Taming Big Data: Integrating diverse public data sources for economic competitiveness analytics

[Demonstration Proposal]

Rodica Neamtu, Ramoza Ahsan, Jeff Stokes, Armend Hoxha, Jialiang Bao, Stefan Gvozdenovic,

Ted Meyer, Nilesh Patel, Raghu Rangan, Yumou Wang, Dongyun Zhang, and Elke A. Rundensteiner

Computer Science Department, Worcester Polytechnic Institute, Worcester MA, USA.

rneamtu|rahsan|jeffstokes|ahoxha|jbao2|sgvozdenovic|tjmeyer|ncpatel|rsrangan|ywang10|dzhang3|rundenst @wpi.edu

ABSTRACT

In an era where Big Data can greatly impact a broad population, many novel opportunities arise, chief among them the ability to integrate data from diverse sources and "wrangle" it to extract novel insights. Conceived as a tool that can help both expert and non-expert users better understand public data, MATTERS [1] was collaboratively developed by the Massachusetts High Tech Council, WPI and other institutions as an analytic platform offering dynamic modeling capabilities. MATTERS is an integrative data source on high fidelity cost and talent competitiveness metrics. Its goal is to extract, integrate and model rich economic, financial, educational and technological information from renowned heterogeneous web data sources ranging from The US Census Bureau, The Bureau of Labor Statistics to the Institute of Education Sciences, all known to be critical factors influencing economic competitiveness of states. This demonstration of MATTERS illustrates how we tackle challenges of data acquisition, cleaning, integration and wrangling into appropriate representations, visualization and story-telling with data in the context of state competitiveness in the high-tech sector.

Categories and Subject Descriptors

H.2.7 [**Database Management**]: Database Administration—*Data warehouse and repository*

Keywords

Big data; data integration; diverse data sources

1. INTRODUCTION

Motivation and Background: In this era of big data, public web resources that host important societal, government, educational and economic data are increasingly plentiful. For example, a website like IPEDS [2] provides complete information in the area of education, including the number and type of STEM degrees granted by each educational institution across the United States, while the Tax Policy Center [3] contains information related to tax policies, rates and trends. These web data sources, typically kept up to date due to both government regulations and active user groups, represent extremely valuable public resources that can be leveraged for many applications, including policy decision making for state prosperity. For instance, in 2013, when the State of Massachusetts introduced a new Sales and Use Tax on computer and software services, organizations such as MHTC and other agencies set out to collect data-driven evidence about the potentially negative impact of such state regulations on the economic health of Massachusetts, and in particular, on the high tech sector. For this, they spent tremendous resources to collect historical data about similar actions in other states and to fight these new policies and eventually they convinced the legislators to repeal many laws. Because of the need for accurate, diverse and meaningful data to support the decision making process, organizations like MHTC are striving to "tame" the data and "wrangle" insights out of such large variety and veracity of public data sites in ways that will give them competitive advantage.

Challenges related to leveraging Data Resources: Although the data is publicly available, challenges in capitalizing on its power include proprietary access specific to each primary source, having to clean and unify the data and its terminology across differently owned data sites, and transforming it into some integrated data product, whose huge power can be harnessed to reveal its hidden correlations and predictive powers. Unfortunately, there is no single site comprehensive enough to provide data of the rich variety required to answer such complex policy questions. The extraction and integration of these metrics from heterogeneous web data sources poses a variety of technical

[1] Massachusetts Technology, Talent and Economy Reporting System.

[2] http://nces.ed.gov/ipeds/

[3] http://www.taxpolicycenter.org/

challenges, including the fact that each independent public source reports data in a different format, at different time intervals, with different levels of granularity and levels of aggregation. Other challenges include data acquisition, data cleaning, unified dynamic data integration and modeling and efficient warehousing. For organizations like MHTC, the cost in resources, expertise and time to collect data from diverse web data sources is tremendous. Thus, over time they had to recognize the need for an atomic big data integration system that pulls these diverse data sets into a unified format, where errors can't creep in and where the quality and accuracy of data is indisputable.

