DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Reference herein to any social initiative (including but not limited to Diversity, Equity, and Inclusion (DEI); Community Benefits Plans (CBP); Justice 40; etc.) is made by the Author independent of any current requirement by the United States Government and does not constitute or imply endorsement, recommendation, or support by the United States Government or any agency thereof.

ORNL/TM-2024/3669 PUBID 225425

VA EDH Advanced Software Pipeline Framework Report: Enhancing Automation, Reproducibility, and Scalability

Hilda B. Klasky
Josh Grant
Midgie MacFarland
Heidi Hanson
Jodie Trafton
Anuj Kapadia

March 2025

OAK RIDGE National Laboratory

ORNL IS MANAGED BY UT-BATTELLE LLC FOR THE US DEPARTMENT OF ENERGY

ORNL/TM-2024/3699 PUBID 225425

Computational Sciences & Engineering Division

VA EDH ADVANCED SOFTWARE PIPELINE FRAMEWORK REPORT: ENHANCING AUTOMATION, REPRODUCIBILITY, AND SCALABILITY

Hilda B. Klasky
Josh Grant
Midgie MacFarland
Heidi Hanson
Jodie Trafton
Anuj Kapadia

March 2025

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831
managed by
UT-BATTELLE LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

ABBREVIATIONS AND ACRONYMS

Acronym	Definition
ORNL	Oak Ridge National Laboratory
AHRQ	Agency for Healthcare Research and Quality
API	Application Programming Interface
ADDRESS Tool	Automated Determination of Detailed Regional and Exact Spatial Segments Address to Census Level Lookup Tool
BISL	Business Intelligence Solutions Laboratory
CDW	Clinical Data Warehouse
CI/CD	Continuous Integration/Continuous Deployment
EDH	Environmental Determinants of Health
FIPS	Federal Information Processing Standards
FY25Q1	Fiscal Year 2025, First Quarter
GDPR	General Data Protection Regulation
GEO ID	Geographical Identifier
HERMIT	**H**armonized **E**ntity **R**esource **M**etadata **I**ntegration **T**ool
HPC	High-Performance Computing
HIPAA	Health Insurance Portability and Accountability Act
MS SQL Server	Microsoft Structured Query Language Server
MINERVA	Multivariate Information Exploration Repository for Veteran Analysis a Geospatial decision support tool.
ORNL	Oak Ridge National Laboratory
PostGIS	Geospatial Database Extension for PostgreSQL
RBAC	Role-Based Access Control
SQL	Structured Query Language
SSL	Secure Sockets Layer
VA	Veterans Affairs

ABSTRACT

The VA Environmental Determinants of Health (EDH) Advanced Software Pipeline Framework is designed to enhance the efficiency, scalability, and security of geospatial data processing workflows. This framework integrates modern data orchestration and containerization technologies, including Prefect for workflow automation, Docker for containerization, and PostgreSQL/PostGIS for geospatial data storage and analysis. It ensures standardized, reproducible, and automated data processing, supporting VA objectives related to substance use risk assessment and recovery research.

The pipeline addresses key scalability and performance challenges through horizontal and vertical scaling, high-performance computing (HPC) integration, parallel processing, task caching, and dynamic resource allocation. These optimizations improve throughput and reduce latency, allowing the system to efficiently manage large and complex datasets. Additionally, security and compliance measures—such as data encryption (SSL), Role-Based Access Control (RBAC), and adherence to GDPR and HIPAA standards—safeguard sensitive information throughout data transmission and storage.

A key implementation of this framework includes the automation of shelter list geolocation workflows, ensuring that up-to-date data is readily available for VA decision-making. Lessons learned from this project include the transition from in-memory processing to incremental storage writes, improving resource management and reliability. Future enhancements aim to expand automation, integrate AI-driven anomaly detection, and incorporate high-performance computing resources. This framework provides a scalable, secure, and adaptable solution for managing geospatial datasets, reinforcing the VA's ability to support clinical and strategic initiatives through data-driven decision-making.

