An infrastructure to support data integration and curation for higher educational research

Linh Bao Ngo, Kimberly Ferguson, Christin Marshall, John McCann, Pengfei Xuan, Yueli Zheng, and Amy 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. The design is implemented using a traditional client-server model and evaluated through a number of descriptive studies.**

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

I. Introduction and Motivation

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. A significant number of these data are collected and stored by government entities such as the National Center of Educational Statistics (NCES), the National Science Foundation (NSF), and the U.S. Census Bureau or private entities such as College Board [2], Thompson Reuters' publication data [3], and US News & World Report [4]. However, utilizing data from different sources that describe different aspects of institutional research could be challenging, even when the data is publicly accessible. This lack of availability and accessibility of data is noted in [5]. Several framework such as NSF's WebCASPAR [6], NCES, or the Census Bureau provide tools to aggregate the data, and to some extend, to perform simple data analysis. Another area of research that also contributes to the storage and maintenance of data repository is library science. The Institution for Social and Policy Studies (ISPS) at Yale University maintains an open access digital collection of social science experimental data and metadata as well as the related processing codes produced by ISPS researchers [7].

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 in order to provide users with a bigger picture across different social aspects.
- Interface to the R statistical framework to provide embedded complex statistical functions as well as allow users to utilize their own codes.
- Back-end connection to allow future integration with high performance computing infrastructure to facilitate computational and data-intensive calculations.
- Open design to allow future extension and integration with infrastructure that supports big data and unstructured data.

This work is based on the foundation of previous work that was built on proprietary tools [8]. This limits access to only people with the proper license or on a specially configured hardware platform. We address this limitation through the use of virtual machine implementation that is portable across different hardware and open source implementations. A significant portion of the work were done by the undergraduate computer science students that participate through the Research Experience for Undergraduates (REU) program by NSF. This contributes to the students' experience in working with data and tools from another discipline.

The remaining of the paper is structured as followed. Section III provides an brief overview on the typical data sources used in this work. Sections II describes some sample use cases in the area of research in higher education and the subsequent hypothesis on user requirements for the infrastructure. This section also describe the initial client-server design for the infrastructure and the implementation of this design. Section III discusses the data sources, how they are ingested into the infratructure, as well as provide several sample descriptive analyses and visualization. Section IV summarizes the paper and discuss the future work.

