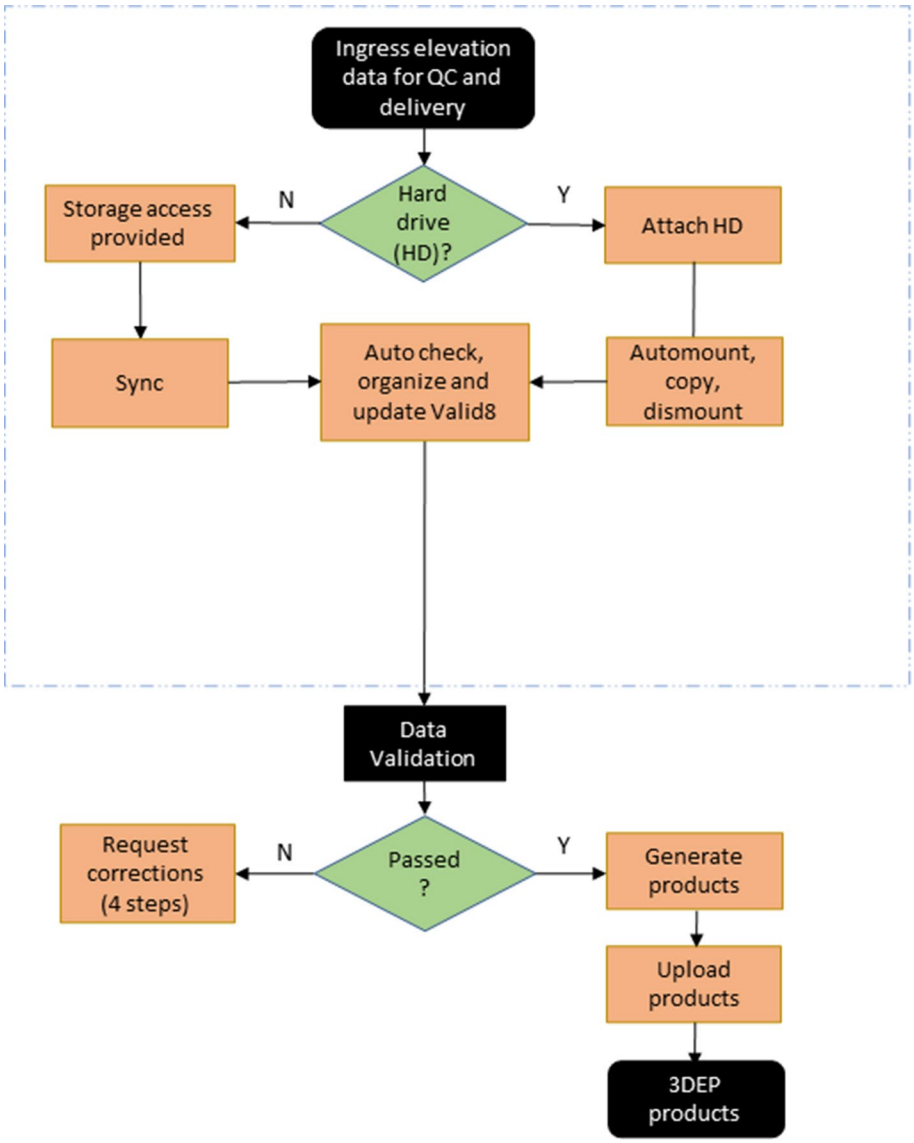

Fig. 5 3DEP's improved process workflow after applying the DMF approach. The blue box encloses the changed elements

  - Bounding Box
  - Polygon Footprint (more computationally complex)
  - Spatial Reference

- Accounting

  - Owner/POC
  - Where/Who did this data come from?

5. Total time for this movement is 2–3 working days, based completely upon physical movements by people.

Phase 2 Stage: Hard drive data at USGS

The DTS provides information about where the data are located and manually places labels on the hard drives.

