SMILE Reconstructed: A User-Friendly Science Gateway for Social Media Research and Collaboration

Chen Wang
National Center for Supercomputing Applications
University of Illinois at Urbana Champaign
Urbana, Illinois, USA
cwang138@illinois.edu

Yong Wook Kim
National Center for Supercomputing Applications
University of Illinois at Urbana Champaign
Urbana, Illinois, USA
ywkim@illinois.edu

Rob Kooper
National Center for Supercomputing Applications
University of Illinois at Urbana Champaign
Urbana, Illinois, USA
kooper@illinois.edu

Joseph T. Yun
Swanson School of Engineering
University of Pittsburgh
Pittsburgh, Pennsylvania, USA
joe.yun@pitt.edu

***Abstract*—The Social Media Macroscope (SMM) project introduces SMILE (Social Media Intelligence and Learning Environment), an open-source science gateway tailored for social media research. Notably, SMILE has recently undergone significant transformations. It has transitioned to utilize Docker containers for enhanced portability and scalability, effectively deployed and orchestrated via Kubernetes and Helm charts. Furthermore, SMILE now leverages CIlogon, a robust identity management platform, for improved identity and access control. These adaptations empower researchers with seamless access to data collection from platforms like Twitter and Reddit, alongside advanced features encompassing natural language processing, sentiment analysis, network analysis, and machine learning classification. To foster collaboration and data sharing, SMILE seamlessly integrates with NCSA Clowder, a widely adopted framework within the science gateway community.**

Keywords—science gateways, research computing infrastructure, cloud computing, data management, container technologies, text mining, natural language processing

I. INTRODUCTION

Despite the availability of numerous services and tools for social media analytics, text mining, and natural language processing, there remain significant deficiencies in terms of usability, accessibility, and transparency. Commercial applications like Brandwatch [1] and Insights in Meta Business Suite [2] may offer user-friendly interfaces but are accompanied by high costs and usage restrictions that primarily cater to marketing needs, making them less accessible to academia. Their lack of transparency hampers the integration of cutting-edge research into practical applications and academic endeavors. Conversely, open-source toolkits like the Python NLTK [3] library provide transparency but require programming skills, limiting their accessibility to non-programmers.

To address these challenges, the Social Media Macroscope (SMM) project [4][5] was initiated in 2017 as a HUBzero-powered [6] social media science gateway. With over 1000 users and 5000 instances of usage worldwide, the tools developed under SMM have demonstrated reliability and appeal. In 2022, we received funding from the National Center for Supercomputing Applications (NCSA) through the Center-Directed Discretionary Research program. This support has allowed us to explore different hosting and deployment strategies, ensuring the tools are more portable and easily deployable both on cloud platforms and on-premises environments.

SMILE, an acronym for Social Media Intelligence and Learning Environment [7][8], is a comprehensive web application within the SMM that tackles the limitations of existing tools. It provides data collection from Twitter and Reddit, along with various natural language processing, sentiment analysis, machine learning classification, and network analysis. Users can access SMILE through a web-based user interface. SMILE is committed to academic research, offering free access for teaching, research, and publication purposes, maintaining transparency through open-source code repositories, facilitating the dissemination of algorithms developed by social scientists and computer scientists, and ensuring compliant data collection by following platform protocols and regulations.

II. MAIN FEATURES

A. Access management integrating with CILogon

CIlogon [9] is an integrated identity and access management platform for science developed at NCSA. Since it supports federated identity management (Shibboleth [10], InCommon [11]) with collaborative organization management (Comanage) with over 5000 identity providers, CILogon makes it easy and secure for users to access cyberinfrastructure resources by authenticating with their home institutions' identities. Many science gateways, such as XSEDE [12] and Globus [13][14], rely on CILogon for authentication and authorization.

We opted to integrate with CILogon due to its lightweight implementation with SMILE, eliminating the need to maintain a user database and dedicate resources to manage account registration and approval. We utilize the standards-compliant OpenID Connect (OAuth 2.0) flow [15], which works smoothly with popular libraries. Upon prompt to login to SMILE, users are redirected to CILogon's page where they can select their home institution's Identity Provider, such as NCSA, and login with their existing account. The user's email address is then recorded by the SMILE app as their unique identifier.