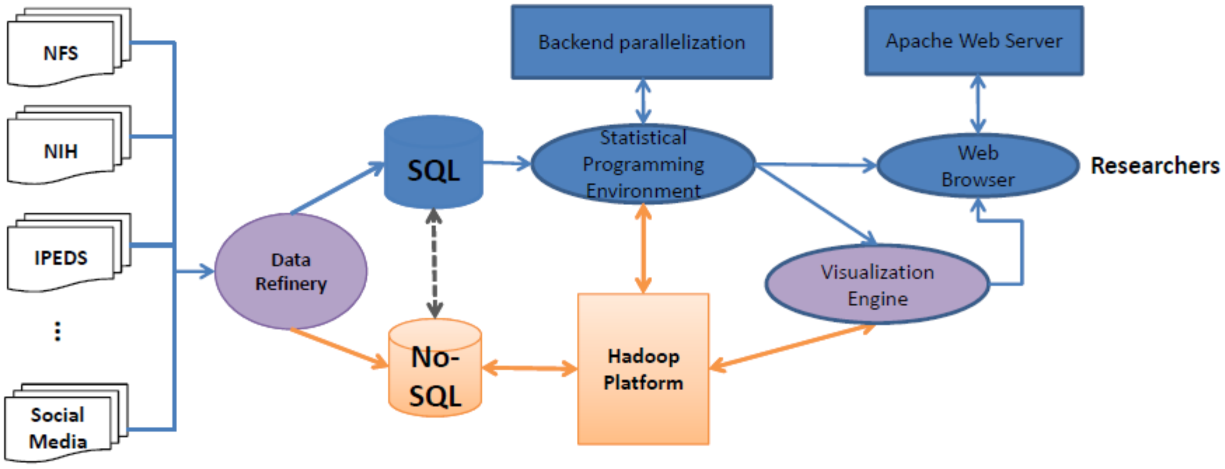

Fig. 1: *Initial client-server design*

II. Initial Design

In earlier work, [8] presents a framework to facilitate efficient extraction and analysis of large scale data. The steps described serve as a model for the implementation of this project. Furthermore, this undertaking seeks to present itself as an actual example in which the framework is employed. The framework exists as an applicable and easy-to-implement approach for the purpose of aggregating and curating data from different sources. The collection of data carries a high degree of variance in format. For instance, certain data sets designate institutions' names as the primary key, whereas other data sets assign each institution a unique ID that serves as the primary key. Moreover, misspellings occur and the naming of institutions differs from source to source due to the human factor. As an example, some sources refer to the Massachusetts Institute of Technology by its full name while other sources refer to it by its abbreviated name, MIT. These discrepancies present obstacles to data analysis and necessitate an organized and well formed process for the correction of such inconsistencies. Another challenge is the difficulty in being able to perform a timely analysis on raw and unstructured higher educational data. To the researchers, they are not always able to immediately identify relationships among the appropriate combination of data fields necessary to answer the science question. While directional uncertainty may be considered inevitable in any research undertaking, it only becomes more and more complex as datasets grow larger. Also, under most circumstances, it is time consuming for the researcher to manually sift through data of this magnitude. In the following sections, an conceptual explanation details the process in which a seemingly incomprehensible dataset is transformed to more suitably present itself to the researcher. While the Unified Data Framework allows for a smooth transition from data ingestion and organization through presentation and analysis, it was built on a set of proprietary solutions and it lacks the inclusion of a statistical environment that can supports complex analysis. In our initial design, we utilize a tradition client-server model that is built upon well known open source components.

A. Use cases and user requirements

Before discussing the initial design, we first look at two sample use cases and then hypothesize a set of user requirements based on these use cases.

1) Investment in high performance computing (HPC) systems: This use case investigate the value realized through investments in high performance computers as quantified by research productivity [9]. An academic institution's investment in HPC systems is measured by entries by that institution on the Top 500 HPC list [10]. Research productivity is represented by publication count and additional external fundings. This use case requires data from the National Science Foundation, the National Institute of Health, the Institute for Scientific Information, the Top 500 HPC List, the Carnegie Foundation, and the Integrated Postsecondary Education Data System. The analysis is done using a longitudinal data analysis approach to find a statistically significant relationship between the HPC investments and the research productivity. With additional data, this study could be extended to analyze the relationship between investment in cyber-infrastructure of institution and the surrounding economic and academic environments.

2) NSF's Experimental Program to Stimulate Competitive Research (EPSCoR): Another use case that we are going to demonstrate as an example is an analysis of the effects of NSF's EPSCoR program on institutions' research productivity and capability. EPSCoR is an effort by NSF since 1980 to provide additional support to a number of states that are lagging behind in term of research competitiveness (as represented by portion of NSF funding aggregated across the entire state). A number of work has been done on studying the effect of EPSCoR, however, the literatures are limited and have not examined the increase in competitiveness of EPSCoR states [11]. An in-depth study of EPSCoR effect will benefit from a diverse set of data that could keep track of different aspects of an institution over time and allow a timely analysis among characteristics such as publication, federal fundings, state fundings, etc ...

