Leveraging National Science Data Fabric Services to Train Data Scientists

Michela Taufer
University of Tennessee
Knoxville, USA
taufer@acm.org

Heberth Martinez
University of Tennessee
Knoxville, USA
hmarti46@utk.edu

Aashish Panta
University of Utah
Salt Lake City, USA
aashish.panta@utah.edu

Paula Olaya
University of Tennessee
Knoxville, USA
polaya@vols.utk.edu

Jack Marquez
University of Tennessee
Knoxville, USA
jmarque4@utk.edu

Amy Gooch
University of Utah
Salt Lake City, USA
amy.a.gooch@gmail.com

Giorgio Scorzelli
University of Utah
Salt Lake City, USA
scrgiorgio@gmail.com

Valerio Pascucci
University of Utah
Salt Lake City, USA
pascucci.valerio@gmail.com

***Abstract*—We document an interactive half-day tutorial in which participants explore the advanced applications of National Science Data Fabric (NSDF) services and strategies for comprehensive scientific data analysis. Targeting researchers, students, developers, and scientists, the tutorial provides valuable insights into managing and analyzing large datasets, particularly those exceeding 100TB. Participants gain hands-on experience by constructing modular workflows, leveraging public and private data storage and streaming solutions, and deploying sophisticated visualization and analysis dashboards. The tutorial emphasizes NSDF's role in supporting visualization conference themes by providing scalable visualization and visual analytics solutions. Our tutorial includes an overview of NSDF's capabilities, addressing common data analysis challenges, and intermediate hands-on exercises using NSDF services for Earth science data. Advanced applications cover handling and visualizing massive datasets requiring high-resolution data management. By the end of the session, attendees have a deeper understanding of integrating NSDF services into their research workflows, enhancing data accessibility, sharing, and collaborative scientific discovery. Our tutorial aims to advance knowledge in data-intensive computing and empower participants to harness the full potential of NSDF in their respective fields.**

***Index Terms*—Data analysis, data visualization, workforce development.**

I. Tutorial Structure

Scientific research often involves dealing with vast amounts of data stored in various public and private remote locations. Researchers frequently prefer to review all the available data remotely before deciding which segments to download, transferring only specific portions of this data to their local computer for closer analysis and visualization. However, every step of this process is challenging: it is difficult to stream the data, identify and deploy tools for data visualization, interact dynamically with the data, explore multiple datasets simultaneously, and decide which relevant data segment to download.

We present a tutorial [1] leveraging the National Science Data Fabric (NSDF) services [2]–[5] to improve how scientific data is accessed, analyzed, and visualized using cloud technologies. The tutorial targets scientists who need to visualize and analyze large scientific datasets interactively. Our tutorial demonstrates how NSDF services enable accessible, flexible, and customizable workflows for multi-faceted analysis and visualization of various datasets.

The tutorial walks through the workflow steps of generating large datasets through modular applications, storing this data remotely, and analyzing and visualizing the data locally to draw scientific conclusions. NSDF services allow users to stream data from public storage platforms like DataVerse [6] or private storage platforms like Seal Storage [7] and access an easy-to-use NSDF dashboard for immediate data interaction. We highlight how to navigate each step of the modular workflow, handle different data formats efficiently for streaming, and use visualization for scientific inference on selected data subsets. The tutorial demonstrates how NSDF services can facilitate fast remote access and gather fine-resolution data to stream for visualization and analysis using examples such as the earth science application SOMOSPIE [8]. SOMOSPIE (SOil MOisture SPatial Inference Engine) accesses, handles, and analyzes raw data from a public source such as the USDA portal [9] into terrain and soil moisture data for precision agriculture, wildfire prevention, and hydrological ecosystems.

II. Tutorial Goals, Structure, and Requirements

Our tutorial (https://github.com/nsdf-fabric/NSDF_Tutorial) [1] aligns directly with the NSDF community's emphasis on advanced computing and data management. Our tutorial offers strategies for handling and analyzing large volumes of data, addressing a common challenge faced by attendees involved in computational and data-intensive research. The target audience of the tutorial includes researchers, students, developers, and scientists. Researchers discover ways to streamline workflows for data-intensive research. Students are introduced to practical data management and analytical techniques within the research landscape. Developers learn to integrate NSDF services for software solutions. Scientists learn methods to enhance data

979-8-3503-5554-3/24/$31.00 ©2024 IEEE
DOI 10.1109/SCW63240.2024.00053

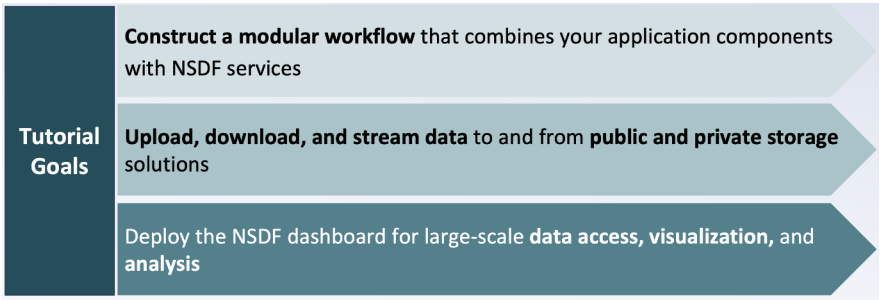


Fig. 1: Visualization of the overarching tutorial goals and objectives.