1. INTRODUCTION

The Veterans Affairs (VA) Environmental Determinants of Health (EDH) Advanced Software Pipeline Framework builds on earlier work[1] to enhance the efficiency, scalability, and reliability of geospatial data processing workflows. This report provides an update on the new pipeline framework developed by the ORNL team.

The VA-EDH Advanced Data Pipeline leverages ORNL's C-HER Ecosystem Architecture for Environmental Determinants of Health to create automated, reproducible, and scalable workflows that address challenges in data processing, Continuous Integration/Continuous Deployment (CI/CD), and machine learning operations. By integrating advanced technologies such as Docker [2] for containerization, Prefect[3] for orchestration, and PostgreSQL/PostGIS[4, 5] or geospatial data storage and analysis, the framework ensures consistent data handling, reduces manual intervention, and enhances system reliability. It also supports dataset imports into an MS SQL Server database [6], aligning with the sponsor's operational environment, where customizations ensure data curation documentation meets specific requirements.

The pipeline framework covers the entire geospatial data management lifecycle—including data ingestion, transformation, validation, storage, and orchestration. Additionally, it incorporates geolocation processing through the Address to Census Level Lookup Tool (ADDRESS Tool), which automatically determines Census geographies from addresses or latitude/longitude coordinates.

The primary objectives of the pipeline are to automate workflows, enhance performance through scalable components, and ensure reproducibility via version control. The following sections describe how this framework supports the sponsor's objectives, its architecture, and ongoing developments.

1.1 OBJECTIVES AND SPONSOR REQUIREMENTS

In response to the sponsor's need for data-driven computational methods to characterize U.S. communities based on substance use risk and recovery factors, this framework:

- **Standardizes Datasets:**
  Ensures all community factors are standardized to a common spatial extent (e.g., U.S. Census Tract or County) for seamless integration with VA geospatial data and alignment with the AHRQ Social Determinants of Health project.
- **Curates Data with Provenance:**
  Uses repeatable, transparent methodologies—supported by configuration files (for APIs, netCDF, CSV, etc.) and versioned Docker containers—to ensure data provenance and reproducibility.
- **Integrates with VA Systems:**
  Enables linkage between social and environmental determinants of health datasets and VA patient and facility locations using GEO IDs such as FIPS. These datasets are transmitted to the VA Clinical Data Warehouse (CDW) and Business Intelligence Solutions Laboratory (BISL) via MS SQL Server, supporting both clinical efforts and strategic planning.

- **Provides Guidance and Code Samples:**
  We plan to provide detailed guidance, recommendations, and sample code to assist in the interpretation and integration of curated data into existing VA data platforms.

2. PIPELINE ARCHITECTURE AND COMPONENTS

Figure 1presents the planned architecture diagram that describes how the VAEDH pipeline framework leveraging the C-HER Ecosystem Architecture for Environmental Determinants of Health. The description of this architecture follows. At the time this report was published most components have been implemented.