Conveyance*: Copy to Server*

1. The data are at USGS, populated in DTS until someone can start the copy.
2. The POC and DTS staff exchange emails about the status, which prompts manual email generation to the Data Validation Team (DVT).
3. DVT connects the drive to a server and uses Robocopy to initiate the transfer.
4. The movement time depends on the bus speed and hardware.
5. Copying 2 to 10 terabytes of data, using USB3, at up to 5 Gbps (Gigabits per second) takes 1 to 2 days.
6. Total time: ~ 1 day to 2 weeks.

Stage: Hard Drive Data on Server

The DTS and emails list data and their locations.

Conveyance: Stage Data for Verification

1. An email from the POC prompts this movement.
2. The DVT manually compares data on disk to data on the server and populates the Work Unit information in Valid8.
3. The DVT organizes the file structure and manually moves data to the structure.
4. This movement may require an hour or two.

Phase 3 Stage: Hard Drive Data Verified Ready for QC on Server

Valid8 provides a record of data and locations.

Usage: Validate (QC) Source Data and Generate Output (Lev8)

1. This usage is initiated by the Valid8 status flag of "Ready for QC."
2. QC reads data from the provided locations over a 40-Gbps connection.
3. Once Valid8 notifies Lev8 that the source data are valid, Lev8 writes a transformed version of the output as final products over the same 40-Gbps connection.
4. This usage time is dominated by the interval required to validate the data and create the products, which is one to three years.

Phase 4 Stage: Public Data Generated Internally

Lev8 populates the internal database with finished products (elevation data products meeting 3DEP standards).

Conveyance: Upload to AWS

1. A flag in Lev8 starts the sync of final products to Amazon Web Services (AWS).
2. A return code is provided in a response string from the sync task.
3. rclone is used for content migration and management.
4. This movement is automated, and the duration depends on the size of the products.
5. Total Time: On average, on the 1-Gbps AWS connection, 300 GB requires 43 min.

Phase 5 Stage: Process Completion

Data are reviewed and published in the USGS ScienceBase data repository, rendering them publicly available.

Results: improved workflow

Based on the DMF analysis, we streamlined the 3DEP workflow by consolidating the number of stages and reducing manual conveyance triggers (Fig. 5). Furthermore, implementing this DMF has reduced the number of manual file conveyances and redundant data copies. The framework also played a pivotal role in identifying the need for a software tool to enforce the data management guidelines and drive additional efficiencies.

Results: new workflow tool

Ultimately, the DMF resulted in an improved workflow and was the impetus for this tool's development and functionality. A Data Orchestration Manager (ADOM) is a user interface designed to enforce data management goals by entwining and extending the open-source file management tool rclone (https://rclone.org/) (Fig. 6).

ADOM overview

The rclone tool provides data movement and management functionality in and between supported end-user-defined endpoints, also known as remotes. For this paper, these remotes can generally be considered nodes.

ADOM takes rclone a step further by wrapping rclone functionality into USGS-specific functions and methods. This allows for the usual and customary rclone features to

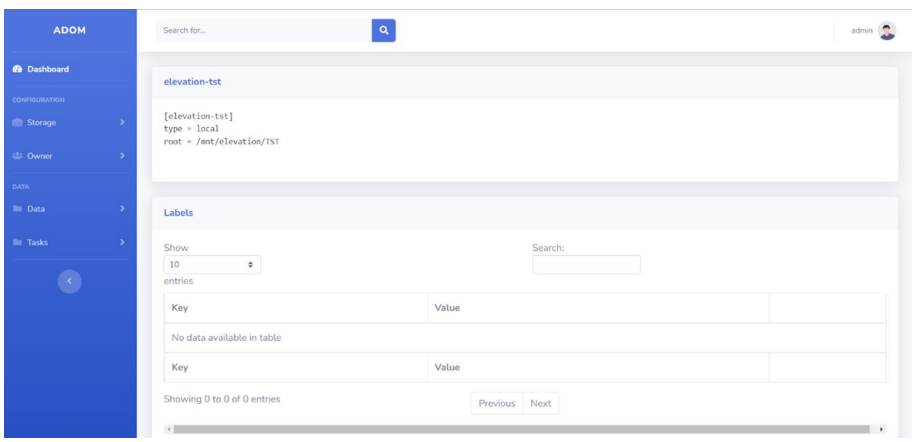

Fig. 6 ADOM user interface

be focused and guided by ADOM, which narrows the scope of functionality to maintain compliance with USGS data management standards.