discovery, sharing, and collaborative efforts. The goals of the tutorial, as depicted in Figure 1, are as follows:

- Construct a modular workflow on top of NSDF: The tutorial demonstrates combining application components with NSDF (National Science Data Fabric) services to create a modular workflow. This goal helps streamline and optimize the management and analysis of scientific data.
- Upload, download, and stream data: Participants will learn how to effectively upload, download, and stream data to and from both public and private storage solutions. This goal emphasizes the importance of efficient data transfer and storage management in handling large datasets.
- Deploy NSDF services such as the NSDF-dashboard: The tutorial includes deploying the NSDF dashboard for large-scale data access, visualization, and analysis. This goal focuses on providing hands-on experience with tools that enable users to interactively work with extensive scientific datasets, enhancing their data analysis capabilities.

The tutorial is structured to cater to varying expertise levels: beginner, intermediate, and advanced. For the beginner level, which constitutes 30% of the content, the tutorial addresses scientists' data analytical needs, introduces data fabric concepts, and explores common data analysis challenges. The intermediate level, covering 40% of the content, involves analyzing practical data using NSDF services for an Earth science dataset. For the advanced level, which makes up 30% of the tutorial, the tutorial provides hands-on experience with complex datasets and demonstrates advanced applications of NSDF services.

Participants must have a foundational understanding of cloud-based storage systems, be familiar with various data formats and visualization tools, and have a GitHub account ready for practical sessions. The half-day tutorial is organized into three sessions. Session 1, lasting 30 minutes, begins with an overview of the NSDF and addresses users' challenges identified through interviews. Session 2, which is 1 hour long, offers a hands-on experience with NSDF services, focusing on visualization and dashboard creation for Earth science datasets [10], [11]. Session 3, lasting 30 minutes, concludes with an interactive Q&A, allowing attendees to discuss applications of NSDF in various research fields.

A continuous narrative is maintained across all sessions to maintain a coherent tutorial, including uniform slides and Jupyter Notebooks. A session chair facilitates each tutorial session, ensuring smooth transitions and guiding participants through the materials. This tutorial was held at the NSDF All Hands Meeting in San Diego with 35 attendees in person, at the University of Delaware with 12 attendees virtually via Zoom, and at the NSDF Webinar with 37 attendees virtually via Zoom.

III. NSDF Platform and Services

The National Science Data Fabric (NSDF) is an interconnected cyber-ecosystem poised to transform the landscape of data management and accessibility. NSDF commits to democratizing data delivery and catalyzing scientific discovery through collaboration with resource providers and users. Users can access NDSF computing, storage, and network services through its entry points, referring to the physical local nodes where a user or program begins data access and analysis [2].

The NSDF testbed integrates a suite of networking (both local and global) [12] [13], storage, and computing services; users access the services through NSDF's entry points across different providers. Figure 2 illustrates the structure of the NSDF's testbed. Entry points enable the interoperability of different applications and storage solutions, facilitating fast data transfer and caching among data sources, community repositories, and computing environments. Thus, they provide the foundation for the NSDF testbed and its services and are also the natural location for integrating FAIR Digital Objects in NSDF. Figure 2 shows the structure of the NSDF testbed with computing, networking, and storage services. For our tutorial, we use the NSDF Dashboard service. The environment we develop for the training builds on other NSDF services, including the NSDF-Plugin, NSDF-FUSE, and NSDF-Catalog services.

A. NSDF-Dashboard

The real-time analysis and visualization of large-scale data are crucial in today's data-driven world, enabling timely decision-making, enhancing the ability to detect and respond to

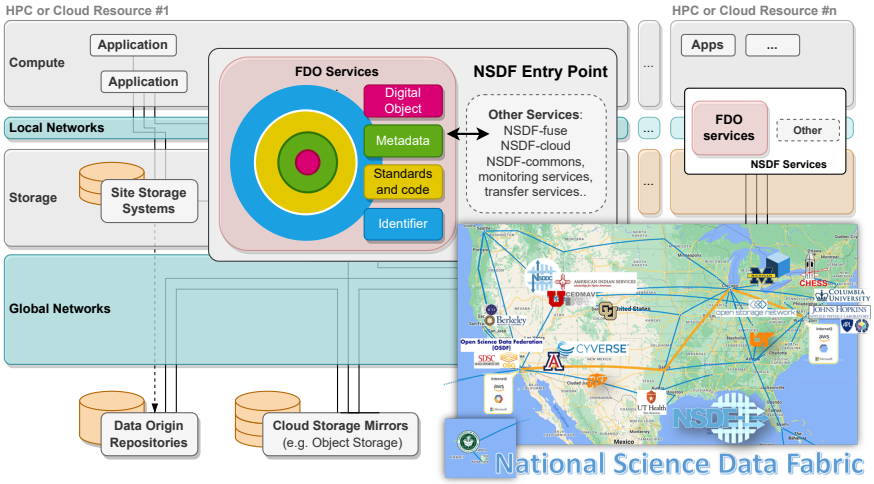


Fig. 2: Structure of the NSDF testbed (with computing, networking, and storage services).

emerging trends, and improving operational efficiency . However, several challenges often hamper the ability to perform on-the-fly analysis [14], [15]. Firstly, the extremely large volume of data generated exceeds the capacity of traditional data processing and storage systems, making it difficult to manage and store [16]. Secondly, limited access to high-performance computational resources restricts the ability to process and interpret large datasets quickly. Additionally, many organizations face bottlenecks due to inadequate infrastructure and outdated technologies not equipped to handle the demands of real-time data analytics [17]–[22]. Consequently, despite the potential benefits, the lack of necessary resources and infrastructure often results in missed opportunities for leveraging data to the fullest extent in real-time applications.