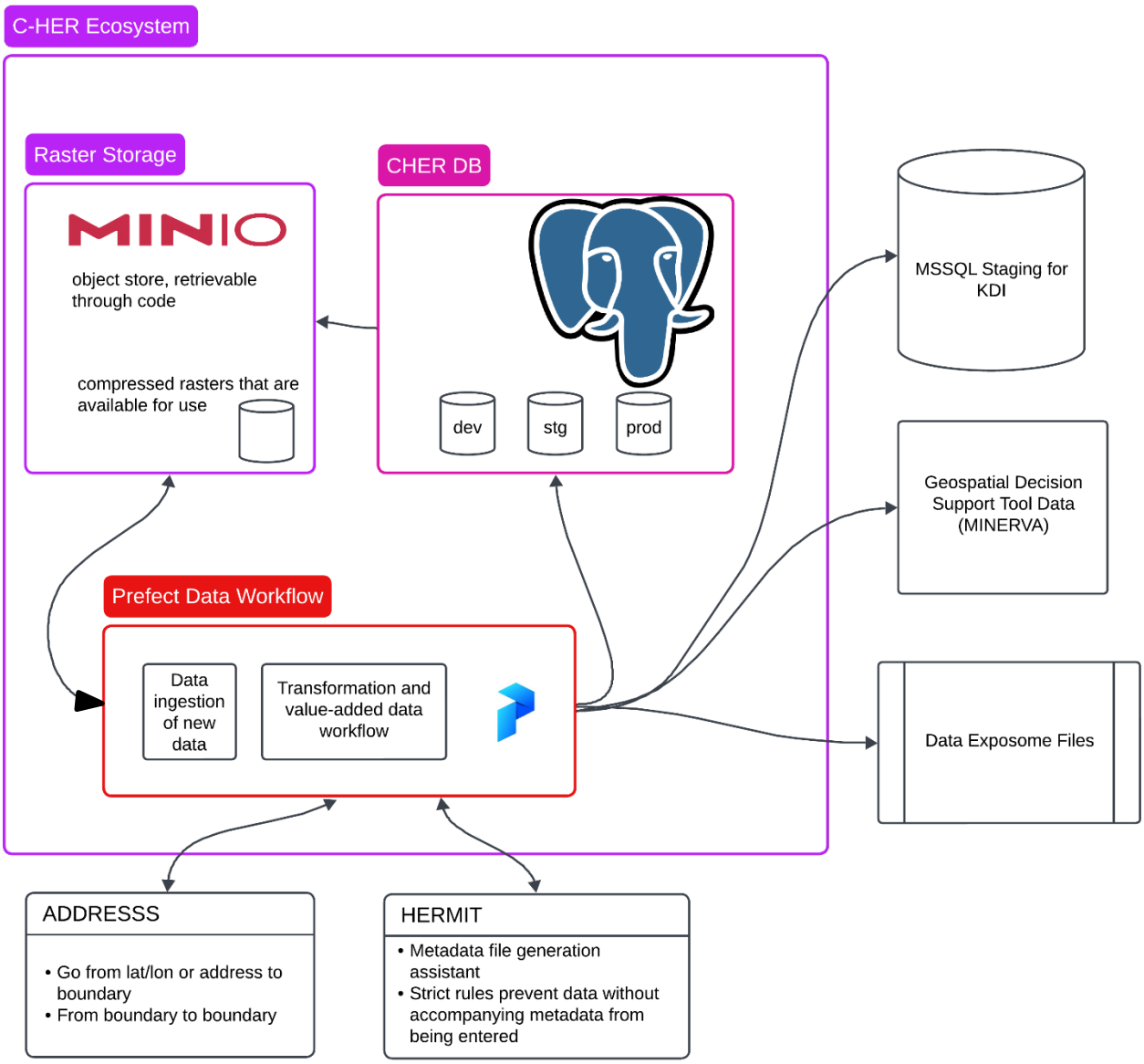

Figure 1. VA-EDH Advanced Data Pipeline: Leveraging C-HER Ecosystem Architecture for Environmental Determinants of Health.

2.1 DATA SOURCES, STORAGE, AND PREFECT WORKFLOW

In the architecture depicted in Figure 1, at the upper left, MINIO[7] serves as the object store for compressed raster datasets. These datasets are retrievable on demand, which supports efficient geospatial analyses that depend on large or frequently updated raster data. Data flows from MINIO and other sources into the Prefect[3] Data Workflow, which automates data ingestion, transformation, and quality assurance (QA)/quality control (QC) processes. Prefect's hybrid execution model allows tasks to be executed locally, in the cloud, or across hybrid environments, ensuring that the pipeline remains flexible and capable of handling new data ingestions or updates with minimal manual intervention. Prefect will be described in depth in the following sections.

The software pipeline is designed to manage the entire geospatial data lifecycle which consists of data ingestion, data transformation and data validation. We describe these stages in the following paragraphs:

Data Ingestion:
Raw geospatial data is extracted from various government agencies (for example, the Centers for Disease Control and Prevention (CDC)) as well as other external sources. Initial validations, including QA and QC checks, are performed to verify geographic identifiers. (Refer to our sponsor reports for a complete listing [8-22]).

Data Transformation:
Once ingested, data proceeds through transformation stage. Spatial transformations—including aggregations and reprojections—are performed using PostGIS functions. Data transformations also include the modifications to format the data to follow our repository standards.