State of the Art: In this light, it is essential that a unique system be developed to overcome these tremendous challenges of dealing with big distributed data and provide easy access, flexible and powerful analytical visual and reporting tools. In general, two possible approaches have been taken to tackle such data integration challenges. In the first, in the loose mediator-based integration the data is kept in its original web sources, while the integration system acquires and combines data relevant to answer a particular user's questions upon demand [5]. In the second, the tight integration approach is to extract, collect, and then replicate all relevant data a-priori into one large integrated big data store, commonly called a data warehouse. This way, analytics tasks can thereafter be conducted directly on this dedicated data store [5]. While the former avoids data replication and directly leverages the storage and maintenance of the source data being kept up to date by its respective originators, it is well recognized that it tends not to be practical for many reasons. Most notably, data warehousing assures that all the labor-intensive tasks of data cleaning, unification, to transformation tasks, each challenging in their own right, are accomplished in an off-line fashion. Furthermore, user-driven data analytics requires visual interaction by the users with the data and thus new real-time performance - not afforded when undertaking the on-demand of the former approach heterogeneous data extraction.

Our MATTERS Approach for Economic Data Integration and Analytics: Driven by the above described big data opportunity, MHTC, WPI and other institutions collaboratively set out to develop MATTERS as an integrative data warehouse based solution on high fidelity state cost and talent competitiveness metrics. The key contributions of MATTERS illustrated in this demonstration include:

First, MATTERS features **robust data acquisition techniques** that break the complex problem of data acquisition into subtasks, generalizing each task into a composable component, automating tedious subtasks when possible and seamlessly plugging the human into the acquisition process.

Second, MATTERS provides an **economic indicator metric model** in time, space (map) and multidimensional metrics. The metric domain specific information can be analyzed from a historical perspective as well as used for comparisons and provides great insights into the main factors contributing to the ranking of states. As a foundation, our unified MATTERS data model achieves the flexibility of withstanding all dynamics related to existing and the additional of new data products and sources.

Third, our **visual analytics Dashboard** offers interactive displays enabling comparative analysis of competitive metrics across space and time. The comparative views are the key to finding correlations between different metrics, their time trends and the overall attractiveness of a state. For the first time, users can compare, correlate and aggregate side by side distinct metrics from public websites on key indicators for a specific state as well as exploring time-based trends from a historical perspective along metric spaces.

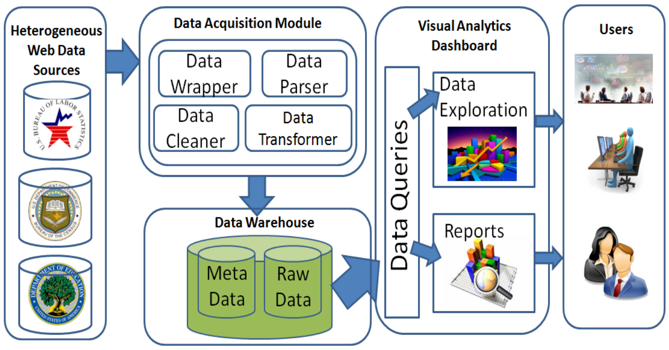


Figure 1: MATTERS Framework

Lastly, our **formative evaluation via a user study and stakeholder case study** involving stakeholders (members of MHTC) and general users (undergraduate students at WPI) has been progressively used to refine the technology to address both usability and usefulness of the technology.