B. Social Media Data

SMILE offers a user-friendly interface for non-programmers to query and collect social media data, with a GraphQL-based [16] data server managing user requests and retrieving data from social media platforms. This data is then stored in an open-source MinIO [17] object storage.

C. Social Media Analytics

Following the collection and transformation of social media data, users are able to submit their data for specific analytics computation. Users can customize their analysis via dropdown menus, radio buttons, text boxes, and slider bars to define specific input parameters. Once submitted, the algorithms are invoked to process the data, with the resulting outputs stored as files in the MinIO storage and subsequently returned to the SMILE server for rendering on the user interface. Each analytics module provides references to the underlying algorithms used, and users are encouraged to refer to the associated academic publications for more comprehensive insights.

At present, SMILE offers support for a variety of analytics features, including topic modeling, entity recognition, phrase mining, network analysis, text classification, natural language processing, and sentiment analysis. As an example: sentiment analysis, also referred to as opinion mining, is the systematic process of identifying, extracting, quantifying, and analyzing affective states and subjective information. SMILE offers 3 sentiment analysis algorithms - the Valence Aware Dictionary and Sentiment Reasoner (VADER) algorithm [18], which is a sentiment analysis modeler developed using Twitter data; the SentiWordNet algorithm [19], which is a lexical resource for opinion mining; and the sentiment classification model using debiasing word embeddings (debiasing) [20] which has a training step that adjusts the embeddings to identify and remove some sources of algorithmic racism and sexism. Both VADER and SentiWordNet algorithms produce sentiment scores classified as neutral, negative, or positive; debiasing produces one overall sentiment score that indicates relative positivity. We provide a pie chart that displays the breakdown of the sentiment of the entire corpus, and we also list the scores for each individual text.

D. Data Sharing

To facilitate data sharing in research-focused scenarios, SMILE offers workflows for sharing analysis outputs with a dedicated instance of Clowder [21] (SMM Clowder [22]), a customizable and scalable data management framework to support any data format in multiple research domains. Clowder serves as the core building block in numerous science gateways, including TERRA-REF [23], 4ceed [24], and Permafrost Discovery Gateway [25]. It offers extensive metadata tracking for raw files, facilitates data visualization, and incorporates built-in access management. By enabling the export of computation results to Clowder, SMILE can leverage the existing infrastructure of Clowder and foster collaboration within the Clowder community.

SMILE incorporates a user-friendly wizard that empowers users to share, organize, and manage data with Clowder by leveraging the robust APIs of the Clowder platform.

Moreover, our objective is to explore the paradigm of bringing computation to data by utilizing the customized information extractor functionality within Clowder, thus advancing knowledge

discovery in immediate proximity to the data. We package the identical analytics functionalities available in SMILE, such as sentiment analysis, into dedicated extractors within Clowder. This empowers users of Clowder to submit raw social media data collected using SMILE to these extractors. Additionally, Clowder provides a preview capability, enabling users to visualize the extracted information and quickly acquire valuable insights.

III. ARCHITECTURE AND TECHNOLOGY STACK

In SMILE, a microservice architecture (see Fig.1) is implemented to leverage its portability and scalability. This is a significant departure from SMILE's previous design, which was a combination of HubZero tool integrated with AWS components such as Lambda for algorithms and S3 for data storage, as detailed in our previous publication [7]. In the reconstructed SMILE, Dockerized components interact with each other over the network, providing users with direct access through a web interface. The SMILE server, developed using the Node.js/Express framework known for its lightweight yet powerful capabilities [26], and employing Pug as the templating engine due to its concise and elegant syntax [27], effectively handles user requests by routing them to various components and managing corresponding responses. The Data Server, developed using GraphQL, takes advantage of its runtime data fulfillment capabilities and the ability to abstract data schemas, allowing for a clean API for diverse social media sources. Each analytics algorithm is encapsulated within an isolated container, enabling serverless computation and facilitating the easy integration of new analytical features. RabbitMQ [28] is employed as the chosen message-broker to facilitate communication between each microservice in SMILE. It is selected for its reliability, widespread adoption, extensive community support, scalability, and flexibility. Data is stored as file-based objects in a MinIO instance, which offers enterprise-grade encryption, identity management, access control, and data protection features. The selection of MinIO was motivated by its compatibility with AWS S3, as it ensures backward compatibility with our previous AWS S3 based data storage approach. Furthermore, the web interface of SMILE provides the capability to export data to our customized instance of the Clowder data management system.