Data Validation:
Rigorous validation checks (such as missing data checks, spatial integrity checks, projection and coordinate reference system validation and attribute consistency) are applied to ensure that the datasets maintain data integrity and consistency as they progress through the pipeline.

Once validated, data undergoes a series of spatial operations—such as transformation and aggregation—using PostGIS functions. These operations prepare the data for efficient querying and subsequent analysis. This structured approach ensures that geospatial data is processed in a robust, automated, and scalable manner, aligning with the overall objectives of the VA-EDH Advanced Software Pipeline Framework.

2.2 GEOLOCATION WITH THE ADDRESS TOOL

The pipeline includes a custom developed Address to Census Level Lookup Tool (ADDRESS Tool) that converts addresses or latitude/longitude coordinates into standardized Census geographies. This geolocation process ensures that each address is accurately mapped to the corresponding administrative boundary, which is critical for regional analysis. This component can also handle boundary-to-boundary lookups, making it indispensable for refining administrative boundaries or enriching datasets with precise location information. Prefect orchestrates these tasks in parallel when possible, reducing processing time for large-scale datasets.

2.3 DATA STORAGE WITH POSTGRESQL/POSTGIS AND MINIO

Validated data is stored in a hybrid storage architecture that combines PostgreSQL/PostGIS and MinIO (as shown at the center of the diagram), which will eventually support three distinct environments—development (dev), staging (stg), and production (prod). This environment separation will ensure robust testing and deployment in alignment with Continuous Integration/Continuous Deployment (CI/CD) principles. The EDH Spatial Data Pipeline relies on PostgreSQL, an open-source relational database management system renowned for its reliability, extensibility, and robust SQL support, as the backbone for structured data storage. PostGIS extends PostgreSQL with native support for geographic objects, enabling efficient querying and manipulation of geospatial data.

In addition to PostgreSQL/PostGIS, MinIO serves as the primary object storage solution within the EDH pipeline, offering scalable and high-performance storage for raster datasets, non-tabular data formats, and raw, untransformed data. MinIO, an S3-compatible object storage system, is particularly well-suited for handling large-scale spatial data such as satellite imagery, LiDAR point clouds, and environmental sensor outputs. By integrating MinIO, the pipeline accommodates semi-structured and unstructured datasets that do not conform to the rigid schema of relational databases, enabling a more flexible and scalable data architecture.

Within this pipeline:

- PostgreSQL/PostGIS is used for structured, tabular, and relational geospatial data, ensuring ACID compliance (Atomicity, Consistency, Isolation, Durability) for reliable data management. PostGIS provides powerful spatial indexing and query optimizations for vector-based datasets, facilitating efficient spatial analysis and geoprocessing.
- MinIO is leveraged for high-volume, unstructured, and semi-structured geospatial data, including raster imagery (GeoTIFFs, NetCDF, etc.), raw untransformed data, and unprocessed files. The object storage paradigm allows for scalable storage and efficient retrieval of large datasets, supporting direct access through APIs and integration with geospatial processing tools.

The synergy between PostgreSQL/PostGIS and MinIO ensures that structured and unstructured geospatial data are efficiently managed and accessible within the EDH pipeline. This hybrid approach eliminates the need for excessive data transformation before storage, preserving raw data integrity while still enabling optimized querying and analysis.

2.3.1 Postgres and PostGIS Databases and Multiple Postgres Database Environments - Sponsor Alignment and Downstream Systems

On the right side of the diagram, three key destinations—**MSSQL Staging for KDI**, **Geospatial Decision Support Tool Data**, and **Data Exposome Files**—illustrate how curated data is made available for various sponsor use cases.

The pipeline's ability to import datasets into Microsoft SQL Server (MSSQL) ensures alignment with sponsor-specific requirements, allowing smooth integration into their existing operational environment.