3) User requirements: 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. In the use cases described previously, the aggregated collection of data covers roughly 200 higher education institutions between the years 1993 to 2009. The total size of data is approximately 4GB in text and spreads across 128 tables, some of which can have up to 400 columns. If we want to look at all the accredited degree-granting institutions in the U.S., the data size can scale up to hundreds of GBs. This is not counting sources such as GIS data and Census data, which can be used to enhanced the complexity of the studies. In other words, higher educational data contains complex associations, a very large number of variables to be considered, and has the potential to become Terabyte-scale data. Therefore, we hypothesize the following requirements to be the foundation of our conceptual design:

1) *Web-based Interface*: An analytical environment will be embedded within a web interface. This is to minimize the amount of efforts it takes to begin taking advantage the infrastructure for users without formal background in computer.
2) *Convenience*: The infrastructure should provide the user with capabilities to do both data querying and statistical analysis.
3) *Extensibility*: While the infrastructure allows users to run their own statistical programs, it should also implement a number of common statistical analysis such as correlations, regressions, mapping, etc ... The implementation of additional analytical methods should require little effort beyond implementing the analytical algorithm itself.
4) *Adaptability*: The infrastructure needs to support both structured and unstructured data in order to take advantage of the recent deluge of social media data for research in higher education.
5) *Scalability*: There should be a high performance computing back-end that would allow users to scale up in term of data size or computational power.

B. Architectural Design and Baseline Implementation

In the initial design of the infrastructure, we follow a traditional client-server model. Figure 1 illustrates an architectural realization of this design. The data is aggregated into the infrastructure through a data refinery mechanism. In an ideal situation, this process is fully automated. However, the current set up only allows for a semi-automated process, as manual efforts must be made to integrate and curate the data. After the refinery process, the data is separated and integrated into two different storage mechanisms for structured data (SQL: statistical data from NSF, IPEDS, ...) and unstructured data (NoSQL: social media data). The embedded statistical programming framework will allow researchers to access the SQL database directly and the NoSQL database through the implementation of the HDFS to support large scale social media data for statistical analysis and visualization of data. The statistical programming framework supports parallelization so that it can be interfaced with a back-end high performance computing environment. The researchers gain access to these data and tools through an web browser. This removes the hardware limitation on the client's side. The contents of the web page can be dynamically configured through the usage of a content management system. Each of these components is implemented within its own KVM virtual machines, and this allows portability and duplication for performance purposes. Among the components of this infrastructure, the SQL Server, the statistical programming environment, and the content management system that controls the web pages form the baseline implementation. For the choice of open source components, we select MySQL as the SQL server, Drupal as the content management system (CMS), and R/Revolution R packages as the statistical programming environment.

- *SQL Server*: Our original database system uses Microsoft SQL Server and is migrated to MySQL server. There are several considerations in making the choice for MySQL. Aside from the fact that MySQL is open source, it is also among the world's most popular open source database management systems, and widely used by many open source projects both academic and industry areas. MySQL is under GPL open source license and has a high-flexibility to integrate with other open source components.
- *Statistical Programming Environment*: R [12] is chosen to be the statistical programming language to use in the analysis. However, the core R program does not provide adequate support on parallel programming, non-sql data access, and web execution. Instead, an open source industrial framework of R, created by Revolution R [13], is implemented. This framework has most, if not all, of the well known R packages that support parallelization installed. For example, there is *multicore* to support multicore programming, *rmpi* for MPI programming, and *rhdfs*, *rhbase*, and *rmr* to support Hadoop/MapReduce. Furthermore, the RevoDeploy framework by Revolution R allows researchers to call on embedded R scripts as well as run their own R source code through a web interface. The majority information requested by the researchers contains in the underlying SQL/NoSQL system and can be accessed through R's function calls.
- *Content Management System*: Drupal is popular and well-supported CMS that has been used to construct a wide variety of websites from small personal web site to enterprise-level web portal. The Drupal website maintains a repository of thousands of user-contributed