In addition to maintaining compliance, the narrowing of scope provided by ADOM limits direct staff access to data while restricting data operations to only automated, standardized, and systemic actions. Source, destination, actions, and conveyance method are defined and fixed at the time of use. Further, control of the ADOM tool itself is constrained to automation calls through a representational state transfer application programming interface (REST API). This scope control removes the possibility of inconsistent manual interaction with large datasets and ensures all data operations are tightly controlled and logged. The tool also requires the existence or creation of a manifest with each data operation.

With the scope regulated by ADOM, the matters of compliance and management are easily standardized. For example, within a typical 3DEP workflow where uncompressed lidar data are delivered, compressed, and ingested for storage, ADOM may be used to read the compressed data from a common internet file system/server message block (CIFS/SMB) share on a lower performance commodity network attached storage (NAS) device before the QC review. ADOM can then egress the data, automatically writing it to high-performance storage before decompression. As described above, this task uses predefined "remote" endpoints and a predetermined programmatic method that is also accessed through ADOM's REST API. Then, when data are staged for the QC phase, the method can be called on the dataset, and the conveyance is launched. When completed, a result code for success or failure allows for either recording the successful move and subsequent post-processing or rescheduling the move until it succeeds.

ADOM function and impact

The current version of ADOM contains the following features and functionality and has been tested successfully:

1. Data Management Framework as input. This is required by default. If a manifest exists, it is verified. If a manifest does not exist, it is created and added to the dataset on disk. This provides the requisite data tracking and maintenance information.
2. Moves, copies, syncs, and validation of manifests. Standard features of rclone are supported, but only those listed are used now. Commands are sanitized via action profiles.
3. Predefined action profiles. These maintain compliance and consistency with workflow and data management standards. Actions are recorded in profiles that can be called with a movement method. Actions are fixed.
4. Predefined movement methods. Supported data movements predefine methods so that developers do not need to.
5. Web administration interface. The web administration interface provides for the setup of user and group access control, individual remote endpoints, and the assignment of rclone configurations by group.
6. A REST API. ADOM is intended to be used via automation only; the REST API is required. Building upon the success of the data management framework, which iden-

tified problematic stages and reduced inefficiencies, the use of ADOM in 3DEP can eliminate manual file movement errors. Redundant copies can be eliminated, and consistent and verified copies can now be ensured due to ADOM, which constrains the work scope and standardizes compliance and management.

Recommendations

Having successfully addressed the immediate needs of maintaining proper scope and compliance for large dataset movement and management for the USGS 3DEP project, ADOM can be extended to other USGS program areas and science centers as a solution for reliable, secure, and programmatic data movements. In addition, we have identified new specific internal use cases for ADOM and our DMF approach, including the performance of on-site and off-site workflow and enterprise data protection tasks in support of disaster recovery, records publishing and retention policies, and movement of large datasets to and from the USGS High-Performance Computing (HPC) platforms.

For example, the entire US Topo dataset (https://www.usgs.gov/programs/national-geospatial-program/us-topo-maps-america), which is published as a collection of the most current topographic map versions, is regularly requested by state and federal emergency management agencies. In addition, many entities request these data electronically, although some users prefer hard-drive media. Yet, the collection of the data for egress is performed by copying only the most current US Topo product from the published products, folders, or buckets. Today, this process proceeds manually. Using a manifest defining the most current US Topo dataset, a request can programmatically alert ADOM to convey the data to any data storage remote endpoint. When the process completes, a result code indicates success or failure to inform further actions.

Current efforts focus chiefly on tasks involving large amounts of data at rest. The enumeration and tracking of these data are paramount to defining the life cycle of USGS data deployed in their products and services today. However, the timely location and conveyance of data still present the principal data management challenges. Using manifests and automation addresses immediate risks and data management concerns, but scalability becomes a substantial concern as these datasets grow and new datasets are created. To this end, it may be beneficial to focus on data discoverability and increased conveyance speed.

