Extended Abstract: VIDIA: a HUBzero Gateway for Data Analytics Education

Jeanette M. Sperhac,* University at Buffalo; Steven M. Gallo, University at Buffalo
*Corresponding author address: Center for Computational Research, University at Buffalo, State University of New York, Buffalo, NY, 14203, USA; email: jsperhac@buffalo.edu

Abstract: *We describe a scientific gateway collaboration undertaken by members of the State University of New York (SUNY) system. The University at Buffalo's Center for Computational Research (CCR) partnered with SUNY College at Oneonta to offer a gateway for teaching data analytics. The result, called VIDIA, hosts open-source tools that have been used by more than 250 students enrolled in 16 SUNY courses. Additional tools enable researchers across the SUNY system to submit larger jobs to CCR's compute cluster. VIDIA supports data-intensive computation for teaching and research at campuses that lack access to traditional high-performance computing (HPC) resources.*

1. VIDIA, a HUBzero gateway

Educators are increasingly interested in teaching data analytics in order to prepare students for our data-centric society. For social scientists, this often entails studying the discussions and controversies circulating in social media. However, these datasets readily grow to a size that greatly exceeds the analytical capability of personal computers running common software tools. Fortunately, gateways can help students and faculty gain access to the computational resources they need to engage in these analyses.

1.1 The collaboration

Faculty at the State University of New York (SUNY) Oneonta wanted to train their students to extract knowledge and insights from large, complex collections of social media data. However, typical of a small college, Oneonta lacked the infrastructure needed to create, manipulate, and analyze such large datasets.

SUNY Oneonta's existing IT infrastructure is representative of many small colleges; in 2013, the storage available to its 7,000 student, faculty, and staff users totaled 4 TB. Software available on the campus was limited to standard Windows applications. Furthermore, Oneonta has no high-performance computing (HPC) capability, limiting the types of analysis that students and researchers could perform.

The University at Buffalo's Center for Computational Research (CCR) comprises more than 170 Tflops of peak performance compute capacity and 4 PB of high-performance storage. CCR's academic HPC cluster offers more than 8000 processor cores with QDR Infiniband. Other CCR resources include an industry HPC cluster, a computer visualization laboratory, and a private cloud to support research and teaching activities. CCR also provides other services to facilitate research, including custom software development and data analytics. To help address SUNY Oneonta's lack of computational facilities, CCR created VIDIA, a gateway that enables Oneonta to integrate data analytics into their curriculum.

1.2 The VIDIA platform

VIDIA is built on the HUBzero® Platform for Scientific Collaboration [1]. This innovative platform enables users to launch and access advanced tools and computations using only a web browser. A hub combines unique middleware with advanced interactive functionality, providing a platform that is much more powerful than an ordinary website. CCR selected the HUBzero platform for VIDIA in part because of their previous experience creating VHub.org, a community cyberinfrastructure platform enabling collaboration in volcanology research, education, outreach, and discovery. [2] An exhaustive description of HUBzero's architecture, capabilities, and installation is beyond the scope of this paper.

The VIDIA gateway allows users to launch computations on CCR's HPC cluster, view results,

Presented at Gateways 2016, San Diego Supercomputer Center, La Jolla, CA, November 2-3, 2016.
https://gateways2016.figshare.com/

and upload content using only a web browser, without having to download, compile, or install code on local systems [3-8]. In addition, users can develop and deploy their own interactive tools, powered by the CCR HPC cluster, and share them with the community. VIDIA is supported by CCR, which provides installation of tools and direct support for educators and their courses.

The VIDIA gateway, hosted at CCR, is itself installed on powerful hardware. It runs on a Dell PowerEdge R720xd server, with 2x 6-core Intel Xeon processors, 48 TB of raw disk space, and 128 GB of memory. For users needing more computational power than a local session affords them, VIDIA can coordinate job submission to CCR's academic HPC cluster using HUBzero software called *submit*. The *submit* software enables users to run codes on the CCR cluster, providing them with access to HPC resources that they do not have at their home institutions.