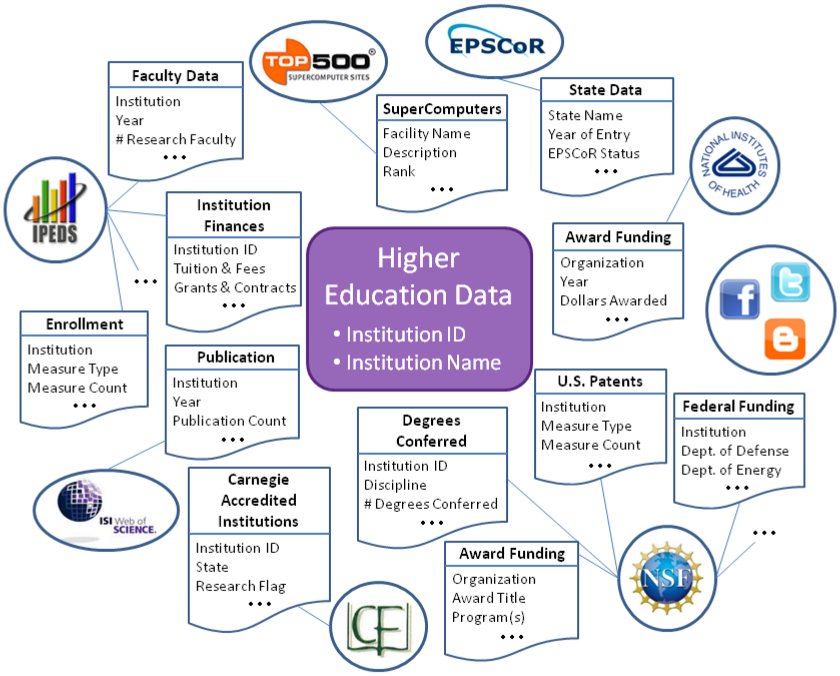

Fig. 2: *Common Data Sources for Research in Higher Education*

extensions (or modules) that can be easily expanded for the functionality of our HDE website. Documents, books and Drupal community are well constructed and active. A well-documented Application Programming Interface (API) is available for programmers to extend the functionality of the software by creating new modules and themes. Drupal could dramatically reduce the life cycle of system development by simplifying site construction and management for our new HDE platform. Furthermore, Drupal is the supported CMS at Clemson University, and we have had experiences in implementing MySQL and Drupal. Typically, the researchers could easily retrieve information via user-friendly Web GUI-based applications without the need for the in-depth data analysis knowledge or database schema background in order to run their own analysis on our server's hardware.

Because all of the components of the baseline implementation are mature and open source, the process of implementation is not very difficult. This is important, as it means that others can duplicate this work easily. One of the few initial difficulties is the different version of apache tomcat supported. The latest tomcat version is 7.0, but the RevoDeploy server version 5.0 we use has tomcat 6.0 embedded in it. This problem has been resolved with RevoDeploy version 6.0 which supports tomcat 7.

III. Data Descriptions and Sample Analysis

In this section, we provide an overall perspective about the data sources that are integrated and curated in our work. In addition, we will also demonstrate a number of descriptive analysis across the data sets.

A. Data Descriptions

The incoming support data sets for higher education data (HED) research are obtained from multiple sources with varying degrees of quality and reliability. A more detailed description of these data sources can be found in [8]. The data sources are aggregated together through the unique institutional identifier (*UnitID*) that is specified by IPEDS and used in the Carnegie Foundation data. This allows the researchers to perform research and analysis across institutions of higher education, and also true when comparing data across sets of institutions with similar characteristics, including data about students, research outcomes, degree programs, and other features of institutions. The attribute that identify the name of the institution, usually called *Institution Name* by a number of different sources, is also used in the aggregation process since sources such as NSF and NIH do not provide the *UnitID* data. Figure 2 illustrates the overall collection of data sources available in the infrastructure. On a side note, while both Carnegie Foundation and IPEDS are considered core sources, the Carnegie Foundation was used as the standard from which institutional names are derived. This is due to the fact that the source at Carnegie Foundation can be accessed more conveniently and contains fewer additional attributes than IPEDS, which makes it easier to update and maintain.

