Towards a Distributed Infrastructure for Data-Driven Discoveries & Analysis

Mohammed Elshambakey*†, Mohamed Khalefa¶, William J. Tolone*, Sreyasee Das Bhattacharjee*,
Huikyo Lee§, Luca Cinquini§ Shannon Schlueter*, Isaac Cho*, Wenwen Dou*, Daniel J. Crichton§
*CS Dept, University of North Carolina Charlotte, USA. {melshamb,mkhalefa,wjtolone,sschluet,sdasbhat,icho1,wdou1}@uncc.edu
†AI Dept, IRI, SRTA-City, Egypt. mshambakey@srtacity.sci.eg
¶University of Loiisville, mohamed.khalefa@louisville.edu
§Jet Propulsion Laboratory/Caltech, USA. {Huikyo.Lee,luca.cinquini,Daniel.J.Crichton}@jpl.nasa.gov

***Abstract*—Big data analytics traditionally involves download of massive amounts of datasets to common server/cluster for processing. Analytic process gets slower with increasing size of required data and network conditions. Data scientists also need explicit access to data locations to download required data. Explicit access to required data may not always be granted due to security reasons. To simplify and accelerate the analytics process on distributed big data with security considerations, we proposed the Virtual Information Fabric Infrastructure (VIFI) for data driven discoveries. Instead of moving large amounts of data to a common place of processing, VIFI allows automatic transfer of required analytics programs to the distributed data locations for in-place processing of relevant data. VIFI allows data scientists to conduct and coordinate complex analytics processes on distributed data repositories using containerization technology and open-source workflow design tools. VIFI alleviates users from having detailed knowledge of distributed data locations, as well as required dependencies, installation and configuration of analytical libraries. In this paper, we demonstrate our current and future work to improve the VIFI architecture using previous and additional uses cases, data management layer that simplifies search of relevant data sets through addition of metadata, integration with security policies at different institutions with the proposed VIFI security layer, and the use of a user-friendly web interface to carry different VIFI activities.**

I. Introduction

The critical task of automatically analyzing large volumes of continuously generated data from multiple heterogeneous sources has received significant attention from academic, scientific, and corporate communities. These structured and unstructured data give rise to complex, *ad-hoc* processing requirements across the data management and analysis lifecycle including cleaning, discovery, preparation, indexing, analysis, exploration, and visualization that must be handled with maximum care to ensure effective revelation of meaningful insights. Such analyses often demand efficient, high-performance computing infrastructures with robust network and storage resources to move and analyze large volumes of data. To this end, we propose an end-to-end comprehensive computing framework *Virtual Information Fabric Infrastructure (VIFI)*, that supports the management and analysis of large volumes of heterogeneous data by enabling complex analytical workflows over distributed information resources across multiple sites. *VIFI* enables these analytical workflows by moving analysis to the data rather than requiring the movement or sharing of large corpora of data. Under VIFI, data owners permit analyses over their data without directly sharing their data. Thus, data scientists leverage the virtual information fabric to execute complex, distributed, ad- hoc analyses. The VIFI approach speeds algorithm implementation and testing processes by utilizing existing shared resources, while ensuring sufficient flexibility and freedom for data scientists. In contrast to transferring massive datasets from distributed repositories to a common site for analysis, VIFI offers an efficient alternative by orchestrating a distributed, analytical workflow, through which users can conduct, execute, and coordinate complex analytics activities in a parallel manner at multiple data sites.

II. VIFI Architecture

VIFI abstracts and isolates access to the underlying shared infrastructure through containerization technology. The proposed VIFI architecture is shown in Figure 1. VIFI consists of the following major components. (1) *Portable Analytic Container (PAC)* is a containerized environment that provides end-users with a standardized, easy-to-use functionality to write, edit, deploy, and execute analytical programs on distributed data repositories. In our demonstration implementation, each PAC is implemented as a Docker Image [1] that contains the dependencies for running the required analytics in Docker Swarm [2]. Thus, PACs allow user to share and re-use analytical programs without worrying about installation and configuration of required dependencies. (2) *Registry Service* is the local and/or central repositories that enables users to store, use and share PACs. Current VIFI implementation uses public Docker repositories like Docker Hub [3] and AWS ECR [4] as the Registery Service. (3) *User-Node* is the access point with a user interface and basic network and compute capabilities that enable users to communicate with VIFI, send analytical requests, and receive results. (4) *Data Sites* are registered organizations in the VIFI fabric that host different aspects of heterogeneous data from various sources and models. (5) *Metadata Server* stores and indexes collected information and annotations about datasets and files, which can be generated using set of registered extractors or manually added by users and data owners. For portability, extractors are containerized as PACs. A metadata server facilitates dataset discovery by data scientists against specified criteria as shown in Figure 2. (6) A *Crawler* performs the data crawling by applying appropriate extractors and transferring extracted data

