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

Award #: 1640818

PIs & Institutions: UNC Charlotte (William J. Tolone – Lead PI, Hadzikadic, Wang, Zadrozny, Dou, El-Shambakey, Das Bhattacharjee), **Caltech** (Djorgovski, Mahabal), **Jet Propulsion Lab** (Crichton, Lee, Braverman, Johnson), **North Carolina A&T** (Cho), and the **SHBE Team** (Zhu, Tao) comprised of **LSU** (CMU, FSU, Lawrence Berkeley National Lab) and **Cleveland State University** (UNT)

NSF CSSI PI Meeting, Seattle, WA, Feb. 13-14, 2020

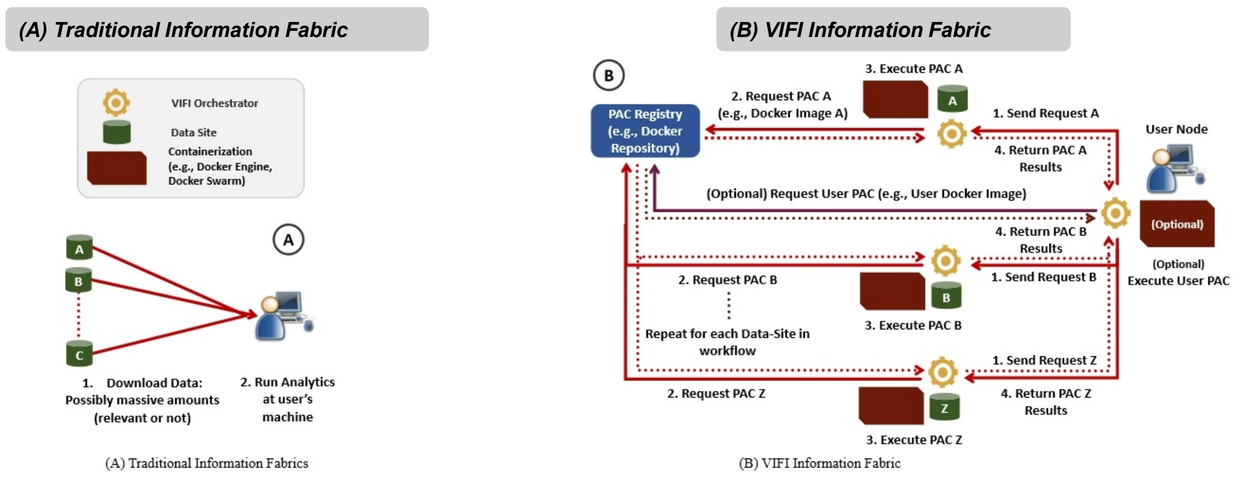
CONCEPT

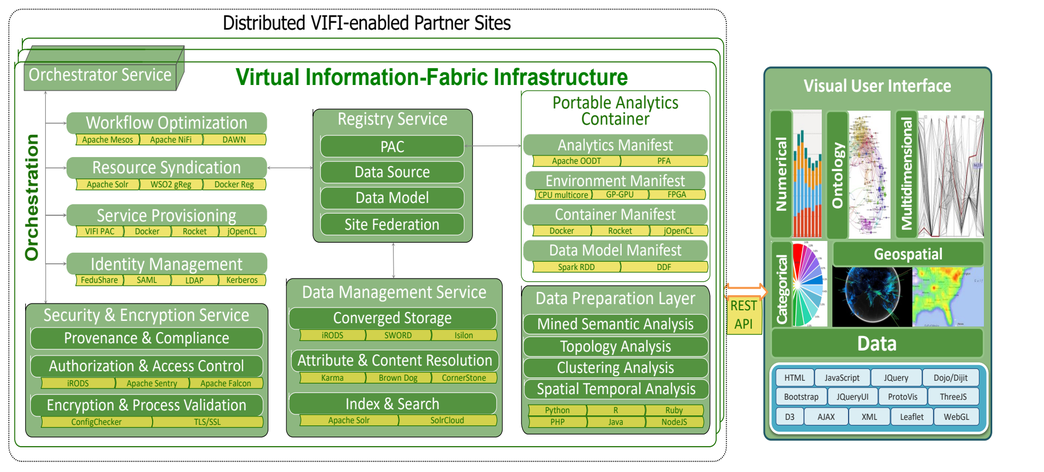
A novel cyberinfrastructure that facilitates data-driven discovery from distributed, fragmented, and un-shareable data without requiring the movement of massive amounts of data or directly exposing sensitive, raw data to end users