The organization of these data sets could be characterized by either SQL or NoSQL data model. The stage of higher education data collection involves many large and complex data problems emerging from its relevant research domains. To simplify data processing and improve data quality, several data collection pipelines are developed and integrated to our HED analytics platform. In this stage, the pipelines scrape, capture, format, and store raw data sets that originate from different data sources. The purpose of this stage is to reduce the data size by filtering noise or by indexing, summarizing, or marking it up so that downstream analytics can manipulate it more efficiently. An example of the pipeline is the ingestion of data sets from IPEDS. After manually exploring and studying the structure of this website, we find there is no central storage location or API to download the exact data sets that we need. Instead it is a number of compressed files distributed through separated HTML pages. However, we can identify the data storage patterns by studying individual HTML pages, which gives us the initial rule for data scraping in the next step. The data from IPEDS also is accompanied by dictionaries that provide information on column's name and formats. This helps automate the data ingestion process. After extracting the compressed files, we need to validate the correctness of original data sets by its own properties. For example, the format of the column headers are specifically formatted characters. There is not an identical length for the numbers in the same column, which might be caused by missing zeros. We create a Java program to automatically validate these original data sets by the different regular expressions. Finally, the normalized data sets are generated to support the Database import. Finally, the database scheme is automatically constructed by SQL statements that are generated by our Java program. To improve the data import performance, we use MySQL built-in data manipulation statement (LOAD DATA INFILE). This MySQL code allowed us to direct the filter to a specific file and filled in the table upon execution.

B. Sample Descriptive Analysis

We provide several sample descriptive analyses that utilize data from different aggregated sources (NSF, IPEDS, NIH, and ISI) to demonstrate the usability of this infrastructure. The data used in these examples only need to be extracted once through the *RMySQL* connector. The analyses are performed using R's embedded correlation function and the open source packages *plm* for panel data analysis and *Benchmarking* for data envelopment analysis.

Correlation Analysis

In the first example, the relationships between several parameters of the different data sets are analyzed using correlation analysis. Table I shows the correlation coefficients for each pairs of parameters. From this table, several observations can be made. First, the number of undergraduate enrollment has little relation with fundings and publications, but it correlates with the number of graduate enrollment. In turn, graduate enrollment slightly correlates with the number of publications. While the number of publications is highly correlated with the sum of all federal fundings, the degree of correlation is highest with NSF fundings. Between the funding sources, NSF and DOE has slightly significant correlation strength, which indicates similarity in institutions that funded by these agencies.

Panel Data Analysis

In the second example, a panel data analysis (longitudinal study) is performed on the following parameters: publication counts, NSF, DOD, DOE, and NIH funding, and EPSCoR status. Panel data analysis allows researchers to look at "repeated observations on the same cross section" of the same set of institutions over time [14]. Previous work indicates a high degree of endogeneity between publication counts and NSF funding; therefore we utilize a 2-stage-least-square (2SLS) estimation method using the number of undergraduate enrollment as an instrument variable. This is due to the fact that the number undergraduate enrollment has almost no correlation with either publication counts or any of the funding sources. We assume that funded project does not produce publication until at least 6 months since the award date, and the publication date usually takes another several months (depending on the type of conference or journal). Therefore, the formula is set up as followed:

- Dependent variable: Publication count
- Independent variable: NIH, DOD, DOE, and NSF funding amounts and EPSCOR status
- Instrument variable for NSF: number of undergraduate enrollment

The length of time is from 1997 to 2007. The EPSCOR status variable is defined as followed. If an institution has never been an EPSCoR state, the value is -1. For EPSCoR institutions, for every year since the year an institution joins the EPSCoR program, the variable is incremented by 1.

Table II, III, and IV show the analytical results for the three cases: all institutions, EPSCoR institutions only, and non-EPSCoR institutions only. It is observed that the publication count of EPSCoR institutions is under significant effect from NIH, NSF funding, and EPSCoR status, but not from DOD and DOE fundings. On the other hand, the two analyses for non-EPSCoR institutions and all institutions show significant effects of all funding sources upon publication counts. In the case of all institutions, the effect of the EPSCoR status variable is not statistically significant.

