Demo: The VIDIA Gateway: a Virtual Infrastructure for Data Intensive Analysis

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: *The Virtual Infrastructure for Data Intensive Analysis (VIDIA) gateway allows campuses with limited computational resources to incorporate data-intensive analysis into their course curricula. We will demonstrate how two powerful open-source tools deployed on VIDIA may be used in the data analytics classroom by beginning or advanced students.*

1. Introduction

Datasets culled from social media or other sources can easily grow to a size that exceeds the analytical capability of common software tools and ordinary desktop computers. Undergraduate institutions often lack the computing infrastructure and support personnel needed to allow students and faculty to create, manipulate, and analyze such large datasets. In order to provide the tools necessary for students to learn data intensive computing and analysis techniques, State University of New York (SUNY) University at Buffalo's Center for Computational Research (CCR) teamed with SUNY College at Oneonta to establish a gateway, Virtual Infrastructure for Data Intensive Analysis (VIDIA), see Fig. 1.

1.1 Collaboration members

The collaboration that created VIDIA is composed of two disparate institutions within the SUNY system: a small college and a flagship research university. SUNY Oneonta's IT infrastructure is typical 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 as well, hampering faculty efforts to bring data analytics to the curriculum.

The University at Buffalo's Center for Computational Research (CCR), a leading academic supercomputing center, has 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. CCR also provides custom software development and data analytics to support faculty research.

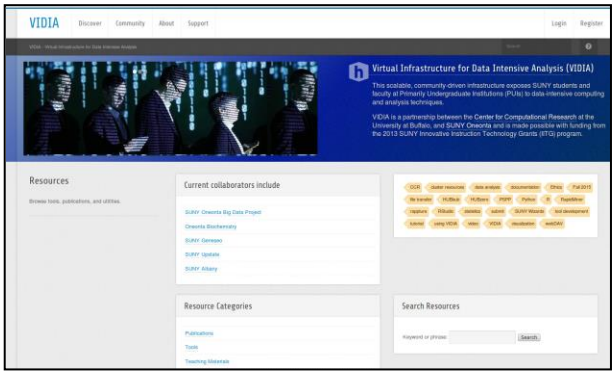

Fig. 1. VIDIA homepage

1.2 Introducing VIDIA

VIDIA, hosted at CCR, is powered by the HUBzero Platform for Scientific Collaboration, originally developed at Purdue University. [1] HUBzero was specifically designed to help a scientific community share resources. Users can upload their own content, launch computations, and view results with a web browser, without needing to download, compile, or install any code. In addition, users can develop and deploy their own software tools, and share them with the community. The tools hosted on the gateway are not just web forms, but powerful graphical tools that support visualization and comparison of results.

The VIDIA gateway 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

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

can coordinate job submission to CCR's academic HPC cluster using HUBzero software called *submit*.

With VIDIA, CCR provides small colleges such as Oneonta, with its limited computing infrastructure, with the capability to analyze large datasets. Using the VIDIA platform, Oneonta has integrated data analytics into Sociology, Political Science and Philosophy coursework. For example, students and faculty capture data using the Twitter Application Programming Interface (API), then analyze it using VIDIA. The VIDIA gateway fosters a collaborative environment where students and faculty can conduct intensive data analysis that is not otherwise possible at their institution. [2-7]

2. Running analysis tools on VIDIA

We have chosen to deploy freestanding Linux desktop applications on the powerful VIDIA server, providing faculty with maintenance-free shared environments for teaching coursework and conducting and guiding research. CCR installs and maintains all tools housed on VIDIA. We work with faculty to find and vet the best open-source tools to support their needs and goals for research and teaching.

The HUBzero platform has proven to be sufficiently robust to support classroom instruction. For example, hands-on VIDIA tool discovery sessions are frequently held in the Oneonta computer labs, with 25 students running concurrent VIDIA tool sessions right in their local browsers. Each virtual container that runs a tool session is preconfigured with all installed software tools and a preset amount of memory, presently 2 GB. Additional tool containers can be configured with different requirements. Each user is allocated a disk quota on VIDIA, and can request more space as needed.

2.1 Installed tools

The current set of installed tools on VIDIA includes:

RapidMiner, a workflow-based environment for machine learning, data and text mining, and analytics. [8]

RStudio, an interactive development environment (IDE) for the R language, offers extensive statistical analysis libraries developed by the R user community. [9]

PSPP, a GNU version of IBM's SPSS software, enables analysis of sampled data. It is written in C and uses the GNU Scientific Library [10]

iPython console, a Qt-based interactive command shell for Python and other languages. [11]

Orange data mining, a workflow-based data visualization, exploration, and analysis environment, can be used in GUI or script mode. [12]

Spyder, an interactive development environment and numerical computing environment for Python. [13]

Workspace, a built-in HUBzero tool, offers a Linux desktop with compilers, scripting languages, utilities, and command line access to the user's home directory on VIDIA.

In this demo we exhibit the functionality of the VIDIA gateway and two of the most widely-used analysis tools in Oneonta's coursework, RapidMiner and RStudio.