Prefect delivers the processed data to the MSSQL staging database, where it is stored before being moved to its final destination for further analysis for use within the VA's CDW system. Datasets are imported from PostgreSQL into an MS SQL Server database, aligning with the sponsor's operational environment. Customizations are implemented to ensure the data curation documentation meets the specific requirements of our VA sponsors. This document: *Dataset Repository for Investigating Suicide Risk Using Social and Environmental Determinants of Health - Manuscript under editorial consideration* [20] outlines how the VA datasets are further customized and imported into the VA's CDW for their use.

Minerva is a key component in our data processing system, acting as the central repository where processed data is stored and accessed. Our Prefect Workflow orchestrator plays a crucial role in creating

the data required for Minerva by automating and managing various steps in the data pipeline. The Prefect Workflow ensures that data is properly processed, transformed, and validated before it is passed to Minerva.

Meanwhile, geospatial decision support tools will leverage the refined data to inform strategic planning, and the Prefect Data Workflow will also allow the creation of Data Exposome Files supports broader analytics related to environmental determinants of health.

This orchestrated process ensures that data flows seamlessly from collection through transformation to storage, enhancing both the efficiency and reliability of the entire system.

2.3.2 Expanding Storage Capabilities with MinIO

The inclusion of MinIO within the EDH Spatial Data Pipeline expands the system's ability to store and manage diverse geospatial datasets beyond what relational databases traditionally support. Raster and non-tabular data—such as drone imagery, climate model outputs, and elevation grids—can be stored in MinIO without the constraints of a structured schema, allowing for more dynamic and scalable storage.

MinIO's compatibility with AWS S3 APIs enables seamless integration with cloud-native geospatial processing tools such as GDAL, Rasterio, and Python processing tools., facilitating direct access to object storage without requiring costly data movement. Additionally, MinIO supports **versioning and immutability**, ensuring that raw datasets remain unchanged while transformed versions can be stored alongside them, enabling reproducible workflows in data science and environmental modeling.

Key advantages of MinIO in the EDH pipeline include:

- Efficient handling of large geospatial datasets that exceed the practical limits of relational databases.
- Direct integration with machine learning and big data frameworks, supporting analytics workflows that require access to raw, high-resolution datasets.
- Optimized retrieval mechanisms for unstructured data through parallel processing and object-based storage architecture.
- Scalability and fault tolerance, ensuring high availability and durability for mission-critical geospatial datasets.

By leveraging both PostgreSQL/PostGIS for structured data and MinIO for unstructured data, the EDH pipeline is positioned to handle the increasing complexity of spatial data ecosystems, ensuring high-performance data access, efficient storage, and robust analytical capabilities.

2.4 ORCHESTRATION WITH PREFECT

Data orchestration is managed by Prefect[23], an open-source workflow orchestration platform built with Python. Prefect's key features include:

- **Hybrid Execution Model:**
  Prefect separates the control plane (which manages workflows) from the execution plane (where tasks are run), allowing tasks to be executed locally, on-premises, or in the cloud with full observability.
- **Automation and Error Handling:**
  Automated task retries and dynamic mapping for parallel processing reduce the need for manual intervention. Prefect's Python-native interface ensures workflows are both readable and maintainable.
- **Version Control and Reproducibility:**
  Integration with Git ensures that all workflow modifications are version-controlled, supporting reproducibility and collaboration among team members

Data orchestration[23] is the process of managing and automating data-driven workflows, ensuring that data moves seamlessly between different systems while maintaining data quality, consistency, and reliability. It plays a pivotal role in modern data engineering by coordinating the complex series of tasks involved in extracting, transforming, and loading (ETL) data, enabling seamless interaction between data sources, processing engines, and storage solutions. Effective data orchestration provides an abstraction layer that hides the complexity of managing dependencies, monitoring data flows, and handling failures, making it easier to develop and maintain data pipelines.

Prefect is the core orchestration tool used in the VA-EDH Spatial Data Pipeline. Prefect is an open-source, next-generation workflow orchestration platform that offers a robust, flexible solution for managing data workflows. It is built to address common issues with data pipelines, such as failed tasks, complex dependencies, and dynamic scheduling. Prefect provides an intuitive framework to build, monitor, and orchestrate complex workflows, while maintaining full observability and control of the data processes.