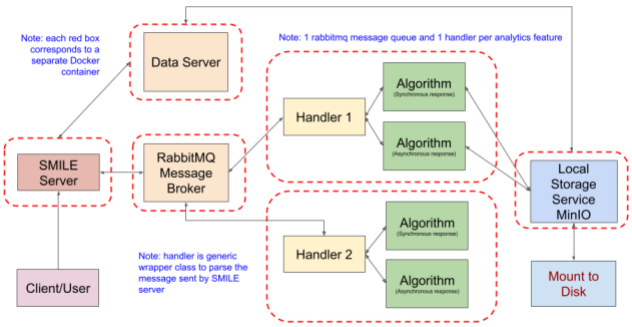

Fig. 1. the archetecture of SMILE

IV. DEPLOYMENT

In recent years, there has been a growing trend in which science gateways are increasingly incorporating container technology, including Docker [29] and Kubernetes [30], into their infrastructure. Docker containers offer an efficient and lightweight solution for packaging and distributing scientific applications, ensuring consistency and reproducibility. Concurrently, Kubernetes serves as a robust orchestration platform, simplifying the management of containerized applications. Its automated scaling, fault tolerance, and optimized resource allocation capabilities align well with the dynamic and resource-intensive nature of scientific workloads.

SMILE is deployed within virtual machines on Radiant. Radiant, an openstack solution offered by NCSA, provides researchers with a robust and flexible cloud computing option supported by a dedicated team of system, security, and network experts. Kubernetes is employed for deployment for SMILE, and Helm charts are utilized to define, install, and manage applications within Kubernetes. The integration of Helm charts streamlines the deployment and management of complex applications through the use of templating and version control mechanisms. All components of SMILE, including the SMILE server, GraphQL data server, algorithms, MinIO instance, and others, are encapsulated as Docker containers and deployed individually within Kubernetes. This approach enables independent management of component activation, scaling, and storage configuration, without causing disruptions to other components. Additionally, the internal networking system provided by Kubernetes enhances the overall security of the system.

Alternatively, for lightweight deployments without scalability considerations, SMILE offers a Docker Compose setup option, providing a simpler learning curve and requiring fewer resources compared to Kubernetes.

V. Conclusion

In conclusion, SMILE serves as an open-source science gateway, providing researchers and students with essential tools for social media data collection, analysis, and visualization. This paper has outlined the ongoing efforts to enhance SMILE's infrastructure by transitioning its components from commercial hosting to the NCSA Radiant computing environment. Through the utilization of Docker and Kubernetes, it enhances the platform's portability and scalability. Moreover, the integration of NCSA tools such as Clowder for data management and CILogon for user sign-on has been discussed.

For future work, we would like to examine the impact of the recent implementation of paid models by major social media companies (e.g., Twitter) on data collection practices and exploring alternative data sources. Additionally, we will focus on developing methodologies for gathering and analyzing multimedia content, utilizing advanced analytics tailored specifically for images, audios, and videos. Furthermore, we will explore the application of AI language models to extract insightful information from social media content, enhancing our understanding of the subject.

