VIFI: Virtual Information-Fabric Infrastructure for Data-Driven Decisions from Distributed Data (NSF DIBBs Award #1640818)

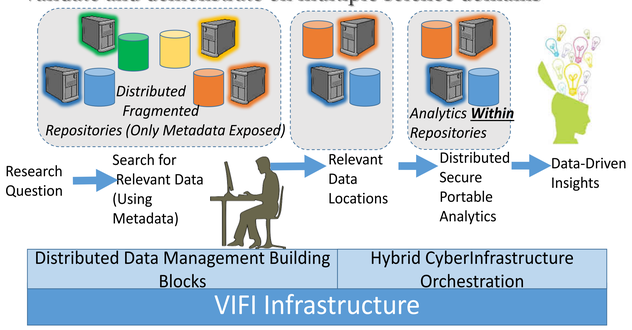
Ashit Talukder, Principal Investigator

W. Dou, Y Tao, Y. Zhu, G. Djorgovski, A. Mahabal, D. Crichton, W. Tolone, M. Hadzikadic, E. El-Shaer, Y. Wang, W. Zadrozny

UNC Charlotte, California Inst. Of Technology, JPL, LSU, Nova Southern U.

GOALS AND OBJECTIVES

- Novel cyberinfrastructure that facilitates data-driven discovery from distributed, fragmented datasets
  - without requiring movement of massive amounts of data
  - without exposing sensitive raw datasets to end users.
- Virtual Information-Fabric Infrastructure (VIFI) allows scientists to search, access, manipulate, and evaluate fragmented, distributed data in the information 'fabric' without directly accessing or moving large amounts of data.
- Validate and demonstrate on multiple science domains



NEW CAPABILITIES

- Distributed analytics at source (vs. moving data to analytics)
- Data-driven innovation by making previously un-shareable data more accessible to data scientists
  - without direct access to rich, diverse datasets
  - while mitigating data owners concerns by not exposing raw data directly to data consumers, and
  - while limiting data movement across networks

Earth Science:

- Ability to integrate reduced simulations with observational measurements at distributed earth science data centers.
- Balance uncertainty associated with scientific inferences from data reductions at distributed sites (simulation & observed measurements).

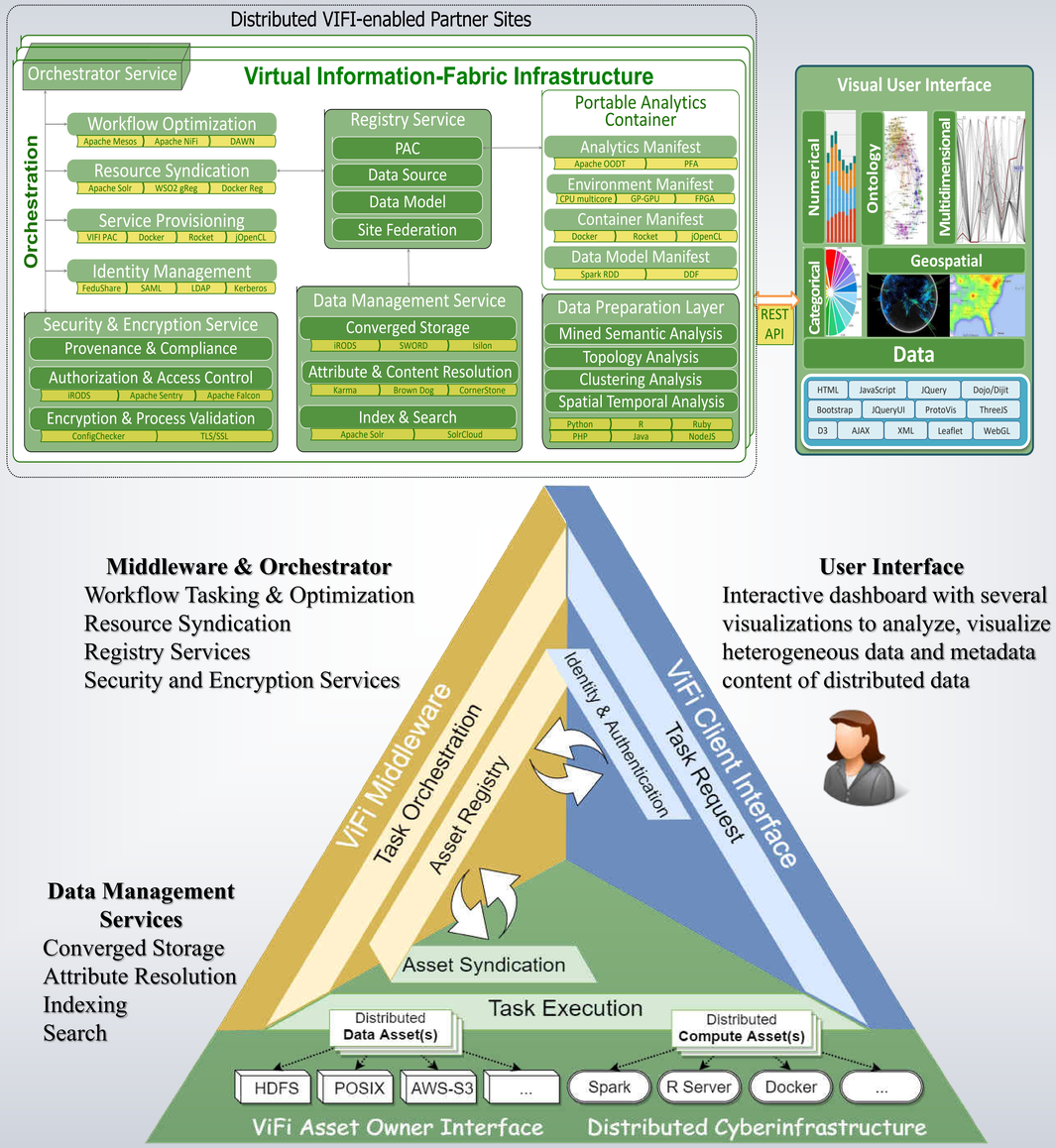
Sustainable Resilient Human-Building Ecosystems (SRHBE):