2. Supporting teaching with VIDIA

In the first semester of this collaboration, CCR deployed three open-source analysis tools to serve four social science courses at Oneonta. During the semester, Oneonta students used VIDIA to run more than 600 analysis sessions. The next semester, CCR deployed an additional tool and supported six Oneonta undergraduate courses. Overall, in the four Semesters between Spring 2014 and Fall 2015, VIDIA supported 16 SUNY courses, serving more than 250 students.

Before VIDIA, no social sciences coursework offered at Oneonta utilized computation-intensive text analysis. Now, the VIDIA platform enables Oneonta students and faculty to conduct analyses that were previously impossible at their institution.

The VIDIA platform enables users to focus on the analysis at hand, rather than the details of the computing environment. Student responses to VIDIA at Oneonta suggest that this approach is bearing fruit. Approximately 80% of POLS-200 students were satisfied with the VIDIA platform and the support they were offered. In Fall 2014, 87% of POLS-200 students felt that the Big Data assignment was a worthwhile learning experience. Furthermore, students noted that the platform enabled them to gain valuable experience in new analytical methods and analysis tools. [9]

Social science provided a starting point; now, other disciplines are benefiting from the gateway. CCR has installed new tools and libraries on VIDIA, and assisted instructors from bioinformatics, comparative literature, and philosophy to incorporate VIDIA into their syllabi.

2.1 Tools

The open-source software tools initially deployed on VIDIA were chosen to support data analytics for the social sciences. CCR staff worked with Oneonta faculty to implement the best open-source tools for introducing students to posing and answering data-oriented questions. To ensure that the tools are easily approachable, all offer intuitive graphical user interfaces (GUIs).

Examples of computational tools deployed on VIDIA include:

RapidMiner, a workflow-based environment for machine learning, data and text mining, and analytics. [10]

PSPP, a GNU version of IBM's SPSS software, enables analysis of sampled data. [11]

RStudio, an interactive development environment (IDE) for the R language, offers extensive statistical analysis libraries developed by the R user community. [12]

2.2 Coursework

Oneonta's undergraduate coursework in the Social Sciences uses tools hosted on VIDIA to introduce techniques for analyzing large datasets. Students might start with a Twitter dataset, then use the RapidMiner tool to preprocess it, evaluate document similarities, explore clustering techniques, and visualize the results.

Examples of undergraduate coursework taught on VIDIA include:

POLS-200, Approaches to Political Science, Dr. Bill Wilkerson: Introduces research methods and techniques in Political Science. Assignments require extensive data collection and computation.

POLS-284, U.S. Foreign Policy, Dr. Brett Heindl: Assigns data analysis exercises that explore the influence of social media on foreign policy.

SOCL-260, Social Class, Dr. Brian Lowe: Assigns analysis of large datasets to examine

sociological theories of stratification.

3. Enabling research with VIDIA

Students and researchers also use VIDIA's tools and software for research. Some use VIDIA to launch larger jobs on CCR's HPC cluster, where VIDIA users ran more than 2000 jobs in the first 9 months of 2016. Access to research-quality computational tools enables faculty members to publish and stay abreast of developments in their field; additionally, it helps faculty expose their advanced students to research methods and the current problems and questions that pose challenges in their field.

Examples of research performed on VIDIA include:

Molecular biology: Doctoral student Nicholas Stam, of SUNY Upstate Medical University, analyzes electron microscopy images using *submit* on VIDIA, thus running codes on up to 256 cores on CCR's HPC cluster.

Sociology: Oneonta undergraduate Matt Hartwell used RapidMiner on VIDIA to study the efficacy of Hashtag Activism for his senior thesis. [13]

Political Science: Oneonta professor Dr. Bill Wilkerson uses R on VIDIA with social media data to quantify public opinion about the United States Supreme Court. [14]

Sociology: Oneonta professor Dr. Brian Lowe, uses RapidMiner on VIDIA to investigate controversies in animal rights, using social media data on whaling and marine mammal captivity. [15]

Population Biology: Oneonta professor Dr. Dan Stich analyzes the effects of dams on fish populations, using VIDIA to submit calculations to CCR's HPC cluster. [16-17]

4. Conclusion

The VIDIA gateway was conceived to support data-intensive computation for research and instruction on campuses that lack access to traditional HPC resources. It puts powerful data analytics tools directly into the hands of students, enabling them to rapidly gather and analyze information and master new techniques. Powered by the HUBzero platform, VIDIA helps level the playing field by removing the barrier to data-intensive computing that small campuses face.