DEA Analysis

The third example is a data envelopment analysis (DEA). DEA is a non-parametric technique originated from operational managements science, which means that it does not suffer from endogeneity and problem with assumptions that parametric techniques such as panel data analysis have. In a nutshell, DEA calculates the ratio between input (resources) and output(productions) and determines whether a firm is efficient in producing the most out of what resources it has. In this example, we use funding sources (NSF, NIH, DOD, and DOE) and number of graduate enrollment as inputs, and publication count as output. The efficiency score is calculated for the years 1997, 2000, 2003, and 2006, and is shown in Table V.

Graphical Representation using GIS

TABLE I

Correlations between different funding sources (NSF, DOD, DOE, NIH, Federal Funding), publication counts, and student enrollments

	Undergraduate	Graduate	Publication	NSF	DOD	DOE	NIH	All Federal Funding
Undergraduate	1.00	0.63	0.20	0.18	0.01	0.08	-0.01	0.09
Graduate		1.00	0.47	0.20	0.26	0.15	0.09	0.39
Publication			1.00	0.51	0.38	0.45	0.36	0.85
NSF				1.00	0.17	0.36	0.05	0.48
DOD					1.00	0.17	0.24	0.67
DOE						1.00	0.16	0.43
NIH							1.00	0.43

TABLE II

Example Longitudinal Analysis for NSF Funding Impact on Publication Counts for All Institutions

2SLS with fixed effects	Dependent variable	Number of observations	Number of groups	R-squared	Adjusted R-squared
	Publication Count	1649	185	0.26993	0.23883
	Coefficient	Std. Errors	T	$P > \lvert t \rvert$	95 % Confidence Interval
NIH(L1)	2.77^{-3}	2.16^{-4}	12.79	$< 2.2^{-16}$	
DOD(L1)	8.93^{-3}	1.13^{-3}	7.88	6.12^{-15}	
DOE(L1)	2.21^{-2}	4.57^{-3}	4.84	1.41^{-6}	
NSF(L1)	1.28^{-2}	2.69^{-3}	4.75	2.19^{-6}	
EPSCoR Status	16.26	6.6386	2.45	0.01439	

F(5,1459) = 182.202 $Prob(> F) < 2.22^{-16}$

TABLE III

Example Longitudinal Analysis for NSF Funding Impact on Publication Counts for EPSCoR Institutions

2SLS with fixed effects	Dependent variable	Number of observations	Number of groups	R-squared	Adjusted R-squared
	Publication Count	457	51	0.41683	0.36575
	Coefficient	Std. Errors	T	$P > \lvert t \rvert$	95 % Confidence Interval
NIH(L1)	1.87^{-3}	2.87^{-4}	6.53	1.88^{-10}	
DOD(L1)	8.18^{-4}	1.46^{-3}	0.55	0.57	
DOE(L1)	1.34^{-3}	3.58^{-3}	0.37	0.70	
NSF(L1)	7.34^{-3}	4.71^{-3}	1.55	0.11	
EPSCoR Status	28.77	4.2382	6.78	4.06^{-11}	

F(5,401) = 56.889 $Prob(> F) < 2.22^{-16}$

TABLE IV

Example Longitudinal Analysis for NSF Funding Impact on Publication Counts for Non-EPSCoR Institutions

2SLS with fixed effects	Dependent variable	Number of observations	Number of groups	R-squared	Adjusted R-squared
	Publication Count	1192	134	0.26881	0.23769
	Coefficient	Std. Errors	T	$P > \lvert t \rvert$	95 % Confidence Interval
NIH(L1)	2.80^{-3}	2.61^{-4}	10.73	$< 2.2^{-16}$	
DOD(L1)	9.46^{-3}	1.37^{-3}	6.89	9.49^{-12}	
DOE(L1)	2.59^{-2}	5.97^{-3}	4.34	1.56^{-5}	
NSF(L1)	1.33^{-2}	3.07^{-3}	4.34	1.53^{-5}	