2. SYSTEM OVERVIEW

The MATTERS framework is depicted in figure 1. The data wrapper extracts data from sources like the US Census Bureau, the Bureau of Labor Statistics, the National Science Foundation etc. The data parser reads the extracted files, while the data cleaner handles the errors, noisy or missing data to ensure correctness and quality of data loaded into the warehouse. It corrects the data entry errors, misspellings (of state names) and contradictory values. The data transformer converts the data of diverse sources into a unified schema and loads it into the data warehouse. This represents the foundation for subsequently allowing the users to slice and dice relevant information to assemble these data particles into high-order uni- and multidimensional objects. After the data is transformed, it is loaded into the warehouse. Meta data stores relevant information about the metrics like data type (numeric, currency, percentage), source name (Milken' s Institute, CNBC etc) and web link along with the category to which this metric belongs to. Each metric can be nested to any level of sub-metrics which in turn may have multiple data attributes. Meta data thus contains the relationship among metrics and sub-metrics while the raw data table stores the actual data values in a column-oriented key-value like approach. The data store also provides the ability for pages, files or views to be stored locally for further use and modeling. Our visual analytics dashboard enables the user to extract, view, analyze and model information for interactive data exploration.

3. KEY INNOVATIONS OF MATTERS

3.1 Robust Data Acquisition Techniques

Data Integration: While every data source provides valuable information in isolation, greater value is gained when integrating heterogeneous data across multiple public data sources [2], [4]. For example, the "Local Tax Burden per capita metric" is computed by integrating tax data from The Tax Policy Center with the state population data from US Census Bureau. Similarly, the "overall ranking" of the State of Massachusetts in selected key metrics is determined by aggregating individual metric scoring to different user-assigned weights. The data acquisition challenge was approached in a modular fashion, by breaking the complex problem into subtasks. We then generalize each task into a component that can be used independently or in combination with other components. We also automate tedious subtasks when possible and seamlessly plug the human into the acquisition process. This way the system can leverage the core technologies implemented in the six distinct data pipelines and re-use them to create a flexible semi-automated tool to allow new data from different data sources to be uploaded with minimal time and effort into the system.

Data cleaning: Conscientious that the quality of input is crucial to the decision making process, our system is designed to clean the data before loading it in the data warehouse [3]. This involves handling noisy, missing or irrelevant data. We built data cleaning tools that limit manual inspection and programming effort. They are extensible to easily cover additional sources. Data cleaning is not performed in isolation but together with schema-related data transformations based on comprehensive meta-data. One example of data cleaning handled by MATTERS is inconsistent state abbreviations where words like "Mas.", "MA", "Mass" all refer to the state "Massachusetts". A global mapping mechanism is incorporated to reconcile such differences. Another objective is to remove errors, inconsistencies across data sources, fill missing data with minimal manual intervention and align data across time dimensions. This cleaning capability is extensible as well as general enough to be applicable to seamlessly support the addition of new data sources.

3.2 Economic Indicator Metric Model

The key element to support the integration of heterogeneous data with diverse schema into one data warehouse is the use of a unified data model that is flexible, allowing for the storage of any type of dataset where data is classified into metrics and sub-metrics. For instance, the "State and Local tax Burden" metric has two sub-metrics, namely "Burden per capita" and "Burden per percent of personal income" where data of each sub-metric can be used to compute information about the main Tax metric. Each metric can be nested to any level of sub-metrics which in turn may have multiple data attributes. Rather than maintaining data in source-specific formats, we use a generalized column-oriented key-value like approach. For example, if the state population data corresponds to several sub-metrics, the integrated view will maintain it only once. Additionally, it will contain the mapping of this data back to the sub-metrics. This way we break down any dataset into its most elementary particles, while at the same time preserving the structure of the entire dataset. This enables assembly of these particles either to reconstruct the original source or to form different high-order data products which can be used for subsequent analytics. The meta-data allows us to "flatten" the data for storage and "reconstruct" and combine it for display and reporting needs. The unified model provides diversity (by supporting a wide variety of data sources within one system), extensibility (by avoiding schema evolution upon the addition of new file formats, new structures, or even new data sources) and generality (by providing any subset of metrics as integrated data product to data analytics service).