To address these challenges associated with on-the-fly visualization of large-scale data, we employ innovative solutions like the NSDF dashboards, which are based on the ViSUS software framework called OpenVisus [23] [24]. The OpenVisus framework allows the interactive exploration of massive scientific data on various hardware ranging from supercomputers to commodity hardware. OpenVisus allows advanced data reorganization using multi-resolution space-filling curves. A key strategy employed by this framework is the Hierarchical Z-Order (HZ-Order) indexing scheme [23]. The reorganization of the data allows efficient access at different resolution levels and ensures that spatially close data points are stored together. The reorganization maximizes the efficiency of modern data storage architectures, enabling swift access and manipulation. By arranging data hierarchically from coarse to fine resolutions, the OpenVisus framework facilitates the efficient real-time processing and access of these large datasets. The caching-enabled framework also allows users to extract any rectangular subsets of the input data progressively. The flexibility of OpenVisus is vital for managing computational resources effectively, enabling users to navigate through large datasets by adjusting the level of detail to match their immediate requirements or the constraints of their computing environment.

Another important aspect of the OpenVisus framework is the support for remote data streaming and out-of-core computations that allow OpenVisus to handle datasets too large to fit into primary memory by analyzing access patterns and reorganizing data for secondary storage. The approach ensures that data can be streamed efficiently, minimizing latency and overhead. By continuously analyzing how data is accessed, OpenVisus can dynamically update the data layout to prioritize frequently accessed data. OpenVisus also includes a storage-oblivious API that allows users to query specific data based on parameters such as region of interest, level of resolution, numerical precision, and amount of data. The API abstracts data storage and access complexities, providing a seamless interface for users to retrieve precisely the data they need. By allowing detailed retrieval queries, the API ensures that users can efficiently access relevant data without needing to understand the underlying storage mechanisms [25]. OpenVisus data management framework supports various industry-standard lossless and lossy compression algorithms such as ZIP, ZLIB, and ZFP with varying precision bits.

Our tutorial uses this dashboard framework to offer a robust array of features suitable for both casual explorers and serious researchers. Users can easily switch between datasets using the dropdown menu, allowing seamless transitions between variables within the dataset. The time slider is a critical tool for navigating through temporal data, enabling users to observe changes and trends over time. Additionally, the dashboard provides tools for taking horizontal and vertical slices of the data, which is beneficial for examining specific cross-sections of datasets. The snipping tool allows users to draw a rectangle within an image, showing a detailed view of the selected region and enabling the download of a NumPy array or a Python script for future data extraction.

To enhance data visualization, users can select from various color palettes, improving the interpretability of complex datasets. The colormap ranges can be manually adjusted or

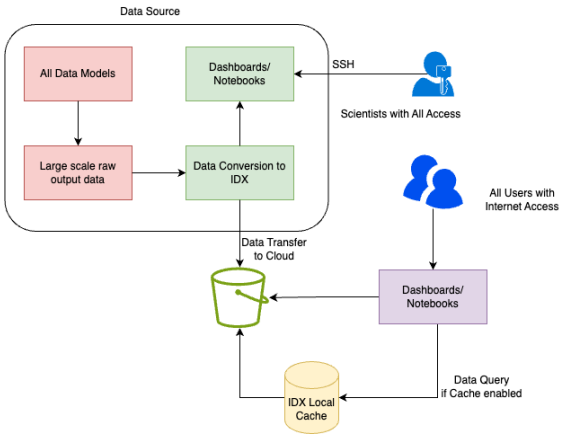


Fig. 3: An example workflow showing data conversion process from different environments. NSDF coordinates with various cloud providers and sources to make the data accessible to all users as required.

set dynamically, giving users control over the visualization's maximum and minimum values. The resolution sliders enable users to adjust the granularity of the data, facilitating both rapid overviews and detailed analyses. The playback functionality allows for automated data walkthroughs, offering a comprehensive view of climate evolution. The time speed control feature lets users adjust the pace of playback, providing flexibility in how data is reviewed over time.

B. Other NSDF Services

To handle the challenges of processing large volumes of data, NSDF integrates several services within its geo-distributed testbed. The NSDF-Plugin provides network monitoring and high-performance data transfer solutions to identify throughput and latency constraints across eight diverse locations in the United States, leveraging resources like Internet2 and Open Science Grid. NSDF-FUSE combines the flexibility of FUSE technology with the robustness of S3-compatible object storage. Through customizable mapping packages, users can seamlessly integrate and manage data across various environments. NSDF-FUSE service enhances data transfer efficiency and scalability, benefiting large-scale data analysis projects. Additionally, the NSDF-Catalog addresses the growing need for accessible scientific data by creating a centralized repository that indexes over 1.59 billion records, facilitating efficient data discovery and interdisciplinary collaboration. Together, these services provide a comprehensive infrastructure to support the scientific community in managing and processing data across distributed platforms.

