10-2012

An Infrastructure to Support Data Integration and Curation for Higher Educational Research

Linh B. Ngo
Clemson University, lngo@clemson.edu

Amy W. Apon
Clemson University, aapon@clemson.edu

Pengfei Xuan
Clemson University

Kimberley Ferguson
Clemson University

Christin Marshall
Clemson University

See next page for additional authors

Recommended Citation

Ngo, Linh B.; Apon, Amy W.; Xuan, Pengfei; Ferguson, Kimberley; Marshall, Christin; McCann, John; and Zheng, Yueli, "An Infrastructure to Support Data Integration and Curation for Higher Educational Research" (2012). *Publications* . 5.
https://tigerprints.clemson.edu/computing_pubs/5

Authors
Linh B. Ngo, Amy W. Apon, Pengfei Xuan, Kimberley Ferguson, Christin Marshall, John McCann, and Yueli Zheng

An Infrastructure to Support Data Integration and Curation for Higher Educational Research

Pengfei Xuan, Kimberly Ferguson, Christin Marshall, John McCann, Linh B. Ngo, Yueli Zheng, and Amy W. Apon
School of Computing, Clemson University, Clemson, South Carolina, USA

***Abstract*—The recent challenges for higher education call for research that can offer a comprehensive understanding about the performance and efficiency of higher education institutions in their three primary missions: research, education, and service. In other for this to happen, it is necessary for researchers to have access to a multitude of data sources.However, due to the nature of their academic training, many higher education practitioners do not have access to expertise in working with different data sources. In this work, we describe a design and implementation for an infrastructure that will bring together the tools and the data to provide access to researchers in the field of higher education institutional research. The infrastructure will include integration and curation for data from different sources, embedded statistical environment, high performance computational back-end, and extensibility for future Big Data and unstructured data.**

***Keywords*-higher education, social science, data integration, data curation**

I. Introduction

Higher education institutions are challenged with increased competition, fiscal difficulty, increased demands for accountability, expansion of diverse needs from the student bodies, and opportunities and difficulties in pervasive new technologies [1]. These challenges call for research that can offer a comprehensive understanding about the performance and efficiency of higher education institutions in their three primary missions: research, education, and service. In other for this to happen, it is necessary for researchers to have access to data that goes beyond describing typical educational characteristics. There exists a large variety of data sources describing not only educational but also different social statistics on the Internet. However, utilizing data from different sources that describe different aspects of institutional research could be challenging, even when the data is publicly accessible. The lack of availability and accessibility of data is noted in [2].

A. Motivation

The design for this infrastructure is motivated through the analytical work in [3]. In this work, the authors analyze data from different sources in order to gain insights to the impacts of investment in high performance computing infrastructure upon research productivity of higher education institutions. In the study, there are approximately 190 institutions, and the aggregated institutional data is roughly 5Gb and comes from multiple sources. The process of transforming the data into a consistent view so that analysis can be performed is difficult and time consuming. As both the analytical work is enhanced and the amount of data is increased, runtime of statistical methods will take longer. For example, a data envelopment analysis [4] that is used for the evaluation of institutional productivity takes one minute for a data set of an academic department with two input and one output variables for 180 institutions. To gain a better statistical evaluation, it is necessary to run this analysis for a much larger set of institutions (roughly 4000 institutions in the U.S.) and across many different sets of input and output variables. Therefore, it is useful to have an infrastructure that could provide higher educational researchers with a unified view of the entire data regardless of sources as well as a strong computational support for complex and time consuming statistical analysis.

B. Vision

Our vision for this work is to design an infrastructure that will bring together the tools and the data and provide access to the researchers in the field of higher education institutional research. Our work improves upon these approaches through the followings:

- Mechanisms to curate and integrate data from different sources.
- Interface to a statistical framework to provide embedded complex statistical functions.
- Back-end connection to allow integration with high performance computing infrastructure.
- Open design to allow future extension and integration with infrastructure that supports big and unstructured data.

This work is based on the foundation of previous data framework that was built on proprietary tools [5]. We extend the previous work through the new implementation using open source platform and the extension of the infrastructure to incorporate Big Data capability and high performance computational backend.