3.3 Visual Analytics Dashboard

The user interface displays information for interpretation, comparison and supports visual interactions for data exploration, thus creating an economic indicator metric model in time, space (map) and multidimensional metrics. The MATTERS dashboard enables the users to put together data from various sources and "wrangle" it into appropriate representations which will lead to better understanding and greater usability of the data. Even though every version of each data source contains valuable information, an aggregated view of data produced over the data sources often offers a greater value than the individual sources. Such reports are aggregated and coupled with information from other publicly available sources to build historical profiles and give insightful information about different states. Our system enables the users to compare side by side for the first time distinct metrics from public files on key indicators for a specific state and different metrics about peer states. Our visual tools allow the exploration of time-based trends from a historical perspective along metric spaces.

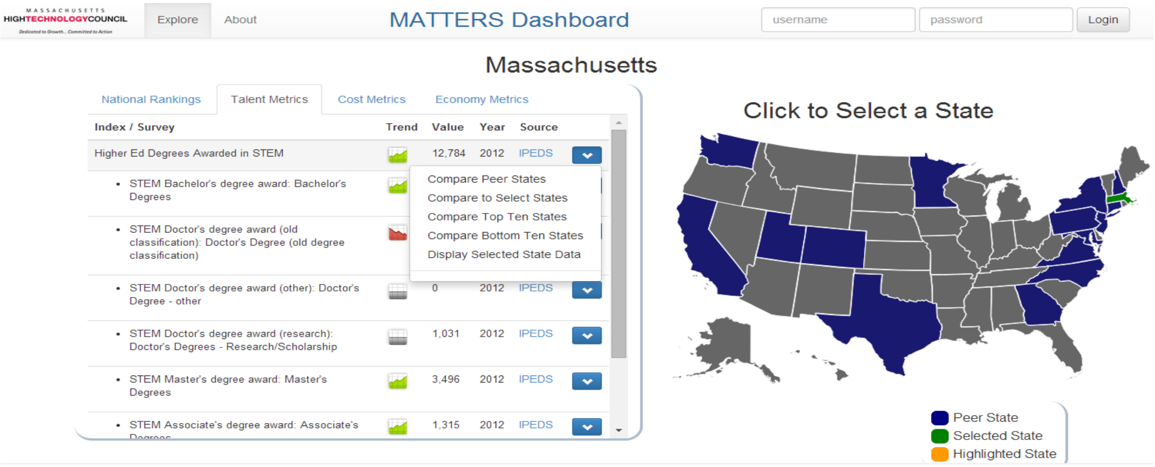
3.4 Formative evaluation via user study and stakeholder study

A usability study involving a general audience of undergraduate students at WPI was conducted. The feedback received has been beneficial in refining the visual and graphic technologies to insure easy and intuitive access to the dashboard. The stakeholder study conducted with members of the MHTC and executives from various high tech companies in MA was instrumental in addressing data quality and representation.

4. MATTERS DEMONSTRATION

We will first demonstrate how MATTERS deals with challenges posed by data acquisition, integration and transformation. As example, let us consider the collection and reporting talent (education) data provided by the Institute of Education Sciences. To create a report that can compare the number of STEM Higher Education degrees awarded over time in five states using the public data sources [4], a customized 20 step process has to be followed for each individual state and each individual year. This is a time and effort intensive endeavor requiring domain specific knowledge and technical expertise, while in addition being subject to human error. Extending this to a large number of states will significantly increase the cost of the task. Considering these issues, our system instead implements robust data acquisition leveraging core technologies to collect the data from the website, clean and transform it to fit the unified model before loading it into the data warehouse. This is an instantiation of the data pipeline that MATTERS uses to collect, clean, integrate, aggregate and display data from public sources. Using this combination of tools, the user

[4] http://nces.ed.gov/ipeds



MATTERS Dashboard

Massachusetts

Index / Survey	Trend	Value	Year	Source
Higher Ed Degrees Awarded in STEM		12,784	2012	IPEDS
• STEM Bachelor's degree award: Bachelor's Degrees				
• STEM Doctor's degree award (old classification): Doctor's Degree (old degree classification)				
• STEM Doctor's degree award (other): Doctor's Degree - other		0	2012	IPEDS
• STEM Doctor's degree award (research): Doctor's Degrees - Research/Scholarship		1,031	2012	IPEDS
• STEM Master's degree award: Master's Degrees		3,496	2012	IPEDS
• STEM Associate's degree award: Associate's Degrees		1,315	2012	IPEDS