Furthermore, in close collaboration with the NCSA industry program, we actively pursue opportunities to engage with research projects, initiatives, and industry partners to foster knowledge sharing, resource sharing, and the development of innovative solutions. The adaptability and compatibility of SMILE with various cloud environments and on-premises resources provide a solid foundation for future collaborations

References

[1] Brandwatch. Brandwatch. https://www.brandwatch.com

[2] Meta. Meta Business Suite. https://www.facebook.com/business/tools/meta-business-suite

[3] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

[4] Yun, J. T., Vance, N., Wang, C., Marini, L., Troy, J., Donelson, C., ... & Henderson, M. D. (2020). The Social Media Macroscope: A science gateway for research using social media data. Future Generation Computer Systems, 111, 819-828.

[5] Social Media Macroscope. Social Media Macroscope. https://smm.ncsa.illinois.edu

[6] HUBzero. HUBzero. https://hubzero.org/

[7] Wang, C., Marini, L., Chin, C. L., Vance, N., Donelson, C., Meunier, P., & Yun, J. T. (2019, September). Social Media Intelligence and Learning Environment: an Open Source Framework for Social Media Data Collection, Analysis and Curation. In 2019 15th International Conference on eScience (eScience) (pp. 252-261). IEEE.

[8] SMILE. SMILE. https://smile.smm.ncsa.illinois.edu

[9] Basney, J., Flanagan, H., Fleury, T., Gaynor, J., Koranda, S., & Oshrin, B. (2019). CILogon: Enabling federated identity and access management for scientific collaborations. Proceedings of Science, 351, 031.

[10] Shibboleth. Shibboleth. https://www.shibboleth.net/

[11] InCommon. InCommon. https://incommon.org/

[12] Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., ... & Wilkins-Diehr, N. (2014). XSEDE: accelerating scientific discovery. Computing in science & engineering, 16(5), 62-74.

[13] Foster, I. (2011). Globus Online: Accelerating and democratizing science through cloud-based services. IEEE Internet Computing, 15(3), 70-73.

[14] Allen, B., Bresnahan, J., Childers, L., Foster, I., Kandaswamy, G., Kettimuthu, R., ... & Tuecke, S. (2012). Software as a service for data scientists. Communications of the ACM, 55(2), 81-88.

[15] Recordon, D., & Reed, D. (2006, November). OpenID 2.0: a platform for user-centric identity management. In Proceedings of the second ACM workshop on Digital identity management (pp. 11-16).

[16] GraphQL. GraphQL. https://graphql.org/

[17] MinIO. MinIO. https://min.io/

[18] Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).

[19] Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Lrec (Vol. 10, No. 2010, pp. 2200-2204).

[20] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29.

[21] Marini, L., Gutierrez-Polo, I., Kooper, R., Satheesan, S. P., Burnette, M., Lee, J., ... & McHenry, K. (2018). Clowder: Open source data management for long tail data. In Proceedings of the Practice and Experience on Advanced Research Computing (pp. 1-8).

[22] SMILE. SMILE Clowder instance. https://clowder.smm.ncsa.illinois.edu/

[23] Burnette, M., Kooper, R., Maloney, J. D., Rohde, G. S., Terstriep, J. A., Willis, C., ... & LeBauer, D. (2018). TERRA-REF data processing infrastructure. In Proceedings of the Practice and Experience on Advanced Research Computing (pp. 1-7).

[24] Nguyen, P., Konstanty, S., Nicholson, T., O'brien, T., Schwartz-Duval, A., Spila, T., ... & Paquin, N. (2017, May). 4ceed: Real-time data acquisition and analysis framework for material-related cyber-physical environments. In 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID) (pp. 11-20). IEEE.

[25] Liljedahl, A. K., McHenry, K., Jones, M. B., Cervenec, J. M., Witharana, C., Budden, A. E., ... & Udawalpola, R. (2020, December). The Permafrost Discovery Gateway: A web platform to enable knowledge-generation from big geospatial data. In AGU Fall Meeting Abstracts (Vol. 2020, pp. C016-08).

[26] Express. Express. https://expressjs.com/

[27] Pug. Pug. https://pugjs.org/api/getting-started.html

[28] RabbitMQ. RabbitMQ. https://www.rabbitmq.com/

[29] Docker. Docker. https://www.docker.com/

[30] Kubernetes. Kubernetes. https://kubernetes.io/