F(4,1054) = 85.1294 $Prob(> F) < 2.22^{-16}$

TABLE V
Efficiency scores of the institutions for the years 1997, 2000, 2003, and 2006

	1997		2000		2003		2006	
Efficient Range	Number of Institutions	Percentage	Number of Institutions	Percentage	Number of Institutions	Percentage	Number of Institutions	Percentage
$0.0 \leq E < 0.1$	1	0.71	0	0.0	0	0.0	0	0.0
$0.1 \leq E < 0.2$	12	8.51	19	10.9	6	3.4	12	6.8
$0.2 \leq E < 0.3$	25	17.73	38	21.8	31	17.7	33	18.8
$0.3 \leq E < 0.4$	22	15.60	24	13.8	29	16.6	31	17.8
$0.4 \leq E < 0.5$	18	12.77	18	10.3	27	15.4	21	11.9
$0.5 \leq E < 0.6$	15	10.64	20	11.5	22	12.6	19	10.8
$0.6 \leq E < 0.7$	9	6.38	7	4.0	9	5.1	15	8.5
$0.7 \leq E < 0.8$	7	4.96	7	4.0	8	4.6	7	4.0
$0.8 \leq E < 0.9$	6	4.26	7	4.0	8	4.6	6	3.4
$0.9 \leq E < 1$	4	2.84	5	2.9	6	3.4	5	2.8
$E = 1$	22	15.60	29	16.7	29	16.6	27	15.3
Total Institution Count								
Mean Efficiency	0.526		0.516		0.549		0.526	

The infrastructure also supports the connection between the data and visualization engines that support large scale data such as Tableau [15] and Datameer [16]. The RevoDeploy server also allows users to generate and display their own graphs based on the analyses using R scripts. In this example, the GIS information of the institutions is utilized in conjunction with NSF funding data to produce visualization of the data for the years 1997, 2000, 2003, and 2006.

IV. Conclusion and Future Work

In this paper, we describe an infrastructure that support the integration of data and tools for the field of research in higher education. The infrastructure allows data collected from different sources to be aggregated into a single data repository and researchers to have access to these data and to perform analysis on the data through an embedded statistical environment with parallelization support. The usability of the infrastructure is proven through a set of sample descriptive analyses done by undergraduate students participating in this work through the Research Experience for Undergraduates program. Future work includes but is not limited to:

- Continued expansion of data sources and enhancement of data ingestion process.
- Deployment of the baseline implementation on a production cluster for community access and performance study.
- Implementation of popular statistical analyses in the RevoDeploy server

Acknowledgment

We would like to acknowledge the work of Vijay Dantuluri, Michael Stealey, and Dr. Stanley Ahalt from RENCI in the acquisition and ingestion of a significant part for 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] College Board, http://www.collegeboard.org/, 2012.

[3] Institute for Scientific Information, http://isiwebofknowledge.com/, 2012.

[4] US News & World Report, http://www.usnews.com/, 2012.

[5] 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.

[6] National Science Foundation, https://webcaspar.nsf.gov/, 2012.

[7] L. Peer and A. Green, "Building an Open Data Repository for a Specialized Research Community: Process, Challenges and Lessons," *The International Journal of Digital Curation*, vol. 7, 2012.

[8] 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.

[9] 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.

[10] Top 500 Supercomputer Sites, http://www.top500.org/, 2012.

[11] Y. Wu, "Tackling undue concentration of federal research funding: An empirical assessment on NSF's Experimental Program to Stimulate Competitive Research (EPSCoR)," *Research Policy*, vol. 39, 2010.

[12] The R Project, http://www.r-project.org/, 2012.