As we concentrate on the accrescent size of the data we create and consume, there could be many data repositories, databases, and services ADOM can access and address, including enormous amounts of active and at-rest structured and unstructured data. Managing access and flow of data requires data namespace management and search features that function within both local and global scope.

Locally in scope, efforts to improve ADOM may emphasize documenting and simplifying management functions. In addition, the maintenance of ADOM endpoints, profiles, and methods could be standardized to reflect popular dataset endpoints, common data management profiles, and regularly used data conveyance methods. Most ADOM users could consume common source data endpoints, maintain common profiles, and operate the same data conveyance methods.

Other possible future efforts to expand ADOM, centering on the geospatial realm, may include the following:

1. allowing index linking to globally available namespace databases,
2. using the Resource Description Framework (RDF) data model relationships or SPARQL protocol and RDF query language (SPARQL) queries as dataset access methods,
3. developing linked indexes and data model relationships as query sources for geospatial features, allowing for novel semantic and ontological methods, and
4. establishing new methods of publishing, indexing, and searching for geospatially referenced data in general.

These efforts can provide novel benefits to the geospatial community by rendering data search and access simpler and more efficient.

More globally in scope, a service to register rclone data endpoints, similar to RDF/SPARQL endpoints, would be beneficial. Future global ADOM releases might include unused datatypes in domain name system (DNS) records to publish datasets, allow parsing, and render them discoverable. Additionally, ADOM functionality might include constraints on at-rest data so that no unauthorized processes or users can access or corrupt it. These improvements may revolutionize the way data are conveyed through processes.

Conclusions

Managing big geospatial data is complex, expensive, time-consuming, and often difficult to understand. Yet, as Li et al. [7] noted, data are a keystone national and scientific resource requiring a framework for proper management. This paper proposes a minimal standard for administering large data programs, emphasizing conveyance. First, to illustrate the merits of our solution, we expressed our ideas by describing a use case of 3D Elevation Program (3DEP) data processing at the USGS NGP/NGTOC. Next, we reviewed the challenges with the current 3DEP workflow and how our DMF approach, which focuses on data conveyance in phases from ingress to egress, addresses these challenges. Finally, we discussed how the ADOM tool enables standardized, efficient data management practices when moving large datasets. Overall, we found that a robust DMF focused on conveyance and ADOM's capabilities streamlined the 3DEP workflows and enforced order on a disordered process. Considering these findings, we plan to use this flexible approach and extendable tool to other big geospatial data management challenges at USGS. We also seek opportunities to collaborate with other entities to help solve similar data issues.

Although the research presented here is limited to the scope of data ingress and egress, data processing workflows, and data management policies at the USGS, additional opportunities to extend DMF and ADOM are substantial. In addition to the previously mentioned research possibilities, DMF and ADOM applicability to general big data management issues provides significant potential for future research. Some research examples include the creation of a DMF specification to extend domain search capabilities and ensure the uniqueness of data across storage domains, the development of alternate data conveyance methods unique to data type, network type, or storage type, and the construction and application of a standardized DMF and data conveyance method library to enable data management as code.