IV. Using NSDF Services for Data Scientists

The four-step modular workflow of our tutorial, depicted in Figure 4, leverages NSDF services to analyze a geospatial dataset generated with GEOtiled [26].

- **Step 1: Data Generation.** Trainees collect Digital Elevation Models (DEMs) from the United States Geological Survey (USGS). They process the DEMs with GEOtiled or link to the data in either public or private storage.
- **Step 2: Conversion to IDX Format.** Trainees convert files from TIFF [27] to IDX format [28], which is used by OpenVisus. This step ensures accuracy is preserved while reducing file size. They store the IDX files in public or private storage.
- **Step 2: Step 3: Static Visualization.** Trainees statically visualize the terrain parameters using OpenVisus. They validate the accuracy of IDX-based images by comparing them with the original TIFF-based images.
- **Step 2: Step 4: Interactive Visualization & Analysis.** Trainees launch the dashboard for interacting with large-scale data, allowing users to access subregions of the original dataset for ad hoc analysis. They provide options to obtain and collect IDX files from either local storage or Seal Storage.

The tutorial can be applied to any application. The default training exercises focus on generating and analyzing high-resolution terrain parameters using the GEOtiled application [26]. Each workflow step is detailed, starting with data collection, then data conversion, static visualization for validation, and finally, interactive visualization for in-depth analysis.

A. Step 1: Data Generation

The first step of the tutorial involves generating data using GEOtiled. Before proceeding to the next step, there are two options for obtaining the data required to generate TIFF files. In the first option, data is generated using the SOMOSPIE application module. SOMOSPIE allows data creation from scratch within a scientific workflow, integrating the GEOtiled library to generate topographic parameters. In the second option, data is accessed from Dataverse public commons, which provides a secure and accessible environment for sharing scientific information publicly.

Terrain parameters, which describe surface form derived from Digital Elevation Models (DEM), can be applied in precision forestry, agriculture, and hydrology fields. In this tutorial, we use elevation, aspect, slope, and hillshading for the CONUS (Contiguous United States) dataset at a resolution of 30 meters. The computational expense of generating high-resolution terrain parameters often hinders their accessibility by the scientific community. We leverage one of the components of the SOMOSPIE (SOil MOisture SPatial Inference Engine) workflow [8], which empowers scientists to generate, predict, and analyze high-resolution topographic data. Figure 5 shows the entire workflow and the GEOtiled Terrain Generation component developed in the tutorial to generate data [26]. GEOtiled computes high-resolution terrain parameters using DEMs and leverages data partitioning to accelerate computation while preserving accuracy. The topographic data considered in this tutorial include elevation, aspect, slope, and hill shading.

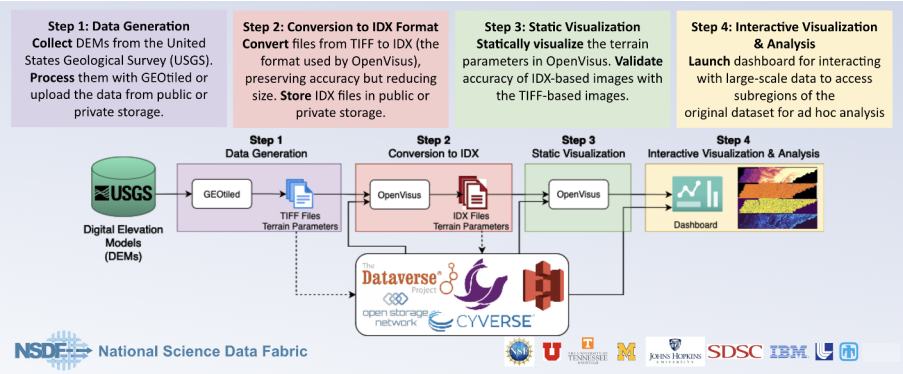


Fig. 4: Workflow depicted in four sequential steps, illustrating the process of data generation, conversion to IDX, static visualization and interactive visualization & analysis.

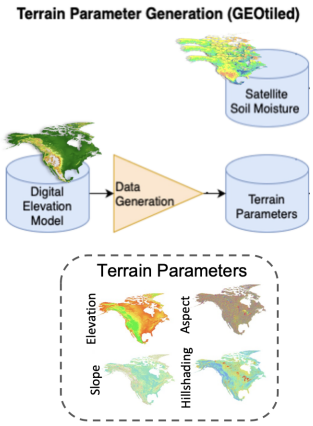


Fig. 5: Workflow illustrating the generation of terrain parameters by GEOTiled..

B. Step 2: Conversion to IDX Format

The second step of the tutorial involves converting the generated data from TIFF to IDX format using OpenVisus. This conversion is crucial for efficient data handling and visualization. OpenVisus is a progressive, cache-oblivious framework designed for large-scale data visualization. The IDX format stores data in a hierarchical Z (HZ) order, which provides efficient, progressive access to large-scale scientific datasets. Converting files from TIFF to IDX reduces file size by approximately 20% while preserving data accuracy. The IDX format offers several advantages:

- Efficient access as it provides progressive access to large datasets, enabling users to interact with the data meaningfully without requiring substantial computational resources;
- Scalability as it supports various running conditions, from personal computers to distributed systems;
- Versatility as the file conversion to IDX is not limited to TIFF; it supports other data formats such as NetCDF, HDF5, RGB, raw/binary, and more; and
- Compression as it supports industry-standard lossless and lossy compression algorithms such as zlib, zfp, and lz4.