- Significantly change the current SRHBE analytic processes in two aspects, i.e., data readiness and service availability,
- Support many multidisciplinary processes including life cycle assessment, energy simulation, multi-hazard mitigation modeling and social/policy modeling.

Astronomy:

- Combine models and light-curve features en masse from multiple sources for surveys without necessitating massive data transfers.
- Allow communication of astronomical data in a transparent manner along with follow-up reassembly of results for multiple disparate sky-surveys and datasets in real-time.

FRAMEWORK AND APPROACH



MILESTONES

Year 1: Initial version of VIFI with core functionalities, demonstrated on data subsets

Prototype VIFI framework to explore possible system architectures

Evaluate the simulated precipitation using observations for Earth Science.

Establish set of surveys and extracted features to be used with VIFI in Astronomy

Finalize implementation requirements for four SRHBE cases; Start recruiting potential users

Years 2-4: VIFI with enhanced security, encryption, distributed data analysis functionalities

Demonstrations on full precipitation data (Earth Science), Identify and confirm rare objects (Astronomy) and expand to additional astronomy data, conduct SRHBE outreach and testing on 4+ SRHBE cases

EXPECTED SCIENCE ADVANCES AND IMPACTS

- Data-intensive scientific analysis at scale over distributed data
- Improved data-driven research outcomes by virtually, seamlessly tying distributed un-shareable data together & ease burden on data scientists

Multidisciplinary Impacts Across Domains

Earth Science:

- More granular (regional-scale), data-rich climate observations
- Ability to generate significantly higher resolution climate models
- Improved understanding in rainfall over the CONUS

Astronomy:

- Enable Citizen-science and crowd-sourcing for wider data access without exchanging high volumes of data.
- Inclusion of formerly sparse and seemingly insignificant data will be possible and aid transient science in a significant manner

SRHBE:

- Overcome multiple data fragmentation problems in research:
  - spatial fragmentation (SF), temporal fragmentation (TF), spatial temporal fragmentation (STF) and data requirements fragmentation (DRF),
  - increase consistency, reliability and accuracy in decision-making by utilizing the processes supported by VIFI.
- Support emergent applications
  - Integrated resilience
  - Sustainability analysis and scaling modeling.

SUSTAINABILITY PLAN

- Open source infrastructure available to all stakeholders
- Improve adoption by incorporating into UNC Data Science Initiative and workshops in Data Analytics conference series
- Integrate into existing system capabilities for climate research from DOE (e.g., Earth System Grid) and NASA (Earth Observing System).
- User Community will establish the appropriate consensus mechanisms to determine VIFI strategy/standards
- VIFI will track and periodically report the number, volume, and types of datasets ingested.
- Metrics will be created to reflect compatibility of datasets for combined tasks and comparative assessments

ACKNOWLEDGMENTS

Funding for this research was provided by the National Science Foundation (NSF) Data Infrastructure Building Blocks (DIBBs) Program under Award #1640818 .