Abbreviations

3DEP	3D Elevation Program
5Ws1H	What, where, who, when, why, and how
ADOM	A data orchestration manager
API	Application programming interface
AWS	Amazon Web Services
CIFS	Common internet file system
CSDGM	Content Standard for Digital Geospatial Metadata
DMF	Data management framework
DataSHIELD	Data Aggregation Through Anonymous Summary-statistics from Harmonised Individual levEL Databases
DEM	Digital elevation model
DNS	Domain name system
DSM	Digital surface model
DTS	Drive tracking worksheet
DVT	Data validation team
FAIR	Findable, accessible, interoperable, and reusable
FGDC	Federal Geographic Data Committee
GB	Gigabyte
Gbps	Gigabits per second
GeoTIFF	Geographic tag image file format
GIS	Geographic information systems or science
HPC	High-performance computing
HTMC	USGS Historical Topographic Map Collection
ifsar	Interferometric Synthetic Aperture Radar
ISO	International Organization for Standardization
LAS	LPC file format
LAZ	Compressed LAS file
lidar	Light detection and ranging
LPC	Lidar point cloud
MB	Megabyte
NAS	Network attached storage
NGP	USGS National Geospatial Program
NGTOC	USGS National Geospatial Technical Operations Center
PDF	Portable document format
POC	Point of contact
OPR	Original product resolution
ORI	Orthorectified radar image
QA	Quality assurance
QC	Quality control
RDF	Resource description framework
REST API	Representational state transfer application programming interface
S3	Amazon Simple Storage Service
SFA	Scientific focus area
SHA-256	Secure hash algorithm
SMB	Server message block
SPARQL	SPARQL protocol and RDF query language
SRT	Shipping and receiving team
TB	Terabyte
US	United States
USB	Universal serial bus
USGS	U.S. Geological Survey

Acknowledgements
Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Author contributions
Conceptualization: STA, BBC, ANB, KGM, PTT. Methodology: STA, BBC, ANB, PTT. Project Administration: STA, BBC. Software: BBC, ANB. Validation: BBC, ANB, PTT. Visualization: STA, KGM, BBC, ANB, PTT. Writing—original draft, review, and editing: STA, KGM, BBC, ANB, PTT. All authors read and approved the final manuscript.

Funding
The USGS National Geospatial Program and National Science Foundation Grant No. 1853864 partly funded this research.

Availability of data and materials
The ADOM tool is available at https://code.chs.usgs.gov/elevsys/services/adom.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable (all data is freely available from public online sources).

Competing interests
No potential conflict of interest was reported by the authors.

Received: 30 January 2023 Accepted: 16 July 2023
Published online: 26 September 2023