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

to the Metadata Server. (7) A *Watchdog* monitors datasets for modification to keep metadata current. (8) *Orchestrator* is a VIFI service for automatic communication and organization between different VIFI components residing at different sites. Current VIFI implementation uses NIFI [5] as the orchestrator as NIFI provides the ability to design distributed workflows with secure site-to-site communication [6].

VIFI components are loosely coupled using REST API. Data owners register their data through **VIFI** User-Node by providing a suitable high-level description with an aim to simplify data discovery processes to be performed by data scientists. In order to assist in the analytic tasks, predefined analytical scripts supplied by the data owners can be leveraged by data scientists as part of their analytical workflows. Alternatively, data scientists can contribute customized analytical scripts and workflows to be shared across the VIFI infrastructure. Unlike most existing cyber-infrastructure systems that provide a complex combination of tools and services for data discovery, retrieval, and analysis -requiring the movement of massive data for analyses and the provisioning of significant computational environments at the data scientist site -VIFI provides a "truly distributed analytics" environment by integrating data analysis within existing infrastructures that support appropriate authorization control for data and metadata access, and execution of models for end users without exposing private datasets. Within a novel architecture, VIFI offers a principled as well as flexible data management and analytic solution that holds significant promise in terms of improved productivity in a secured, shared environment with reduced administrative efforts.

III. VIFI Current and Future Work

In order to demonstrate VIFI suitability across multiple application domains, the VIFI framework is presently customized and deployed in three use cases from Earth Science, Astronomy, and Structural Engineering domains. Primary VIFI implementation of the Earth Science use case is compared against the traditional data fabric approach in [7]. In the Earth Science use case, joint probability distribution functions (JPDF) of model data ($\approx$ 21 Gigabytes) and observational data ($\approx$ 11 Gigabytes) are compared to evaluate the performance of derived models. The observation data resulted from the Integrated Multi-satellite Retrievals for Global Precipitation Measurement (GPM IMERG, or simply GPM [8]). GPM has provided a gridded precipitation over the globe every 30 minutes since 2014 by combining multiple satellites and ground gauge observations. Model data is the output of climate simulation from NASA Unified WRF (NU-WRF, [9]) at multiple resolutions (e.g., 4km, 12km, and 24km) at different remote servers. The Earth Science use case was implemented using AWS services (EC2 [10], [11], ECR [4], and S3 [12], [11]). Comparison results show that transfer time under VIFI approach ($\cong$ 1 sec) is much less than transfer time under traditional data fabric approach ($\cong$ 246 sec). VIFI is expected to show much better performance when deployed in real institutions with lower network speeds and smaller bandwidths than AWS. Analytics transfer to data locations using PACs under VIFI will be more useful in case of frequent data streaming.

At present, we are working on extending VIFI data management capabilities further within a distributed setting, by employing containerized machine learning algorithms. With a stronger yet adaptable security layer, VIFI expands its security layer design to integrate different security policies across different data sites and user nodes. The proposed VIFI security layer architecture is shown in Figure 3. VIFI security layer aims to provide proper authentication and authorization, single sign-on, identity mapping of different accounts of the same user, and detailed identification of security credentials of each user. VIFI security layer uses well known security standards like OAuth 2.0 [13], OpenID Connect (OIDC) [14], SAML 2.0 [15], Role/Attribute Based Access Control (RBAC/ABAC) [16], [17] and JSON Web Token (JWT) [18]. Other goals of VIFI security layer include secure communication and content encryption between different registered institutions, data privacy by processing encrypted data, protection from harmful programs, and proper containerization security. In the present phase, VIFI is being employed for a wide range of analytical use cases including: astronomical analyses from huge, distributed observational data collections; optimization of energy consumption for the design of sustainable resilient human-building ecosystems; and model vs. observational earth science data analyses. With a user-friendly web interface to simplify access to different functionalities, VIFI offers a comprehensive, cost effective, and risk averse solution to the challenge of distributed analyses across massive heterogeneous data.