One of the standout features of Prefect is its "hybrid execution model[3]". Unlike traditional orchestrators, Prefect separates the control plane, which manages the workflow, from the execution plane, where tasks are actually performed. This means users can run workflows wherever they want—on-premises, in a cloud environment, or across a hybrid infrastructure—without sacrificing observability or control. Prefect flows are defined using Python, which makes it easy for data engineers to create workflows as code. By providing native Pythonic constructs, Prefect allows for more readable and maintainable workflows, thus simplifying complex ETL processes.

Prefect's emphasis on resilience and reliability is also key for the VA-EDH project. The platform supports task retries, enabling failed tasks to be automatically re-executed, which is essential in environments where network issues or other transient errors may occur. Prefect's dynamic mapping functionality allows for parallelization, which is extremely useful in large-scale geospatial data processing, where tasks like data ingestion, transformation, and spatial operations can benefit from concurrent execution to improve performance. Prefect's observability features, such as task-level logging and a visual interface for monitoring pipeline states, provide a clear view of the data flows and help pinpoint and resolve issues more quickly.

The Prefect orchestration engine significantly contributes to the scalability of the VA-EDH Spatial Data Pipeline. It allows the pipeline to handle diverse workloads, ranging from routine data updates to more intensive geospatial analysis tasks. Prefect's compatibility with various cloud and on-premises compute environments makes it suitable for the distributed and containerized nature of the pipeline, which relies on Docker.

2.5 CONTAINERIZATION WITH DOCKER

Docker is utilized to containerize every component of the pipeline—including Prefect flows, PostgreSQL/PostGIS instances, and supporting microservices. This containerization guarantees a consistent environment across development, testing, and production, while also enabling scalable and independent deployment of individual pipeline components.

The VA-EDH project employs Docker to containerize all components of the pipeline, including Prefect flows, PostgreSQL/PostGIS instances, and supporting services. Docker's containerization allows each component to run in an isolated environment, ensuring consistency across development, testing, and production environments. Containers encapsulate all dependencies, which makes deploying the pipeline as a set of reproducible, portable microservices easier. This modularity enhances the scalability and flexibility of the overall architecture, as each container can be independently scaled, maintained, or replaced without disrupting the entire system.
Microservices architecture [24], facilitated by Docker, allows each part of the EDH Spatial Data Pipeline to handle a specific functionality—such as orchestration, data storage, or monitoring—which can be developed and deployed independently. This modular approach reduces the coupling between components, making the entire system more resilient and easier to update or extend with new features.

2.6 DATA FLOW, COMMUNICATION, AND SCHEDULING

Data flows seamlessly through the pipeline stages either sequentially or concurrently, depending on processing requirements. Key elements include:

- **Dynamic Mapping and Parallel Processing:**
  Prefect’s dynamic mapping feature allows for concurrent task execution, which is essential for processing large-scale geospatial datasets efficiently.
- **Incremental Data Writes:**
  By incorporating incremental writes to storage, the pipeline avoids excessive memory usage, thus handling larger datasets without overwhelming system resources.
- **Scheduling and Orchestration:**
  Workflow stages can be triggered based on time-based schedules, event-driven triggers, or manual intervention, offering flexibility to adapt to varying data loads and requirements.

2.7 SCALABILITY, PERFORMANCE, AND SECURITY

Efficient data processing frameworks must be designed to handle increasing workloads while maintaining performance and security. The VA EDH Advanced Software Pipeline Framework incorporates **scalability, performance optimization, and security best practices** to ensure reliability in diverse computational environments. This section outlines the framework’s ability to **scale efficiently**, optimize

processing performance, and **secure sensitive data**, making it well-suited for handling large-scale geospatial datasets and complex analytical workflows.

2.7.1 Scalability and Performance

- **Scalability:**
  The framework supports both horizontal scaling (adding more containers) and vertical scaling (increasing resource allocation) to handle varying workloads. It also accommodates high-performance computing (HPC) environments when necessary.
- **Performance Optimization:**
  Techniques such as parallel processing, task caching (to reuse results and avoid redundant processing), and dynamic resource allocation are employed to optimize throughput and reduce latency.