References

1. Vert G, Stock M, Jankowski P, Gessler P. An architecture for the management of GIS data files. Trans GIS. 2002;6:259–75. https://doi.org/10.1111/1467-9671.00110.
2. Yang C, Huang Q, Li Z, Liu K, Hu F. Big Data and cloud computing: innovation opportunities and challenges. Int J Digit Earth. 2017;10:13–53. https://doi.org/10.1080/17538947.2016.1239771.
3. Bartoněk D. Solving big GIS projects on desktop computers. Kartogr i geoinformacije. 2019;18:44–62. https://doi.org/10.32909/kg.18.32.4.
4. Hariri RH, Fredericks EM, Bowers KM. Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data. 2019;6:44. https://doi.org/10.1186/s40537-019-0206-3.
5. U.S. Office of Management and Budget. Coordination of geographic information and related spatial data activities (Circular A-16 revised). 2002.
6. Varadharajan C, Faybishenko B, Henderson A, Henderson M, Hendrix VC, Hubbard SS, et al. Challenges in building an end-to-end system for acquisition, management, and integration of diverse data From sensor networks in watersheds: lessons from a mountainous community observatory in East River, Colorado. IEEE Access. 2019;7:182796–813. https://doi.org/10.1109/ACCESS.2019.2957793.
7. Li D, Gong Y, Tang G, Huang Q. Research and design of mineral resource management system based on big data and GIS technology. In: 2020 5th IEEE Int Conf Big Data Anal. IEEE; 2020. p. 52–6. https://doi.org/10.1109/ICBDA49040.2020.9101268.
8. Jankowski P. Integrating geographical information systems and multiple criteria decision-making methods. Int J Geogr Inf Syst. 1995;9:251–73. https://doi.org/10.1080/02693799508902036.
9. Airinei D, Homoncianu D. The architecture of a complex GIS & spreadsheet based DSS. J Appl Comput Sci Math. 2010;4:9–13.
10. Shah N, Agrawal S, Oza P. Data ingestion and analysis framework for geoscience data. Singapore: Springer; 2021. p. 809–20. https://doi.org/10.1007/978-981-15-8297-4_65.
11. Arenas D, Atkins J, Austin C, Beavan D, Egea AC, Carlysle-Davies S, et al. Design choices for productive, secure, data-intensive research at scale in the cloud. 2019; Available from: http://arxiv.org/abs/1908.08737
12. Kulawiak M, Kulawiak M, Lubniewski Z. Integration, processing and dissemination of LiDAR data in a 3D Web-GIS. ISPRS Int J Geo-Information. 2019;8:144. https://doi.org/10.3390/ijgi8030144.
13. Li Z, Hodgson ME, Li W. A general-purpose framework for parallel processing of large-scale LiDAR data. Int J Digit Earth. 2018;11:26–47. https://doi.org/10.1080/17538947.2016.1269842.
14. Kitchin R. Big Data, new epistemologies and paradigm shifts. Big Data Soc. 2014;1:205395171452848. https://doi.org/10.1177/2053951714528481.
15. Ferreira D, Vale M. Geography in the big data age: an overview of the historical resonance of current debates. Geogr Rev. 2022;112:250–66. https://doi.org/10.1080/00167428.2020.1832424.
16. Kitchin R. Big data and human geography: opportunities, challenges, and risks. Dialogues Hum Geogr. 2013;3:262–7. https://doi.org/10.1177/2043820613513388.
17. Bikakis N, Papastefanatos G, Papaemmanouil O. Big Data exploration, visualization and analytics. Big Data Res. 2019;18:100123. https://doi.org/10.1016/j.bdr.2019.100123.
18. Pencheva I, Esteve M, Mikhaylov SJ. Big Data and AI—a transformational shift for government: so, what next for research? Public Policy Adm. 2020;35:24–44. https://doi.org/10.1177/0952076718780537.
19. Wyly E. The new quantitative revolution. Dialogues Hum Geogr. 2014;4:26–38. https://doi.org/10.1177/2043820614525732.
20. Arribas-Bel D, Reades J. Geography and computers: past, present, and future. Geogr Compass. 2018;12:e12403. https://doi.org/10.1111/gec3.12403.
21. Goodchild MF. Geographical information science. Int J Geogr Inf Syst. 1992;6:31–45. https://doi.org/10.1080/02693799208901893.
22. Yano K. GIS and quantitative geography. GeoJournal. 2000;52:173–80. https://doi.org/10.1023/A:1014252827646.
23. Li X, Zheng D, Feng M, Chen F. Information geography: the information revolution reshapes geography. Sci China Earth Sci. 2022;65:379–82. https://doi.org/10.1007/s11430-021-9857-5.
24. Singleton A, Arribas-Bel D. Geographic data science. Geogr Anal. 2021;53:61–75. https://doi.org/10.1111/gean.12194.
25. Baumann P, Mazzetti P, Ungar J, Barbera R, Barboni D, Beccati A, et al. Big Data analytics for Earth Sciences: the EarthServer approach. Int J Digit Earth. 2016;9:3–29. https://doi.org/10.1080/17538947.2014.1003106.
26. Goodchild MF. The quality of big (geo)data. Dialogues Hum Geogr. 2013;3:280–4. https://doi.org/10.1177/2043820613513392.