Future work of VIFI includes resource management and scheduling based on current and expected workloads in order to optimize concurrent analytical requests processing using different tools (e.g., DAWN [19] and IReS [20]). VIFI will use logging and monitoring capabilities to provide real-time progress and run-time errors to users. Additional and more sophisticated use cases will be exploited to increase and improve VIFI functionalities.

Acknowledgment

Funding for this research was provided by the National Science Foundation (NSF) Data Infrastructure Building Blocks (DIBBs) Program under Award number 1640818. A portion of this work was performed by the Jet Propulsion Laboratory, California Institute of Technology under contract to the National Aeronautics and Space Administration. The VIFI team is thankful to Prof. Ashit Talukder for his leadership on the project from Oct 2016 to July 2017.

References

[1] https://docs.docker.com/engine/.
[2] https://docs.docker.com/engine/swarm/.
[3] https://hub.docker.com/.
[4] https://aws.amazon.com/ecr/.
[5] https://nifi.apache.org/.
[6] https://nifi.apache.org/docs/nifi-docs/html/user-guide.html#site-to-site.
[7] A. Talukder, M. Elshambakey, S. Wadkar, H. Lee, L. Cinquini, S. Schlueter, I. Cho, W. Dou, and D. J. Crichton, "Vifi: Virtual information fabric infrastructure for data-driven discoveries from distributed earth science data," in *The 3d IEEE Conference on Cloud and Big Data Computing (CBDCom)*, In Press.