ACKNOWLEDGEMENT

Funding provided by the National Science Foundation
Data Infrastructure Building Blocks Program under Award #1640818

VIFI ARCHITECTURE





TRADITIONAL DATA FABRIC ARCHITECTURE

1. Ad-hoc search for relevant data from online search
2. Required to move the data to a central location for analytics
3. Extended time spans for transferring massive datasets to the User Site
4. Storage requirements for storing massive datasets on the User Site
5. All computations are executed at the User Site
6. All data are transferred to the User Site – including data that may not be relevant to the analysis
   Scientist must manually install the algorithms (including all dependencies) at the User Site

VIFI DATA FABRIC ARCHITECTURE

1. All phases of the scientific analysis lifecycle (compute and data transfer) are executed by an orchestrator, without manual intervention or a-priori knowledge on the scientist's part
2. Scientific algorithms are encapsulated in reusable PACs that are seamlessly deployed to any VIFI-enabled Site
3. Computations are distributed to multiple servers, where each server has direct access to the data
4. Only a subset of the data (i.e., the derived analytical results from PAC execution) are migrated, drastically reducing the data transfer times/costs
5. Infrastructure scaling to new data sources is achieved by VIFI software installation

EXPECTED SCIENCE ADVANCES & IMPACTS

General

- Data-intensive, scientific analysis at scale over distributed data
- Improved data-driven research outcomes by virtually and seamlessly integrating and analyzing distributed, un-shareable data
- Reduced data management burdens on data scientists when conducting multidisciplinary impact analysis across domains
- New analytical methods for distributed, fragmented data
- New interactive methods for data discovery, exploration, and analysis

Earth Science

- More granular (regional-scale), data-rich climate observations
- Enhanced ability to generate significantly higher resolution climate models
- Improved understanding for critical climate questions – e.g., predicted rainfall over the CONUS

Astronomy

- Expanded citizen-science and crowd-sourcing through wider data access without the associated data transfer costs
- Integrated analysis of sparse (seemingly insignificant) data to aid transient science in a significant manner

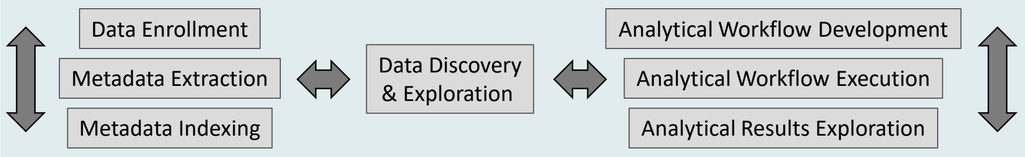
Sustainable Human-Building Ecosystems

- Mitigation of multiple data fragmentation problems: Spatial fragmentation (SF), temporal fragmentation (TF), spatial-temporal fragmentation (STF) and data requirements fragmentation (DRF),
- Increase consistency, reliability, and accuracy in decision-making through VIFI-enabled analyses
- Cultivation of emergent applications via virtual information fabrics
- Integrated resilience analysis for human-building ecosystems
- Sustainability analysis and modeling at scale