[13] Revolution Analytics, http://www.revolutionanalytics.com/, 2012.

[14] J. Woolridge, *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, 2010.

[15] Tableau Software, http://www.tableausoftware.com/, 2012.

[16] Datameer, http://www.datameer.com/, 2012.

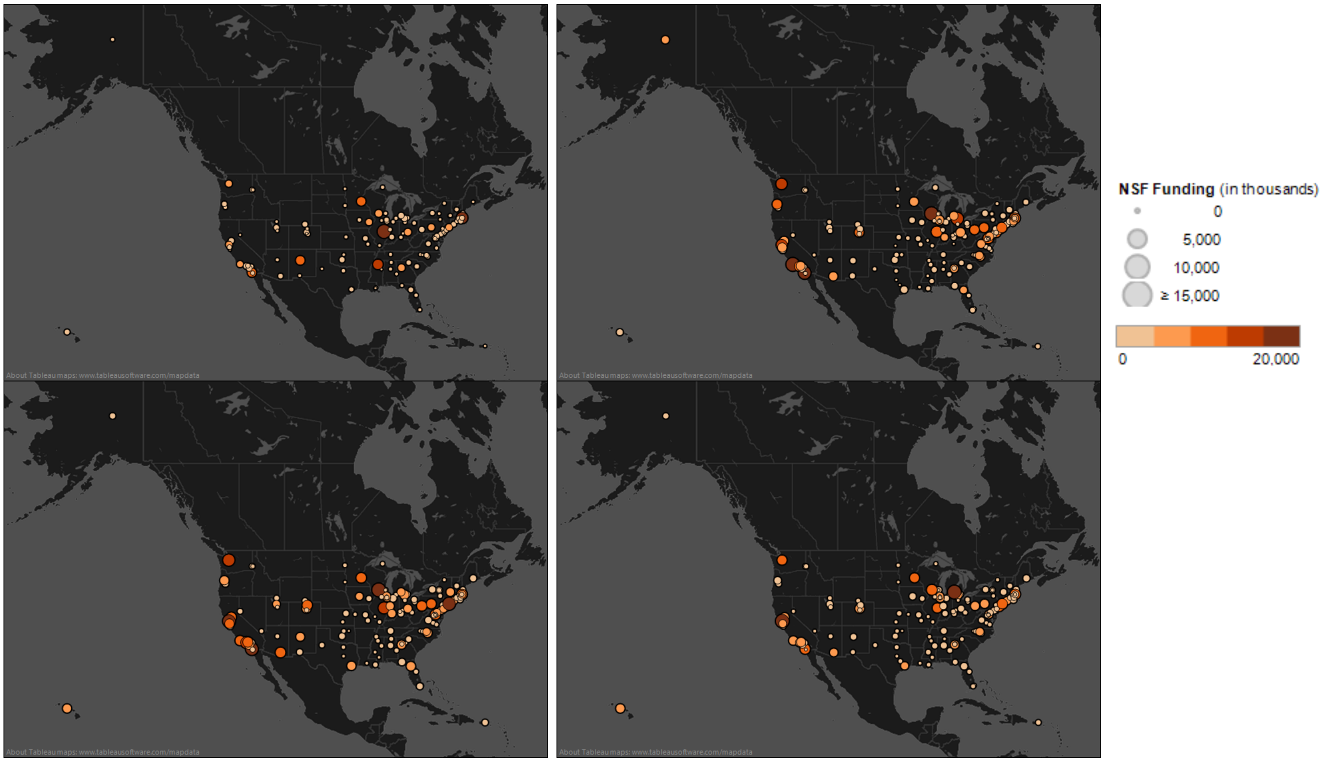
NSF Funding (in thousands)
0
5,000
10,000
≥ 15,000
0
20,000
About Tableau maps: www.tableausoftware.com/mapdata

Fig. 3: *NSF Funding per institution for the years 1997, 2000, 2003, and 2006 (left to right, top to bottom)*