2.2 RapidMiner tool

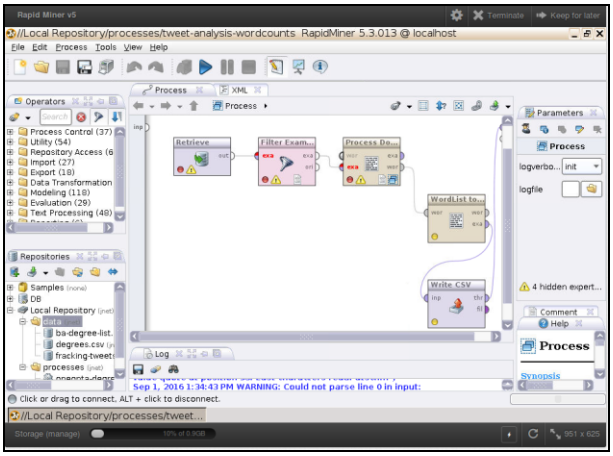

Fig. 2. RapidMiner text analytics workflow

The RapidMiner tool enables students to construct visual workflows that perform text analytics, data mining, or machine learning on their datasets. [8] Students assemble workflows out of operators that represent different parts of the analysis. Underlying the workflows, configuration files govern the computations that each operator performs.

RapidMiner is used extensively for business

analytics, and enables users with limited coding background to create and refine quantitative analyses of large datasets. The version 5.3 release is open-source, and has been a central part of the Oneonta data analytics program for the social sciences. In their assignments, students are provided with example workflows and are asked to adapt them for datasets of their choice. More advanced students construct workflows on their own. Another benefit of RapidMiner is its comprehensive documentation set, which is also deployed on VIDIA alongside the tool.

We will demonstrate the local use of a text analytics workflow using a Twitter dataset, see Fig. 2. The dataset is first processed to remove stop words and punctuation, then word counts are evaluated. The workflow can be successively refined until the analysis yields insights into the data. Finally, result data can be exported, and visualizations performed on the results, using RapidMiner's graphical tools.

2.3 RStudio tool

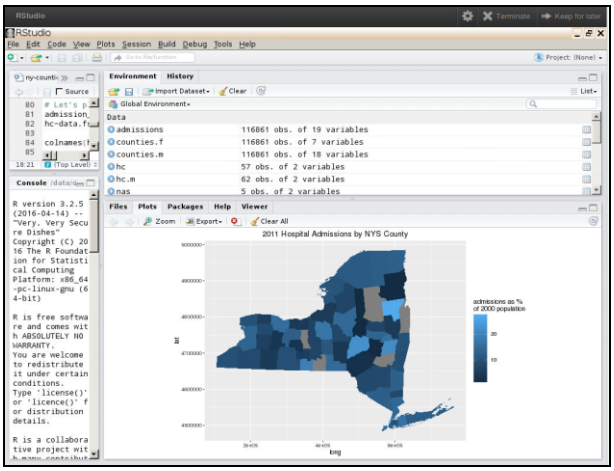

Fig. 3. RStudio GIS visualization

The RStudio tool enables students to learn the functionality of the R statistical language and the community-contributed packages that vastly extend its capabilities. It provides an interactive development environment (IDE) that supports a full-featured editor, a console window, and a display of the variables, objects, and functions that are presently in scope. The IDE also offers R help pages, package loading controls, and plot displays. [9]

We will showcase RStudio on VIDIA by locally loading and processing a 2.5 million record hospitalization dataset, see Fig. 3. We will relate the dataset to a geographic information system (GIS) dataset and visualize the results. [14-15] At CCR, we use this dataset in an annual summer workshop on data analytics and visualization. [16]

3. 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 provides powerful data analytics tools that enable students to rapidly gather and analyze information and master new technologies. Powered by the HUBzero platform, VIDIA helps remove the barrier to data-intensive computing that small campuses face.

4. Acknowledgments

The pilot VIDIA project was funded by a SUNY Innovative Instruction Technology Grant (IITG), 2012, renewed 2013. [17-18] The authors thank Tom Furlani and Bob DeLeon for helpful conversations.

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

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

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

[5] 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.
[6] J.B. Greenberg, S.M. Gallo, "Undergraduate Social Science Tools: An Upgrade." SUNY Wizard, Syracuse, NY, 19-21 November 2013.
[7] G. Fulkerson, S.M. Gallo, "Big Data and Social Science: Potential, Goals, and Challenges." SUNY Critical Issues Conference, New York, NY, 29 October 2013.
[8] rapidminer.com
[9] rstudio.com
[10] https://www.gnu.org/software/pspp/
[11] https://ipython.org/
[12] http://orange.biolab.si/
[13] https://pythonhosted.org/spyder/
[14] New York State Statewide Planning and Research Cooperative System (SPARCS): https://www.health.ny.gov/statistics/sparcs/
[15] SPARCS 2011 dataset: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/pyhr-5eas
[16] The Eric Pitman Annual Summer Workshop in Computational Science: http://www.buffalo.edu/ccr/outreach/k-12-outreach/summer-workshop.html
[17] 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/).
[18] VIDIA: Virtual Infrastructure for Data Intensive Analysis (http://commons.suny.edu/iitg/virtual-infrastructure-for-data-intensive-analysis-vidia-2/).