SUSTAINABILITY PLAN

- Distributing VIFI as an open source infrastructure that is available to all stakeholders
- Improving adoption by extending open access to user communities and providing workshops at appropriate conferences and symposia
- Integrating into existing system capabilities for climate research from DOE (e.g., Earth System Grid) and NASA (Earth Observing System)
- Cultivating user community that will establish appropriate consensus mechanisms to determine VIFI strategy/standards
- Tracking and periodically reporting the number, volume, and types of datasets ingested into VIFI
- Introducing tools and metrics to assess the compatibility of datasets for integrated analyses and comparative assessments

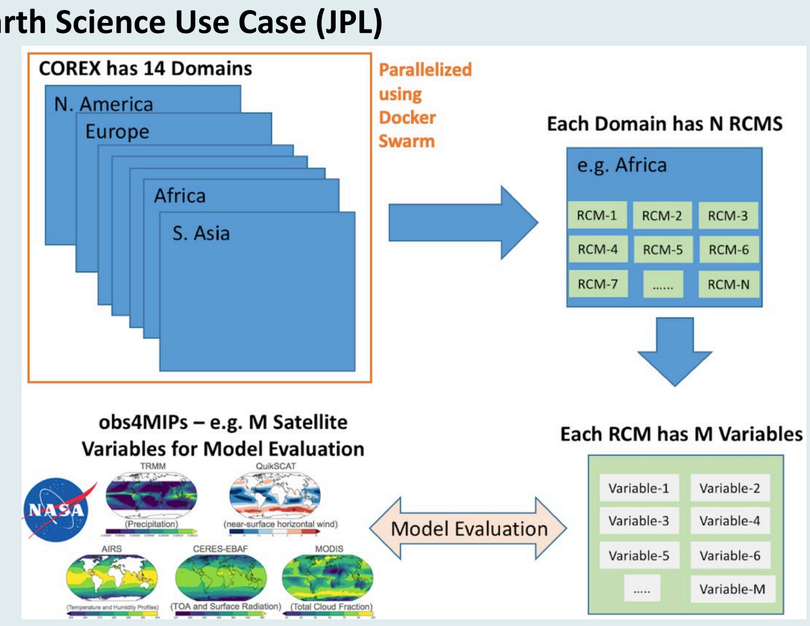
DATA-DRIVEN ANALYTICAL WORKFLOWS



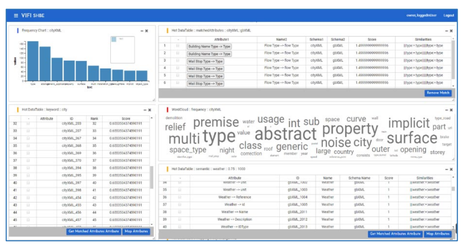
USER CASE SCENARIOS / END-TO-END USER EXPERIENCE

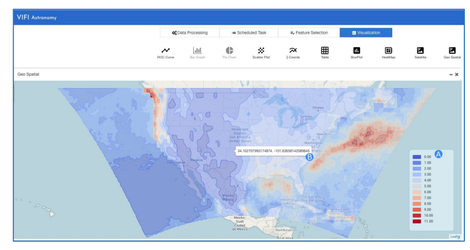
(some data in screenshots are illustrative)

Earth Science Use Case (JPL)