2.7.2 Security and Compliance

Data security is paramount. The pipeline employs encryption for data both in transit (using Secure Sockets Layer (SSL) handshakes) and at rest. Strict access controls, including Role-Based Access Control (RBAC), ensure that data handling complies with regulatory standards such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA).

3. CASE STUDY

One application of the pipeline is the crowdsourcing of shelter lists from an online source, first implemented in the first quarter of Fiscal Year 2025 (FY25Q1)[16]. The pipeline was used to collect shelter information, which was then processed using the Address to Census Level Lookup Tool (ADDRESS) to geolocate addresses. The ADDRESS tool converted the addresses into latitude and longitude coordinates and provided the corresponding Census tract, resulting in a geolocated dataset of shelters.

This workflow is now fully automated and can be executed at regular intervals to ensure that the VA has access to the most up-to-date shelter data. Since FY25Q1, most datasets have been processed through this pipeline. However, some datasets were provided directly to the sponsor without pipeline processing due to data-sharing restrictions from the source providers.

4. CHALLENGES AND LESSONS LEARNED

The implementation of the EDH Spatial Data Pipeline faced challenges in integrating various technologies and managing system resources. Earlier versions of the pipeline relied on an in-memory mechanism for data delivery, which became unsustainable when datasets exceeded system memory. To address this, incremental writes to storage were introduced, improving resource management and enabling the pipeline to handle larger datasets more efficiently. Additionally, modularizing components using Docker enhanced maintainability, while Prefect's dynamic mapping significantly improved processing efficiency. These lessons have guided best practices and will inform future pipeline enhancements.

5. CONCLUSION

The VA EDH Advanced Software Pipeline Framework meets sponsor requirements by providing a robust, data-driven solution that characterizes U.S. communities through automated, standardized, and reproducible workflows. By integrating key technologies—Docker for containerization, Prefect for orchestration, PostgreSQL/PostGIS for geospatial data storage, MinIO for raw and unstructured data storage, and the ADDRESS Tool for geolocation—the framework enhances clinical decision-making and strategic planning within VA systems.

Future improvements may focus on increasing automation of QA/QC procedures, integrating high-performance computing (HPC) resources, leveraging AI for advanced analytics, and implementing anomaly detection for proactive monitoring. Ultimately, this framework delivers a scalable, secure, and flexible solution for high-quality data management, benefiting both the VA and other government agencies.

6. REFERENCES

[1] H. B. Klasky, Whitehead, Matthew, Hamaker, Alec, Johnson, Evelyn, and Kapadia, Anuj., "VA-Environmental Determinants of Health's Software Pipeline Framework.," 2023. [Online]. Available: https://www.osti.gov/biblio/1999090

[2] "Docker Inc., "Docker Overview," [Online]. ." https://www.docker.com/ (accessed November, 2022.

[3] Prefect. "Prefect Hybrid Execution Model." https://www.prefect.io/ (accessed 2025).

[4] P. G. D. Group. "PostgreSQL." https://www.postgresql.org/ (accessed 2024).

[5] P. Project. "PostGIS." https://postgis.net/ (accessed 2024).

[6] M. Corporation. "Microsoft SQL Server." https://www.microsoft.com/en-us/sql-server (accessed 2024).

[7] "MinIO, "MinIO Documentation," [Online]." https://min.io (accessed November, 2024.

[8] B. Christian *et al.*, "VA EDH Data Curation Documentation – FY21, Rev. 2, ORNL/SPR-2021/2366," Oak Ridge National Laboratory, United States, 2022. [Online]. Available: https://www.osti.gov/biblio/1854468-va-edh-data-curation-documentation-fy21-rev

[9] B. Christian *et al.*, "VA EDH Data Curation Documentation FY22-Q1, Rev. 2, ORNL/SPR-2022/2316," Oak Ridge National Laboratory, United States, 12 2021. [Online]. Available: https://www.osti.gov/biblio/1854460-va-edh-data-curation-documentation-fy22-q1-rev