5. Acknowledgments

The pilot VIDIA project was funded by a SUNY Innovative Instruction Technology Grant (IITG), 2012, renewed 2013. [18-19] The authors thank Tom Furlani and Bob DeLeon for helpful conversations.

6. References

[1] M. McLennan, R. Kennell, "HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering," Computing in Science and Engineering, 12(2), pp. 48-52, March/April, 2010.

[2] J.l. Palma, l. Courtland, S. Charbonnier, R. Tortini , G.A. Valentine (2014) "Vhub: a knowledge management system to facilitate online collaborative volcano modeling and research." Journal of Applied Volcanology 3:2, doi: 10.1186/2191-5040-3-2.

[3] J.M. Sperhac, J. Greenberg, "VIDIA – A Virtual Infrastructure for Data Intensive Analysis", SUNY Conference on Instructional Technologies, SUNY Geneseo, Geneseo, NY, May 2015.

[4] J. Greenberg, J.M. Sperhac, "Teaching Big Data analysis in the Social Sciences using a HUBzero-based platform." HUBbub, Indianapolis, IN, 29 September 2014.

[5] J. Greenberg, "The Creation of a Big Data Analysis Environment for Undergraduates in SUNY." Interactive Learning Technologies SALT Conference, Reston, VA, 15 August 2014.

[6] B. Lowe, S.M. Gallo, J.M. Sperhac, J. Greenberg, "Providing Undergraduates with a Virtual Infrastructure for Data Intensive Analysis." SUNY Conference on Instructional Technologies, Ithaca, NY, 30 May 2014.

[7] J.B. Greenberg, S.M. Gallo, "Undergraduate Social Science Tools: An Upgrade." SUNY Wizard, Syracuse, NY, 19-21 November 2013.

[8] G. Fulkerson, S.M. Gallo, "Big Data and Social Science: Potential, Goals, and Challenges." SUNY Critical Issues

Conference, New York, NY, 29 October 2013.
[9] W. Wilkerson, "Using 'Big Data' in a Political Science Research Methods Course: A Description and initial Assessment of a Social Media Analysis Assignment." Fifteenth Annual American Political Science Association Teaching and Learning Conference, Washington, D.C. 16-18 January 2015.
[10] rapidminer.com
[11] https://www.gnu.org/software/pspp/
[12] rstudio.com
[13] M. Hartwell, "Exploring Hashtag Activism." New York State Political Science Association Conference, SUNY Plattsburgh, Plattsburgh, NY, April 2015.
[14] W.R. Wilkerson, "Using 'Big Data' in a Political Science Research Methods Course: A Description and Initial Assessment of a Social Media Analysis Assignment", submitted to: Political Science and Politics.
[15] B.M. Lowe, J.B. Greenberg, "Bringing Qualitative Analysis to 'Big Data': Some Preliminary Findings", 32nd Annual Qualitative Analysis Conference, Brescia University College, London, Ontario, Canada, June 24 – 26, 2015.
[16] D. S. Stich, T. Sheehan, and J. Zydlweski. "Projected effects of dam passage performance standards on American shad (*Alosa sapidissima*)". Oral presentation at the Annual Meeting of the New York Chapter of the American Fisheries Society, Cooperstown, NY, 2016.
[17] D. S. Stich, T. Sheehan, and J. Zydlweski. "Assessing the sensitivity of population demographics associated with system-wide changes in passage efficiency for American shad". [In preparation]
[18] Big Data on a Small(er) Campus: Use of Large-Scale Text Analysis by a Comprehensive Primarily Undergraduate Institution (http://commons.suny.edu/iitg/big-data-on-a-smaller-campus-use-of-large-scale-text-analysis-by-a-comprehensive-primarily-undergraduate-institution/).
[19] VIDIA: Virtual Infrastructure for Data Intensive Analysis (http://commons.suny.edu/iitg/virtual-infrastructure-for-data-intensive-analysis-vidia-2/).