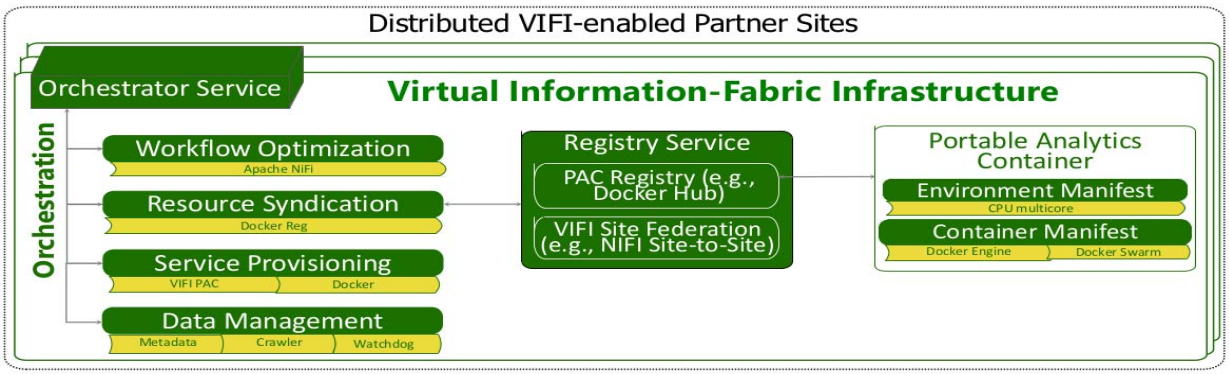

Fig. 1. VIFI architecture framework

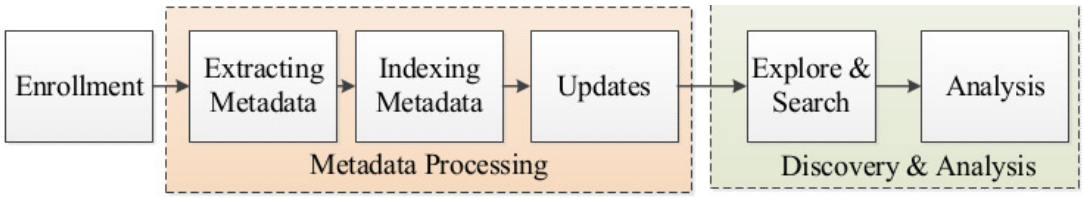

Fig. 2. Data life cycle in VIFI

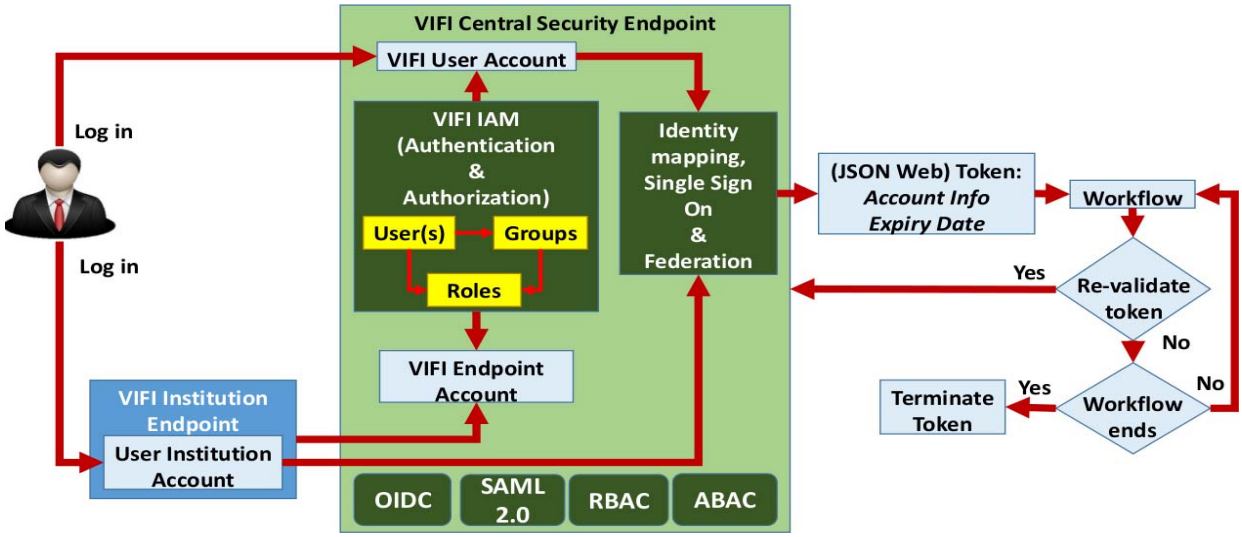

Fig. 3. VIFI proposed security layer architecture

[8] G. J. Huffman, D. T. Bolvin, D. Braithwaite, K. Hsu, R. Joyce, C. Kidd, E. J. Nelkin, and X. P., "Nasa global precipitation measurement (gpm) integrated multi-satellite retrievals for gpm (imerg)," *ATBD Version 4.5*, 2015.

[9] C. D. Peters-Lidard, E. M. Kemp, T. Matsui, J. A. Santanello, S. V. Kumar, J. P. Jacob, T. Clune, W.-K. Tao, M. Chin, A. Hou, J. L. Case, D. Kim, K.-M. Kim, W. Lau, Y. Liu, J. Shi, D. Starr, Q. Tan, Z. Tao, B. F. Zaitchik, B. Zavodsky, S. Q. Zhang, and M. Zupanski, "Integrated modeling of aerosol, cloud, precipitation and land processes at satellite-resolved scales," *Environmental Modelling and Software*, vol. 67, pp. 149–159, 2015.

[10] A. Inc, *Amazon Elastic Compute Cloud (Amazon EC2)*. http://aws.amazon.com/ec2/#pricing: Amazon Inc., 2008.

[11] J. Murty, *Programming Amazon Web Services - S3, EC2, SQS, FPS, and SimpleDB*. Farnham: O'Reilly, 2008.

[12] https://aws.amazon.com/documentation/s3/.

[13] https://oauth.net/2/.

[14] http://openid.net/connect/.

[15] http://docs.oasis-open.org/security/saml/Post2.0/sstc-saml-tech-overview-2.0.html.

[16] https://csrc.nist.gov/projects/role-based-access-control.

[17] https://csrc.nist.gov/Projects/Attribute-Based-Access-Control.

[18] https://jwt.io/.

[19] H. Lee, L. Cinquini, D. Crichton, and A. Braverman, "Optimization of system architecture for big data analysis in climate science," in *IEEE International Conference on Big Data (Big Data)*, Oct 2015, pp. 2169–2172.

[20] K. Doka, N. Papailiou, D. Tsoumakos, C. Mantas, and N. Koziris, "Ires: Intelligent, multi-engine resource scheduler for big data analytics workflows," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '15. New York, NY, USA: ACM, 2015, pp. 1451–1456.