27. Marr B. Big Data: using SMART big data, analytics and metrics to make better decisions and improve performance. USA: Wiley; 2015.
28. Madden S. From databases to Big Data. IEEE Internet Comput. 2012;16:4–6. https://doi.org/10.1109/MIC.2012.50.
29. Pasupuleti P, Salmone PB. Data lake development with big data: explore architectural approaches to building data lakes that ingest, index, manage, and analyze massive amounts of data using big data technologies. Birmingham, UK: Packt Publishing; 2015.
30. Li Z, Yang C, Jin B, Yu M, Liu K, Sun M, et al. Enabling Big Geoscience Data analytics with a Cloud-based, MapReduce-enabled and service-oriented workflow framework. PLoS ONE. 2015;10:e0116781. https://doi.org/10.1371/journal.pone.0116781.
31. Verma JP, Agrawal S. Big Data analytics: challenges and applications for text, audio, video, and social media data. Int J Soft Comput Artif Intell Appl. 2016;5:41–51.
32. Tsai C-W, Lai C-F, Chao H-C, Vasilakos AV. Big data analytics: a survey. J Big Data. 2015;2:21. https://doi.org/10.1186/s40537-015-0030-3.
33. Kappelman LA, Zachman JA. The enterprise and its architecture: ontology & challenges. J Comput Inf Syst. 2013;53:87–95. https://doi.org/10.1080/08874417.2013.11645654.
34. Sowa JF, Zachman JA. Extending and formalizing the framework for information systems architecture. IBM Syst J. 1992;31:590–616. https://doi.org/10.1147/sj.313.0590.
35. Babinski G. System Modelling for Effective GIS Management. Geogr Inf Sci Technol Body Knowl. 2018;2018. Available from: https://gistbok.ucgis.org/bok-topics/system-management.
36. Yao X, Mokbel M, Ye S, Li G, Alarabi L, Eldawy A, et al. LandQ^{v2}: a mapreduce-based system for processing arable land quality Big Data. ISPRS Int J Geo-Information. 2018;7:271. https://doi.org/10.3390/ijgi7070271.
37. Marchal L, Primet PV-B, Robert Y, Jingdi Zeng. Optimal bandwidth sharing in grid environments. In: 2006 15th IEEE Int Conf High Perform Distrib Comput. IEEE; 2006. p. 144–55. https://doi.org/10.1109/HPDC.2006.1652145.
38. Polverini M, Cianfrani A, Baiocchi A, Listanti M, Salvatore V. From raw data packets to ingress egress traffic matrix: the distributed MapReduce-based solution. NOMS 2018–2018 IEEE/IFIP Netw Oper Manag Symp. IEEE; 2018. p. 1–6. https://doi.org/10.1109/NOMS.2018.8406288.
39. Aljumaily H, Laefer DF, Cuadra D. Big-Data approach for three-dimensional building extraction from aerial laser scanning. J Comput Civ Eng. 2016;30. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000524.
40. Giachetta R. A framework for processing large scale geospatial and remote sensing data in MapReduce environment. Comput Graph. 2015;49:37–46. https://doi.org/10.1016/j.cag.2015.03.003.
41. Zhou S, Yang X, Li X, Matsui T, Liu S, Sun X-H, et al. A Hadoop-based visualization and diagnosis framework for earth science data. In: 2015 IEEE Int Conf Big Data (Big Data). 2015. p. 1911–6.
42. Li Z, Hu F, Schnase JL, Duffy DQ, Lee T, Bowen MK, et al. A spatiotemporal indexing approach for efficient processing of big array-based climate data with MapReduce. Int J Geogr Inf Sci. 2017;31:17–35. https://doi.org/10.1080/13658816.2015.1131830.
43. Gao S, Li L, Li W, Janowicz K, Zhang Y. Constructing gazetteers from volunteered Big Geo-Data based on Hadoop. Comput Environ Urban Syst. 2017;61:172–86. https://doi.org/10.1016/j.compenvurbsys.2014.02.004.
44. Mohammed EA, Far BH, Naugler C. Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends. BioData Min. 2014;7:22. https://doi.org/10.1186/1756-0381-7-22.
45. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Commun ACM. 2008;51:107–13. https://doi.org/10.1145/1327452.1327492.
46. Sverdlik Y. Google Dumps MapReduce in favor of new hyper-scale analytics system. Data Cent Knowl. 2014.
47. Varadharajan C, Hendrix VC, Christianson DS, Burrus M, Wong C, Hubbard SS, et al. BASIN-3D: a brokering framework to integrate diverse environmental data. Comput Geosci. 2022;159:105024. https://doi.org/10.1016/j.cageo.2021.105024.
48. Hagedorn S, Götze P, Sattler K-U. Big spatial data processing frameworks: Feature and performance evaluation. In: Proc 20th Int Conf Extending Database Technol. 2017. p. 490–3.
49. Budin-Ljøsne I, Burton P, Isaeva J, Gaye A, Turner A, Murtagh MJ, et al. DataSHIELD: an ethically robust solution to multiple-site individual-level data analysis. Public Health Genomics. 2015;18:87–96. https://doi.org/10.1159/000368959.
50. Wilson RC, Butters OW, Avraam D, Baker J, Tedds JA, Turner A, et al. DataSHIELD—new directions and dimensions. Data Sci J. 2017;16. https://doi.org/10.5334/dsj-2017-021.
51. Hare TM, Rossi AP, Frigeri A, Marmo C. Interoperability in planetary research for geospatial data analysis. Planet Space Sci. 2018;150:36–42. https://doi.org/10.1016/j.pss.2017.04.004.
52. Bolstad P. GIS Fundamentals: a first text on geographic information systems. 3rd ed. White Bear Lake: Eider Press; 2008.
53. Devarakonda R, Prakash G, Guntupally K, Kumar J. Big federal data centers implementing FAIR data principles: ARM Data Center example. In: 2019 IEEE Int Conf Big Data (Big Data) [Internet]. IEEE; 2019. p. 6033–6. https://doi.org/10.1109/BigData47090.2019.9006051.
54. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.
55. Michener WK. Ten simple rules for creating a good data management plan. PLOS Comput Biol. 2015;11:e1004525. https://doi.org/10.1371/journal.pcbi.1004525.
56. Wolf M, Logan J, Mehta K, Jacobson D, Cashman M, Walker AM, et al. Reusability first: Toward FAIR workflows. In: 2021 IEEE Int Conf Clust Comput [Internet]. IEEE; 2021. p. 444–55. https://doi.org/10.1109/Cluster48925.2021.00053.
57. Wayne L. ISO Geospatial Metadata: The 411 on 19115. North Rockies Chapter: URISA; 2015.
58. Chrisman N. Exploring geographic information systems. 2nd ed. New York: Wiley; 2002.