Figure 2: MATTERS Dashboard Main View.

doesn't need domain specific knowledge, nor expertise. Instead, the visual dashboard will employ easy to use dialog boxes to perform comparisons by states and in time. Wrappers collect data directly from the source, cleaning modules clean, fill in missing data, functions store any relevant information about the data representation (units, dollars, etc) and towards this end the data is transformed to fit the unified format. The system uses basic entities like state, time and metrics, to store all relevant information, that can be composed as needed to achieve the desired level of granularity for displaying and analytics. The time and effort are minimal, the data is much less prone to being affected by human errors and the results benefit from a large array of display and reporting tools.
Next we will demonstrate how the MATTERS dashboard can be used as an economic indicator metric model in time, space and across many multidimensional metrics. It provides an interactive visual analytic framework and enables the users to perform analysis across states, time and metrics. The main view of the dashboard contains the map for state selection (RHS of Figure 2) and a metric data display (LHS of Figure 2). The state oriented view displays the data for selected states, state ranking in all metrics and source of the data. The time oriented view enables the users to explore historical time series data of a metric in different visual display types. It provides insights in finding the trends, deviations and similar time series data for a selected state. The trends for each metric describe whether the metric value is improving, declining or remaining steady over the years. The metric oriented view offers users the capability to perform comparative analysis, correlations, and associations and aggregate the results across different metrics. Figure 3 shows the state oriented view for metric "Higher Education Degrees Awarded in Stem metric" with selected states as Massachusetts, New York, Maryland, California and Texas.

Conclusion: MATTERS extracts, integrates and models rich economic, financial, educational and technological information from renowned heterogeneous web data sources This demonstration showed how we tackle challenges of data acquisition, cleaning, integration, mining, and "wrangling" into appropriate representations, visualization and storytelling with data in the context of state competitiveness in the high-tech sector.

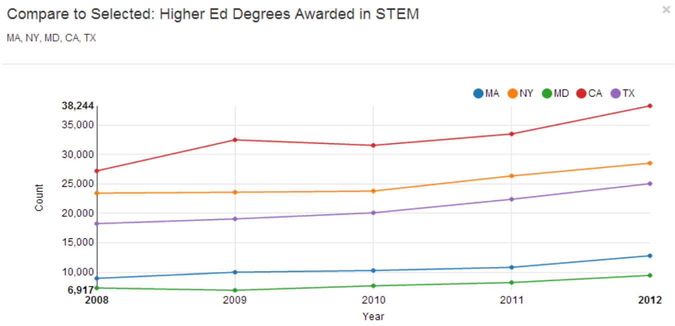


Figure 3: Visual comparative display of economic indicators across time

5. ACKNOWLEDGMENT

We thank C. Anderson and M. Catizone from MHTC for this project opportunity. We also thank WPI, Fulbright, USAID and NSF IIS 1018443 for partial support of this project.

6. REFERENCES

[1] Yigal Arens, Chin Y Chee, Chun-Nan Hsu, and Craig A Knoblock. Retrieving and integrating data from multiple information sources. *International Journal of Intelligent and Cooperative Information Systems*, 2(02):127–158, 1993.

[2] Andrea Calì, Diego Calvanese, Giuseppe De Giacomo, and Maurizio Lenzerini. Data integration under integrity constraints. In *Seminal Contributions to Information Systems Engineering*, pages 335–352. Springer, 2013.

[3] Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.

[4] Mary Roth and Wang-Chiew Tan. Data integration and data exchange: It's really about time. In *CIDR*, 2013.

[5] Patrick Ziegler and Klaus R Dittrich. Data integration problems, approaches, and perspectives. In *Conceptual Modelling in Information Systems Engineering*, pages 39–58. Springer, 2007.