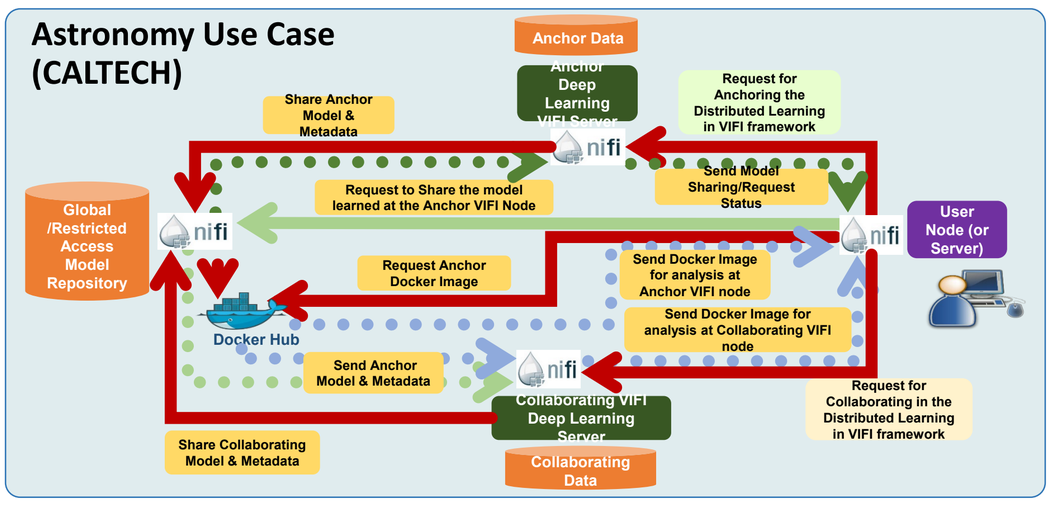


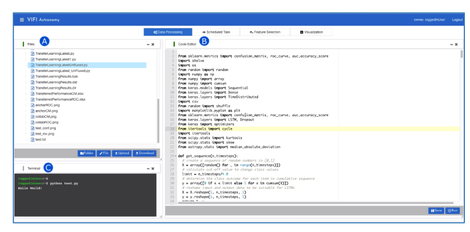
Multi-model, Multi-variable Evaluation v. Satellite Observations (adapted from Lee et.al, 2018) - Coordinated Regional Climate Downscaling Experiment (CORDEX); Regional Climate Model (RCM) simulations; Regional Climate Model Evaluation System (RCMES)





Astronomy Use Case (CALTECH)





SIGNIFICANT MILESTONES

- **Project Y1**
  - Established internal mechanisms to create an open-source toolkit
  - Demonstrated baseline capabilities for distributed orchestration of a federated infrastructure for portable analytics
  - Validated the baseline capabilities against initial use-cases from multiple application domains
- **Project Y2**
  - *Server and Orchestration* → Expanded orchestration capabilities to including sequential and parallel workflows
  - *Data Discovery and Management* → Prototyped/integrated a data management framework with capabilities for meta-data extraction, semantic search and matching, and data owner infrastructure virtualization design
  - *Middleware and Security* → Designed/demonstrated the baseline integration of the VIFI security architecture
  - *User Experience* → Designed/demonstrated the baseline, end-to-end user experience for data owners and users
- **Project Y3**
  - *Server and Orchestration* → Extended server and orchestration to support (across distributed, fragmented data): i) iterative analytical workflows; and, ii) the transfer of machine learning models across VIFI sites; Introduced container caching, logging, and server optimizations
  - *Data Discovery and Management* → Advanced data discovery and management with support for semantic search and semantic match capabilities
  - *Middleware and Security* → Designed and implemented baseline middleware server to mediate interactions between UI and orchestration services
  - *User Experience* → Enhanced the end-to-end UI design and implementation; Designed and implemented a baseline VISREC system that extracts metadata from derived data and facilitates the construction of interactive visualizations based on metadata, user preferences, and associated requirements
- **Use Case Teams**
  - Demonstrated expanded analytics capabilities and new approaches to analyses
  - Opened new research directions – e.g., framework for optimizing analytical costs vs. uncertainty; multi-view, generative, transfer learning for distributed time series classification (**best paper award**, *The 4th Big Data Transfer Learning Workshop, IEEE Big Data 2019*).
  - Validated the VIFI against workflows from multiple domains while advancing research in each domain
  - Made distributed, fragmented, un-shareable data and analytical techniques more accessible
  - Adding new use cases – e.g., analyzing Clinical Practice Guidelines (medical)

UNC CHARLOTTE
College of Computing and Informatics