The conversion process involves reading the TIFF files using Python functionalities and writing them in IDX format using the OpenVisusPy library. This step generates a metadata IDX file that provides a brief structural data layout, enabling progressive and cached data access that is beneficial for handling large datasets. Step 2 highlights the efficiency and practicality of using the IDX format for large-scale data visualization. By leveraging the capabilities of OpenVisus, we ensure that the converted data is organized, accessible, and ready for subsequent analysis and visualization stages in the tutorial.

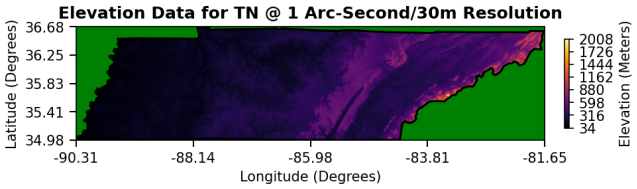
C. Step 3: Static Visualization

The third step of the tutorial involves the static visualization of terrain parameters using OpenVisus. This step is crucial for validating the accuracy of the converted data and ensuring that the data conversion process has preserved the integrity of the original data. In this step, participants statically visualize the terrain parameters (elevation, aspect, slope, and hillshading) using OpenVisus. The goal is to compare the visual representations of the original TIFF-based images with the converted IDX-based images to ensure accuracy and consistency. There are two options for obtaining and visualizing the data:

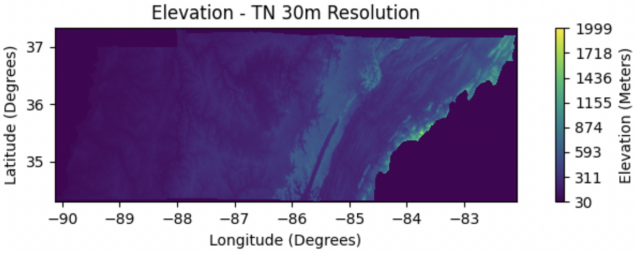
- **Option A: Local Storage** Load the dataset from local storage to perform the static visualization. This option allows participants to validate the data under controlled and known conditions
- **Option B: Seal Storage** Access the data from Seal Storage, a private storage service, to guarantee the use of validated and verified data.

The static visualization process involves loading the data into OpenVisus and comparing specific portions of the original and converted images using scientific metrics. This comparison ensures that the conversion to IDX format has not compromised the quality and accuracy of the data. Figure 6 illustrates the process of static visualization, highlighting the steps involved in validating the accuracy of the IDX-based images against

the original TIFF-based images. By completing this step, participants will gain confidence in the integrity of the converted data and be prepared to move on to the interactive visualization and analysis phase. This step underscores the importance of thorough validation in scientific workflows to ensure reliable and accurate data analysis.



(a) Visualization of TIFF file using GEOtiled.



(b) Visualization of the IDX file derived from TIFF conversion.

Fig. 6: Static visualization of terrain parameters using GEOtiled and OpenVisus.

D. Step 4: Interactive Visualization & Analysis

The fourth step of the tutorial involves interactive visualization and analysis of the terrain parameters using the NSDF dashboard. The fourth step allows participants to interact with large-scale data dynamically, enabling in-depth exploration and analysis. Participants launch the NSDF dashboard to interact with the high-resolution terrain data. The dashboard facilitates remote access to large datasets, allowing users to zoom into specific areas, select and crop subregions of interest, and save the data locally in a Python-compatible format for further analysis. There are two options for accessing and visualizing the data:

- Option A: Local Storage. The participants utilize the local storage to fetch the data, providing quick access for interactive exploration.
- Option B: Seal Storage. The participants access the data from Seal Storage, enabling the visualization of large datasets stored in the cloud without requiring local storage resources.

The interactive visualization process includes several key features:

- **Dynamic interaction** The participants zoom, pan, and select subregions within the dataset, providing a flexible and detailed exploration of the terrain parameters.
- **Geographical regions** The tutorial visualizes and analyzes two specific geographical regions: the State of Tennessee and the Contiguous United States (CONUS), both at a 30-meter resolution.
- **Ad-hoc analysis** The dashboard allows users to perform ad hoc analysis on selected subregions, facilitating detailed scientific discovery and insights.

Figure 7 showcases the NSDF dashboard in action, highlighting its capabilities for interactive visualization and analysis. Users can explore large-scale data efficiently, selecting only the subregions of interest for in-depth examination. This step emphasizes the power of interactive tools in managing and analyzing extensive datasets. By leveraging the NSDF dashboard, participants can visualize complex data structures, perform detailed analyses, and apply the insights to their specific research applications. The interactive capabilities provided by the NSDF dashboard are crucial for modern scientific workflows, enabling researchers to make data-driven decisions effectively.