59. Yang J, Zhang Z, Zhang N, Li M, Zheng Y, Wang L, et al. Vehicle text data compression and transmission method based on maximum entropy neural network and optimized Huffman encoding algorithms. Complexity. 2019;2019:1–9. https://doi.org/10.1155/2019/8616215.
60. Geospatial Media and Communications. A Research Study To Understand Use Of USGS-NGP Geospatial Data and Products by Private Mapping Companies [Internet]. 2020. Available from: https://www.usgs.gov/media/files/use-ngp-geospatial-data-products-private-mapping-companies.

Publisher's Note

S. T. Arundel STA received her Ph.D. in Geography from Arizona State University in 2000. She was an assistant and then associate professor at Northern Arizona University until 2009 when she joined the U.S. Geological Survey, where she is a research geographer in the Center of Excellence for Geospatial Information Science. Her research emphasizes automating natural feature mapping and modeling using deep learning and other techniques and improving workflows by maximizing automation. She was the team lead in automating the 3D Elevation Program, which superseded the National Elevation Dataset.

K. G. McKeehan KGM received his Ph.D. in Geography from Michigan State University in 2022 and his Master of Science degree from the University of Wisconsin-Madison in 2018. In addition, KGM is a Research Physical Scientist and a Mendenhall Research Fellow at the USGS Center of Excellence in Geographic Information Science.

B. B. Campbell BBC is the Lead Program Analyst in the National Geospatial Technical Operations Center (NGTOC) at the U.S. Geological Survey. He researches and improves efficiency to automate computing, storage, and networking.

A. N. Bulen ANB is a GIS software developer for the USGS NGTOC in Rolla, Missouri, primarily focusing on developing software for processing 3D elevation data. ANB received his education in Computer Sciences from the University of Missouri, Rolla.

P. T. Thiem PTT is a developmental computer scientist in the Center of Excellence for Geospatial Information Science at the U.S. Geological Survey. His work deals with low-level and system organization issues, computational logistics, parallel processing, GIS processing, and infrastructure. He was formerly a senior developer for the U.S. Topographic Map Contours and the 3D Elevation Program.