Massachusetts Economy and Technology Index System

[Extended Abstract]

Ramoza Ahsan, Rodica Neamtu, 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

100 Institute Road, Worcester MA, USA.

rahsan|rneamtu|jeffstokes|ahoxha|jbao2|sgvozdenovic @wpi.edu

tmathmeyer@gmail.com, ncpatel|rsrangan|ywang10|dzhang3|rundenst @wpi.edu

ABSTRACT

In the era of big data, economic competitiveness is assured by decision making leveraging insights gained from large scale yet granular data sets from a rich diversity of areas. In this light, the METIS [1] system collaboratively developed by a team at WPI, the Massachusetts High Tech Council and other institutions, emerges as an analytic platform offering dynamic modeling capabilities. The integrative data source is based on high fidelity cost and talent competitive metrics. METIS extracts, integrates and models rich economic, financial, educational and technological information from renowned public heterogeneous web data sources ranging from The US Census Bureau, The Bureau of Labor Statistics, to Institute of Education Sciences. The METIS technology creates a powerful tool that allows intuitive analysis of the key factors that drive Massachusetts cost and talent competitiveness relevant to high tech companies.

1. INTRODUCTION

METIS [1] is a dynamic dashboard with modeling capabilities designed to assist and enrich the decision making process across a set of metrics relevant to high tech companies. Examples of metrics integrated by METIS include state and local tax burden per capita and personal income from The US Census Bureau [2], unemployment insurance payroll tax from The Tax Policy Center [3], total employment from Bureau of Labor Statistics [4], unemployment rates from the Bureau of Labor Statistics and STEM degrees produced from The Institute of Education Sciences [5]. The extraction and integration of these economic metrics from such diverse web data sources poses a variety of technical challenges. Challenges include robust data acquisition, unified dynamic data integration and modeling, efficient warehousing, and effective analytics services. With these advanced analytics capabilities including time line analysis, trend detection and predictive analytics, METIS promises to enable policy makers and lobbyists to take data-driven approaches to drive and change policies and rules anchored in economic, political and educational realities.

The quest to overcome these challenges and channel these resources into useful analytics is compounded by the need to create diverse high-level data views most appropriate for specific analytics. As a foundation, our unified METIS data model with meta data achieves the flexibility of withstanding all dynamics related to existing and the addition of new data products and sources.

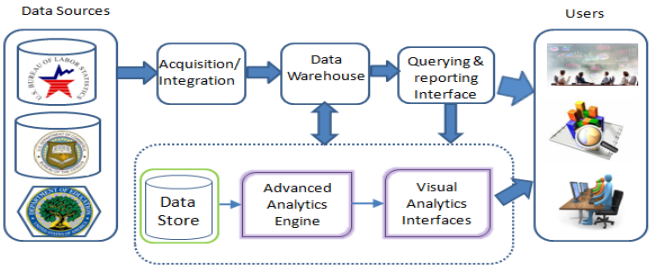

Figure 1: METIS work flow.

2. SYSTEM OVERVIEW

The overall system work flow designed to integrate the datasets and provide analytical, visualization and reporting capabilities is depicted in Figure 1. The data acquisition component retrieves, cleans and integrates data from diverse web data sources. The data warehouse then houses the transformed data in a unified manner, and allows the

[1]Massachusetts Economy and Technology Index System.
[2]http://www.census.gov/govs/statetax
[3]http://www.taxpolicycenter.org
[4]http://www.bls.gov
[5]http://nces.ed.gov

ACM SIGMOD '14, June 27, 2014, Snowbird, Utah, USA

users to slice and dice relevant information to assembly these data participles into high-order uni- and multidimensional objects. The data store provides the ability for pages, files or views to be stored locally for further use and modeling. The advanced analytics engine enables data modeling and mining. The querying and reporting interfaces enable the users to extract, view, analyze and model information, while the visual analytics interface is the medium for interactive data exploration.

3. METIS DATA ACQUISITION PIPELINE

With a huge quantity and variety of economic data in the world, it is rare for data to be entirely contained in a single data source or to be managed by a single authority. Also each data source may provide only partial knowledge about the economic metric. Thus the integration of data across multiple sources is often needed to reveal insights and patterns that may otherwise be hidden [1]. While critical, this task of accurately and consistently combining heterogeneous data sources poses many challenges.

Data Cleaning: To ensure quality of input to the decision making process, the data in the data warehouse must be "cleaned" which involves handling noisy, missing or irrelevant data. One example of data cleaning handled by METIS is inconsistent state abbreviations where words like "Mas.", "MA", "Mass" all refer to the state "Massachusetts". To reconcile such differences, a global mapping mechanism is incorporated. 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 to seamlessly support the addition of new data sources.

Data Integration: While every data source provides valuable information in isolation, greater value is gained when integrating this heterogeneous data across data sources. For example, we compute the "Local Tax Burden per capita metric" by integrating tax data from "The Tax Policy Center" with the state population data from "US Census Bureau".The overall ranking of the State of Massachusetts in selected key metrics is determined by aggregating individual metric scoring to different user-assigned weights. Towards this end, the cleaned data is transformed, normalized, and consolidated into a unified data model, as explained below.

4. METIS MODELING AND MANAGEMENT

To support the integration of this heterogeneous data with diverse schema into one data warehouse to support its cross-source analytics, a unified data model is designed, called the METIS model. The METIS unified data model is flexible, allowing for the storage of any type of dataset where data is classified into metrics and submetrics. 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 submetric 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 submetrics, the integrated view will maintain it only once.Additionally it will contain the mapping of this data back to the submetrics. 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, as explained below.

5. METIS DATA ANALYTICS

METIS aims to provide **descriptive data mining** (to explain the data and extract interesting properties and interrelationships) as well as **predictive data mining** capabilities (to construct a set of models to infer the behavior of new data set or the predicted effect of changes on the data for what-if analytics [2]). Some of the analytics tasks essential for METIS are explained below.

Classification allows us to compare states across different metrics to identify key features likely to improve the overall ranking of the State of Massachusetts. We do this by constructing a concise summarization of the stored data as well as data distribution information such as variance.

Time Series analysis provides mechanisms to analyze large sets of time series data to find interesting characteristics like trends, deviations or similar sequences. Each metric's historical time series are saved and subsequently can be used to predict the ranking of Massachusetts in the high technology employment metric for the coming year.

Prediction Task aims to predict the most plausible values of some missing data or value distribution of certain attributes. Using statistical analysis, it can be used for instance to predict the potential for a state's high tech employment using data of previous years.

Visual Interfaces display information for interpretation, comparison and support visual interactions for data exploration. Sifting through huge amounts of data, especially through the results of data mining that can be bigger than the data itself, is a complicated task leading to the challenge of choosing and controlling the analytics, as well as interpreting the results.

6. CONCLUSION

METIS extracts, transforms and loads data from public heterogeneous sources into one data warehouse. The unified data model affords METIS the flexibility of assembling the particles of data from different data sources useful for subsequent mining. METIS is in its incipient phase and is being developed in collaboration with MHTC into a highly interactive dashboard with modeling, predictive and data mining capabilities to inform and help the process of policy making.

7. ACKNOWLEDGMENT

We thank Christopher Anderson and Mark Catizone from Massachusetts High Tech Council for vision of this project as well as for continued guidance and support. We also thank WPI, Fulbright, USAID and NSF IIS 1018443 for partial support of this project.

8. REFERENCES

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

[2] Daniel Deutch, Zachary G Ives, Tova Milo, and Val Tannen. Caravan: Provisioning for what-if analysis. In *CIDR*, 2013.