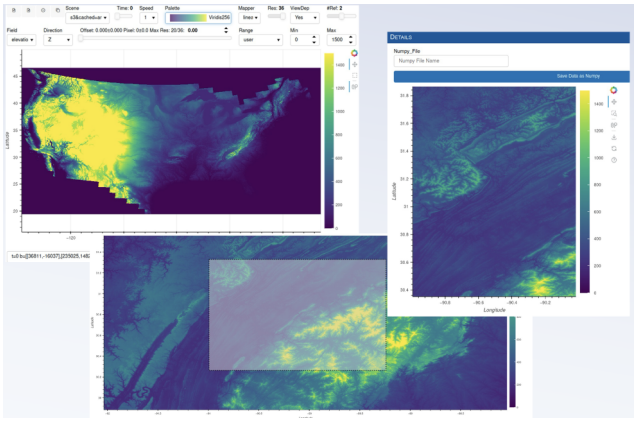


Fig. 7: Dashboard interface showcasing facilities and interactive visualization using OpenVisus to visualize large amount of data (CONUS).

E. Discussing the Tutorial Outcomes

The final part of the tutorial focuses on discussing the outcomes and answering participants' questions. This step consolidates the learning experience and ensures participants can apply the knowledge gained to their research needs. By the end of the tutorial, participants should be able to combine their application components with NSDF services to create a modular workflow that enhances data management and analysis. They should have acquired the skills to upload, download, and stream data to and from both public and private storage solutions, facilitating efficient data handling and accessibility. Participants should also be proficient in deploying the NSDF dashboard for large-scale data access, visualization, and analysis, enabling them to interact dynamically with extensive datasets. To ensure that these outcomes are met, the tutorial concludes with a discussion session where the following questions are posed to the participants:

- Application modularity: Can your application leverage APIs to integrate seamlessly with NSDF services?

- NSDF services: How can NSDF services enhance your data management and analysis processes?
- Data analysis practices: What type of analysis do you perform on your data? How can the NSDF dashboard support your analytical needs?
- Data characteristics: How large is your data? What methods do you use to access, share, and store your data?
- Storage solutions: Can your data benefit from utilizing both private and public storage solutions provided by NSDF?

V. Results

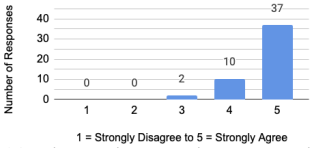
We conducted the tutorial on leveraging National Science Data Fabric (NSDF) services for large-scale scientific data analysis and visualization at various venues, engaging a diverse audience, including computer science experts, domain science experts, the general public, and students. Our sessions included the National Science Data Fabric All Hands Meeting at the San Diego Supercomputer Center, a research group at the University of Delaware, a public webinar, and a University of Tennessee Knoxville class. The total number of participants across these sessions was 108, with their professional backgrounds and modalities detailed in Table I.

A. Participant Feedback

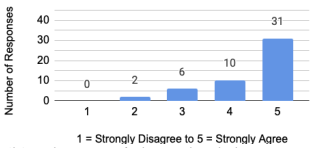
The feedback from the tutorial sessions was overwhelmingly positive. We gathered both qualitative and quantitative data through surveys, capturing the attendees' user experience and technology exposure. Participants across different demographics appreciated the clarity of the presentation and the ease of using the modular workflow. Comments from participants included:

- "The text was pretty clear, so I felt comfortable making decisions," said a domain scientist.
- Another domain scientist found the tutorial "excellent."
- Undergraduate students described the tutorial as "very easy to follow," "clear," and "very smooth and easy."

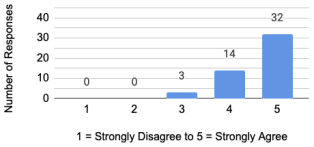
We collected data on various tutorial aspects, focusing on user experience and technology exposure. Key areas evaluated included the tutorial's difficulty level, whether it met attendees' expectations, the knowledge level of the presenters, the ease of following the tutorial, and the likelihood of recommending the tutorial to others. We also assessed the software stack's attainability, the dashboard's capability, the use case's impact, and the methodology's generality. From the user experience point of view, the tutorial was rated positively for its clarity and ease of use. Participants felt comfortable making decisions based on the modular workflow presented. Overall, the tutorial was considered easy to follow and understand. About the technology exposure, the methodology and tools presented were deemed effective for other datasets and study cases. The NSDF dashboard was recognized for enabling meaningful visualization and analysis. Overall, attendees appreciated the ability to upload, download, and stream data efficiently using NSDF services. The charts in Figure 8 illustrate selected survey questions and responses regarding user experience and technical exposure.



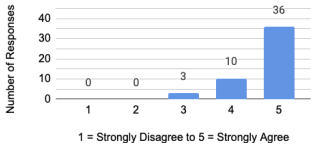
(a) The study case demonstrated the visualization and analysis capabilities of NSDF.



(b) The tutorial methodology can be generalized for other datasets and study cases.



(c) The dashboard enabled meaningful visualization and analysis.



(d) The workflow was easy to follow and understand.

Fig. 8: Questions from the tutorial survey that capture the user experience and technical exposure evaluation from the attendees across all the sessions.

B. Impact on Curriculum

The tutorial impacted the curriculum at the University of Tennessee Knoxville. Real-world applications were integrated into coursework using Jupyter Notebooks and newly developed software packages, bridging the gap between theoretical knowledge and practical application. Our effort enriched the learning experience for students, who gained hands-on data management and analysis skills. The successful integration of real-world applications into university coursework further underscores the tutorial's effectiveness in bridging the gap between theory and practice.

