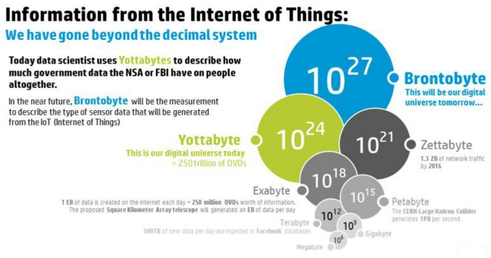
2013 SUNY Critical Issues in Higher Education Conference

Building a Smarter University: Big Data, Innovation and Ingenuity

Oct. 29-30, 2013

Dr. Brian Lowe, Dr. Greg Fulkerson, Dr. Brett Heindl, Dr. Achim Koedderman, Mr. James Greenberg

Introduction



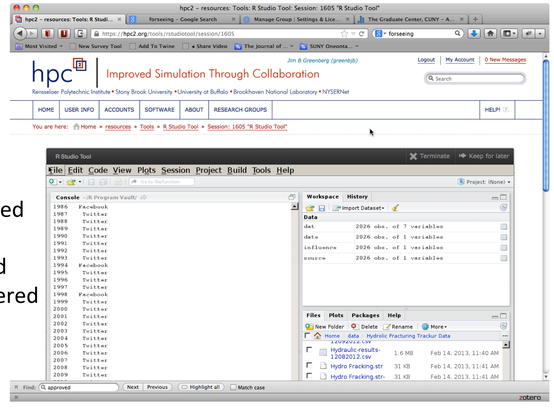
The term "Big Data" typically refers to datasets so large they are difficult to manage and analyze using traditionally available tools such as relational database management systems (RDBMS), statistical software such as R or SPSS, or Microsoft Excel. In fact, big data is more than just the amount of the data – it also includes the need to perform complex analytics on the data beyond the "advanced counting techniques"[1] provided by standard RDBMS, including trend analysis, clustering, and principal component analysis (PCA). Primarily undergraduate institutions (PUIs) such as SUNY Oneonta typically have neither the computing and networking infrastructure, nor the support personnel, needed to manage and do the analysis of these large multi-terabyte datasets. This poster outlines the Information Technology Components of a collaborative project between SUNY Oneonta and the Center for Computational Research (CCR) at the University at Buffalo to provide instructors and their students at SUNY Oneonta with readily deployable pedagogical and research strategies for grappling with big data in their respective disciplines. In essence, this project's plan is to infuse a "highlight reel" of data science into the undergraduate social science programs (sociology, political science, and philosophy) at SUNY Oneonta. This poster highlights three of the four major components of the project: 1) the software used to create and analyze social media data from multiple sources; 2) the information technology (IT) infrastructure and software environment created to support the creation and analysis of these large datasets; and 3) the collaborations the project has established with its business partners.

Component One – Software to Create and Analyze Large Data Sets from Social Media

Big data can best be characterized by the three "V" components: Volume (the sheer amounts of data available), Velocity (the need to process large numbers of data items in a relatively short time period, even as they change), and Variety (multiple types of data taken from many sources).With support from a 2012 Tier 2 SUNY Innovative Instruction Technology (IITG) grant, SUNY Oneonta explored and evaluated the use of numerous data capture software packages (including Discovertext, GNIP, Trackur, ContentSeer, and Topsy), eventually deploying Trackur because of its low cost, ease of use, and ability to export datasets into Excel format. During the fall 2012 and spring 2013 semesters, faculty at SUNY Oneonta integrated use of Trackur into courses in sociology, political science and philosophy, centered upon a social-scientific examination of how moral claims and discourses are created, sustained, altered and challenged within the electronic sphere. Faculty members captured and analyzed data using locally available software (Excel, SPSS) as well as software applications written in-house that access the Twitter Application Programming Interface (API). Given Trackur's limitations (access to only 3,000 records at once), the capture of enough data to monitor social media trends was identified as a barrier to progress (also noted under Component Three below). In spring 2013, the College expanded its capacity for data capture and analysis to a senior sociology seminar and two political science courses, using both "home-grown" software and a donation by IBM of its MAP suite of tools.

TOPSY
Real-time search for the social web

Component Two – Virtual Infrastructure for Data Intensive Analysis (VIDIA): System-Ness & Synergy



Having overcome the initial hurdle of lack of tools and training for dataset analysis, faculty members faced their next hurdle: the limitations of the College's current technology infrastructure. SUNY Oneonta's existing IT infrastructure is typical of the SUNY comprehensive colleges: total storage shared by its ~6,000 student and 900 faculty/staff users is currently 4 TB, with available software limited to the standard set of Windows and Macintosh applications, including SPSS, R, Minitab, Atlas.ti, and SAS. To overcome this limitation, the College partnered with CCR on a 2013 Tier 3 IITG grant titled "Virtual Infrastructure for Data Intensive Analysis (VIDIA)." CCR, a leading academic supercomputing facility, maintains over 8,000 processing cores and 500 TB of storage, and has extensive experience both with the development of virtual organizations (including the VHub volcanology community and the New York State High Performance Computing Consortium, or HPC[2]), as well as analysis tools for high-performance computing users. The IITG proposal has been funded and is in its initial planning stages. As part of the VIDIA project, CCR and SUNY Oneonta will develop a scalable, community-driven infrastructure to expose students and faculty at SUNY Oneonta (and eventually other PUIs) to data-intensive computing and analysis techniques. The environment, not typically available to PUIs, will include an initial set of open-source data analysis tools, storage space, and seamless access to CCR's HPC facilities for analysis. In addition to deploying the environment, CCR staff will train SUNY Oneonta educators in utilizing the platform. While we will deploy an initial set of high-priority tools, tested by SUNY Oneonta social science faculty, eventually, the broader educational community will be encouraged to provide tools it has developed, to provide content, and to utilize the environment for courses and workshops.

Component Three – Curriculum Integration (see faculty posters)

Component Four– Business Collaborations

Business collaborations have been critical to the success of the project. IBM Corporation has provided SUNY Oneonta with complimentary academic use of its Modeler Premium software, which has given the faculty (and students, beginning in the fall of 2013) access to a comprehensive set of data-mining and text-analysis tools that the campus would not have otherwise been able to provide. Microsoft's FUSE Labs has also granted access to its collection of social media data; finally, Content Savvy has made its ContentSeer product available to the project team at deep discount. These tools, and the support of these companies, are the foundations of this effort.

IBM

DELL

CCR Center for Computational Research
www.ccr.buffalo.edu