[10] B. Christian *et al.*, "VA EDH Data Curation Documentation FY22-Q2, Rev. 2, ORNL/SPR-2022/2391," Oak Ridge National Laboratory, United States, 3 2022. [Online]. Available: https://www.osti.gov/biblio/1862127-va-edh-data-curation-documentation-fy22-q2-rev

[11] H. Klasky, Sparks, K., Logan, J., Hamaker, A., Whitehead, M., Peluso, A., Hanson, H., Watson, R., & Kapadia, A.,, "VA EDH Data Curation Documentation FY23-Q1, ORNL/SPR-2022/2694," Oak Ridge National Laboratory, United States, 2022. [Online]. Available: https://doi.org/10.2172/1909101

[12] H. Klasky, Sparks, K., Peluso, A., K., Logan, J., Hamaker, A., McGee, M., VanDerslice, J., Hanson, H., Watson, R., and Kapadia, A.,, "VA EDH Data Curation Documentation FY23-Q3 ORNL/SPR-2023/2930 PUB ID 195499," Oak Ridge National Laboratory, United States, 2023.

[13] H. Klasky, Sparks, Kevin, Peluso, Alina, Myers, Aaron, Hamaker, Alec, McGee, Michael, Zhang, Jonathan, Logan, Jeremy, Hanson, Heidi, Watson, Rochelle, and Kapadia, Anuj., "VA EDH Data Curation Documentation FY23-Q4. ORNL/SPR- 2023/3097. PUB ID 202517," 2023. [Online]. Available: https://www.osti.gov/biblio/2204567

[14] H. Klasky, K. Sparks, and J. Logan, Tuccillo, Joe, Whitehead, Matthew, Hamaker, Alec, Hanson, Heidi, Watson, Rochelle, and Kapadia, Anuj., "VA EDH Data Curation Documentation - FY22-Q3, ORNL/SPR-2022/2487," Oak Ridge National Laboratory, United States, 2022. [Online]. Available: https://www.osti.gov/biblio/1876283-va-edh-data-curation-documentation-fy22-q3

[15] H. Klasky *et al.*, "VA EDH Data Curation Documentation FY24-Q4," Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), United States, 2024. [Online]. Available: https://www.osti.gov/biblio/2472692
https://www.osti.gov/servlets/purl/2472692

[16] H. Klasky *et al.*, "VA EDH Data Curation Documentation FY25-Q1," Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), United States, 2024. [Online]. Available: https://www.osti.gov/biblio/2502173
https://www.osti.gov/servlets/purl/2502173

[17] H. Klasky *et al.*, "VA EDH Data Curation Documentation (FY24-Q2)," Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), United States, 2024. [Online]. Available: https://www.osti.gov/biblio/2341397
https://www.osti.gov/servlets/purl/2341397
[18] H. Klasky *et al.*, "VA EDH Data Curation Documentation (FY24-Q1)," Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), United States, 2023. [Online]. Available: https://www.osti.gov/biblio/2229216
https://www.osti.gov/servlets/purl/2229216
[19] H. Klasky *et al.*, "VA EDH Data Curation Documentation FY23-Q2 ORNL/SPR-2023/2857," Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), United States, 2023.
[20] H. B. Klasky, Hanson, H., Sparks, K., Whitehead, M., Blair, C., Trafton, J.A., and Kapadia, A.,, "Dataset Repository for Investigating Suicide Risk Using Social and Environmental Determinants of Health - Manuscript under editorial consideration," 2022. [Online]. Available: https://www.osti.gov/biblio/1999090
[21] H. B. Klasky *et al.*, "VA EDH Data Curation Documentation FY22-Q4, ORNL/SPR-2022/2587," Oak Ridge National Laboratory (ORNL), Oak Ridge, TN United States, 2022.
[22] H. Klasky *et al.*, "VA EDH Data Curation Documentation (FY24-Q3)," Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), United States, 2024. [Online]. Available: https://www.osti.gov/biblio/2404618
https://www.osti.gov/servlets/purl/2404618
[23] "Prefect, “Prefect Documentation,” [Online]." https://docs.prefect.io/v3/get-started/index (accessed November, 2024.
[24] "Microservice Architecture." https://microservices.io/ (accessed November, 2022.).