VI. Conclusion

Our paper presents a tutorial towards solving some of the challenges for data scientists dealing with large volumes of data. We created the tutorial to provide insights into managing and analyzing large datasets. Our paper presents how NSDF services address common pain points in data analysis for Earth science data and advanced applications, including handling and visualizing massive datasets in domains requiring high-resolution data management. Attendees of the documented tutorial sessions gained hands-on experience constructing modular workflows, leveraging public and private data storage and streaming solutions, and deploying visualization and analysis dashboards for scientific discovery. We present survey results of some tutorial sessions for different audiences that show the benefits they obtained by using our tutorial with the different tools that we presented. The outcome of our survey highlights how the tutorial successfully demonstrated using NSDF services for large-scale data analysis and visualization and receiving positive feedback from a diverse audience. Attendees acquired practical skills in constructing modular workflows, handling large datasets, and deploying visualization dashboards, enhancing their ability to perform complex data analyses in their respective fields.

TABLE I: Number of participants and their professional backgrounds across the tutorial presentations.

Tutorial	Modality	Audience	Number of participants
National Science Data Fabric All Hands Meeting, San Diego Supercomputer Center	In-person	Computer science experts	25
Research group, University of Delaware	Virtual	Domain science experts	15
National Science Data Fabric Webinar	Virtual	General public	36
Class at the University of Tennessee Knoxville (undergraduate and graduate students)	In-person	Undergraduate and graduate students	32
Total Participants			108

ACKNOWLEDGMENT

This research is supported by the National Science Foundation (NSF) awards #2138811, #2103845, #2334945, #2138296, and #2331152. The work presented in this paper was partly obtained using resources from ACCESS TG-CIS210128. We thank the Dataverse, Seal Storage, and Vargas Lab, led by Dr. Rodrigo Vargas.

REFERENCES

[1] H. Martinez, A. Panta, P. Olaya, G. Laboy, J. Ashworth, G. Scorzelly, J. Marquez, V. Pascucci, and M. Taufer, "Tutorial: Using NSDF for End-to-End Analysis of Scientific Data," https://github.com/nsdf-fabric/NSDF_Tutorial, Feb. 2024.

[2] J. Luettgau, H. Martinez, P. Olaya, G. Scorzelli, G. Tarcea, J. Lofstead, C. Kirkpatrick, V. Pascucci, and M. Taufer, "NSDF-services: Integrating Networking, Storage, and Computing Services into a Testbed for Democratization of Data Delivery," in *Proc. of the IEEE/ACM 16th International Conference on Utility and Cloud Computing*, 2023, pp. 1–10.

[3] P. Olaya, J. Luettgau, N. Zhou, J. Lofstead, G. Scorzelli, V. Pascucci, and M. Taufer, "NSDF-FUSE: A Testbed for Studying Object Storage via FUSE File Systems," in *Proc. of the 31st International Symposium on High-Performance Parallel and Distributed Computing*, 2022, pp. 277–278.

[4] J. Luettgau, G. Scorzelli, V. Pascucci, G. Tarcea, C. R. Kirkpatrick, and M. Taufer, "NSDF-Catalog: Lightweight Indexing Service for Democratizing Data Delivery," in *Proc. of the 2022 IEEE/ACM 15th International Conference on Utility and Cloud Computing (UCC)*, 2022, pp. 1–10.

[5] J. Luettgau, P. Olaya, N. Zhou, G. Scorzelli, V. Pascucci, and M. Taufer, "NSDF-Cloud: Enabling Ad-Hoc Compute Clusters Across Academic and Commercial Clouds," in *Proc. of the 31st International ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC)*, 2022, pp. 1–2.

[6] Dataverse Project, ""Dataverse: An Open Source Research Data Repository Platform"," https://dataverse.org, 2024, accessed: 2024-06-25.

[7] Seal, ""Seal Storage Solutions"," https://sealstorage.io/, accessed: 2024-06-25.

[8] D. Rorabaugh, M. Guevara, R. Llamas, J. Kitson, R. Vargas, and M. Taufer, "SOMOSPIE: A Modular SOil MOisture SPatial Inference Engine Based on Data-Driven Decisions," in *Proc. of the 2019 15th International Conference on eScience (eScience)*, 2019, pp. 1–10.

[9] United States Department of Agriculture, ""USDA Portal"," https://www.usda.gov, accessed: 2024-06-25.

[10] G. Tarcea, B. Puchala, T. Berman, G. Scorzelli, V. Pascucci, M. Taufer, and J. Allison, "The Materials Commons Data Repository," in *Proc. of the 18th IEEE International Conference on e-Science (eScience)*, 2022, pp. 1–2.

[11] R. M. Llamas, M. Guevara, D. Rorabaugh, M. Taufer, and R. Vargas, "Spatial Gap-Filling of ESA CCI Satellite-Derived Soil Moisture Based on Geostatistical Techniques and Multiple Regression," *Remote. Sens.*, vol. 12, no. 4, p. 665, 2020.

[12] J. Luettgau, H. Martinez, G. Tarcea, G. Scorzelli, V. Pascucci, and M. Taufer, "Studying Latency and Throughput Constraints for Geo-Distributed Data in the National Science Data Fabric," in *Proc. of the 32nd International Symposium on High-Performance Parallel and Distributed Computing*, 2023, p. 325–326.