II. Architectural Design

The ecosystem of higher education institutions, in both educational and research aspects, is complicated. Knowledge is intangible, and the effects of knowledge creation and exchange usually take a long period (years) to make visible impacts. Higher educational data contains complex associations, a very large number of variables to be considered, and has the potential to become Big Data. the general principles of system design are closely followed by these requirements:

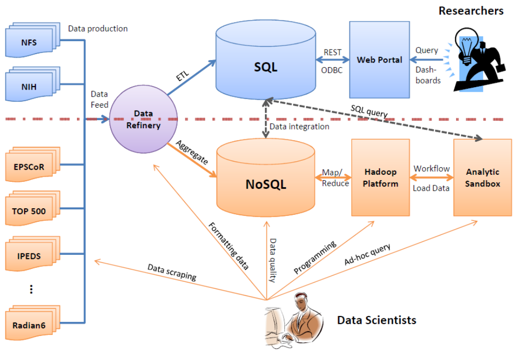

Fig. 1: Infrastructure to Support Higher Educational Research

- An easy-to-use interface with the lightweight platform-independent deployment on the client side.
- Built-in advanced analytic modules that enable users to perform fast, ad-hoc investigation and answer their own questions.
- The infrastructure should provide embedded statistical functions as well as allow custom statistical codes.
- Massive and multi-structured data sets support.
- High-level flexibility, availability, and scalability by using high performance computing and cloud computing resources.

The analysis of higher education data involves not only large volumes of multi-structured data sets but also many different technologies, specifically Hadoop based ecosystem. The emerging of Big Data paradigm in recent HED research requires the new generation computing infrastructure with offering both high capability and scalability [6]. Figure 1 illustrates the perceived analytics ecosystem for HED research that supports multi-structured data sets processing and analyzing. The blue area shows the relational structured data environment under the relational database management system (RDBMS) and Hadoop-based Big Data analysis environment for unstructured data.

A. Structured data processing (SQL):

All structured data sets (e.g. NFS, HIH data sets or refined data sets from the NoSQL environment) are properly normalized and imported to the RDBMS environment. Typically, the researchers could easily retrieve information via user-friendly Web GUI-based applications without the need for the in-depth data analysis knowledge, programming skills, or database schema background. In our current design, we select MySQL as the RDBMS, Drupal as the content management system (CMS) and R as the statistical programming environment.

B. Unstructured data processing (NoSQL):

The NoSQL environment supports large-scale and complex Ad-hoc queries that provide the data scientists with the in-depth exploration and study of underlying raw data sets from varieties of structures and data sources. The data scientists are able to completely sidestep the schema constraints under the structured data and freely develop data processing programs to explore, analyze, and model data in analytic sandboxes. The aggregated analysis results from NoSQL computing environment can be shared from the front-end through the polling of results to the production SQL querying environment. In our current design, we select Hadoop-based ecosystem as our NoSQL Big Data processing platform. This includes the usage of Pig Latin for expressing data analysis, HBase as the database for unstructured data, Oozie as the workload engine for analysis tasks, and Hadoop HDFS and MapReduce as the underlying distributed computing infrastructure [7]. The combination of SQL and NoSQL expands the horizon for analytical inquiries, providing exceptionally high price-performance and new, flexible data processing framework for our ongoing study in higher educational research.

III. Initial Implementation and Future Work

With the assistance of REU students Ferguson, Marshall, and McCann, we have completed the initial implementation of the infrastructure including the installation of the structured data processing mechanisms. As discussed in Section II, the structured data processing architecture mainly consists of MySQL, Drupal, and R. With this implementation, we were able to recreate many of the previous analysis in [3] as well as enhance these analysis with additional data. Our next steps involve both the implementation of the non-structured data processing infrastructure as well as the continuous addition of more data sources. This implementation will be wrapped in virtual machines for portability and tested through the resources of the ExTENCI project [8].

This work was supported in part by NSF Grant #1212680. We acknowledge the work of Vijay Dantuluri, Michael Stealey, and Stanley Ahalt from RENCI in the acquisition and ingestion of a significant part of the data repository.

References

[1] J. M. Gappa, A. E. Austin, and A. G. Trice, *Rethinking faculty work: Higher education's strategic imperative*. San Francisco: Josey-Bass, 2007.

[2] R. K. Toutkoushian and K. Webber, *University Rankings: Theoretical Basis, Methodology and Impacts on Global Higher Education*. Springer, 2011, ch. Measuring the Research Performance of Postsecondary Institutions.

[3] A. Apon, S. Ahalt, V. Dantuluri, C. Gurdgiev, M. Limayem, L. Ngo, and M. Stealey, "High performance computing instrumentation and research productivity in u.s. universities," *Journal of Information Technology Impact*, vol. 10, no. 2, 2010.

[4] R. W. Shephard, *Theory of Cost and Production Functions*. Princeton: Princeton University Press, 1970.

[5] L. Ngo, V. Dantuluri, M. Stealey, S. Ahalt, and A. Apon, "An architecture for mining and visualization of u.s. higher educational data," in *Information Technology: New Generations (ITNG), 2012 Ninth International Conference on*, april 2012, pp. 783 –789.

[6] L. Soares, "The rise of big data in higher education," http://www.educause.edu/library/resources/rise-big-data-higher-education, 2012.

[7] T. White, *Hadoop: The Definitive Guide*. O'Reilly Media Inc., 2010.

[8] P. Avery et. al., "Extending Science Through Enhanced National Cyber-Infrastructure," https://sites.google.com/site/extenci/, 2012.