[13] M. Taufer, H. Martinez, J. Luettgau, L. Whitnah, G. Scorzelli, P. Newell, A. Panta, P. Bremer, D. Fils, C. R. Kirkpatrick, and V. Pascucci, "Enhancing Scientific Research with FAIR Digital Objects in the National Science Data Fabric," *Computing in Science &; Engineering*, vol. 25, no. 05, pp. 39–47, 2023.

[14] N. Zhou, G. Scorzelli, J. Luettgau, R. R. Kancharla, J. Kane, R. Wheeler, B. Croom, P. Newell, V. Pascucci, and M. Taufer, "Orchestration of Materials Science Workflows for Heterogeneous Resources at Large Scale," *International Journal of High-Performance Computing Applications (IJHPCA)*, vol. 3-4, no. 37, pp. 260–271, 2023.

[15] R. M. Llamas, L. Valera, P. Olaya, M. Taufer, and R. Vargas, "Downscaling Satellite Soil Moisture Using a Modular Spatial Inference Framework," *Remote Sensing in Geology, Geomorphology and Hydrology*, vol. 14, no. 13, p. 3137, 2022.

[16] P. Olaya, D. Kennedy, R. Llamas, L. Valera, R. Vargas, J. Lofstead, and M. Taufer, "Building Trust in Earth Science Findings through Data Traceability and Results Explainability," *IEEE Trans. Parallel Distributed Syst. (TPDS)*, vol. 34, no. 2, pp. 704–717, 2023.

[17] J. Luettgau, G. Scorzelli, V. Pascucci, and M. Taufer, "Development of Large-Scale Scientific Cyberinfrastructure and the Growing Opportunity to Democratize Access to Platforms and Data," in *Proc. of the 25TH International Conference On Human-Computer Interaction (HCII)*, 2023.

[18] C. Roa, M. Rynge, P. Olaya, K. Vahi, T. Miller, J. Goodhue, J. Griffioen, D. Hudak, S. Knuth, R. Llamas, R. Vargas, M. Livny, E. Deelman, and M. Taufer, "End-to-end Integration of Scientific Workflows on Distributed Cyberinfrastructures: Challenges and Lessons Learned with an Earth Science Application," in *Proc. of the 15th IEEE/ACM International Conference on Utility and Cloud Computing (UCC)*. IEEE Computer Society, 2023, pp. 1–10.

[19] P. Olaya, J. Luettgau, C. Roa, R. Llamas, R. Vargas, S. Wen, I.-H. Chung, S. Seelam, Y. Park, J. Lofstead, and M. Taufer, "Enabling Scalability in the Cloud for Scientific Workflows: An Earth Science Use Case," in *Proc. of IEEE CLOUD*, 2023, pp. 1–10.

[20] D. Kennedy, P. Olaya, J. Lofstead, R. Vargas, and M. Taufer, "Augmenting Singularity to Generate Fine-grained Workflows, Record Trails, and Data Provenance," in *Proc. of the 18th IEEE International Conference on e-Science (eScience)*, 2022, pp. 1–2.

[21] T. Kitson, P. Olaya, E. Racca, M. R. W. II, M. Guevara, R. Vargas, and M. Taufer, "Data Analytics for Modeling Soil Moisture Patterns across United States Ecoclimatic Domains," in *Proc. of the 2017 IEEE International Conference on Big Data (BigData)*, 2017, pp. 4768–4770.

[22] R. McKenna, V. K. Pallipuram, R. Vargas, and M. Taufer, "From HPC Performance to Climate Modeling: Transforming Methods for HPC Predictions into Models of Extreme Climate Conditions," in *Proc. of the 11th IEEE International Conference on eScience (eScience)*, 2015, pp. 108–117.

[23] V. Pascucci, G. Scorzelli, B. Summa, P.-T. Bremer, A. Gyulassy, C. Christensen, S. Philip, and S. Kumar, "The visus visualization framework," in *High Performance Visualization*, 2012, pp. 439–452.

[24] B. Summa, G. Scorzelli, M. Jiang, P.-T. Bremer, and V. Pascucci, "Interactive Editing of Massive Imagery Made Simple: Turning Atlanta into Atlantis," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 2, pp. 1–13, 2011.

[25] A. Panta, X. Huang, N. McCurdy, D. Ellsworth, A. Gooch, G. Scorzelli, H. Torres, P. Klein, G. Ovando-Montejo, and V. Pascucci, "Web-based visualization and analytics of petascale data: Equity as a tide that lifts all boats," 2024. [Online]. Available: https://arxiv.org/abs/2408.11831

[26] C. Roa, P. Olaya, R. Llamas, R. Vargas, and M. Taufer, "GEOtiled: A Scalable Workflow for Generating Large Datasets of High-Resolution Terrain Parameters," in *Proc. of the 32nd International Symposium on High-Performance Parallel and Distributed Computing*, 2023, pp. 311–312.

[27] ""TIFF (Tagged Image File Format)"," https://www.loc.gov/preservation/digital/formats/fdd/fdd000022.shtml, 2024, accessed: 2024-06-25.

[28] S. Kumar, V. Pascucci, V. Vishwanath, P. Carns, M. Hereld, R. Latham, T. Peterka, M. E. Papka, and R. Ross, "Towards Parallel Access of Multi-Dimensional, Multi-Resolution Scientific Data," in *Proc. of the 2010 5th Petascale Data Storage Workshop (PDSW)*, 2010, pp. 1–5.