Environment for Datasets Processing and Visualization Using SciDB

Rui Wu Sergiu M. Dascalu Frederick C. Harris, Jr
Department of Computer Science and Engineering
University of Nevada
Reno, USA
{rui, dascalus, fred.harris}@cse.unr.edu

Abstract

When a scientist analyzes certain datasets, there are three common steps—import, visualize, and process data. There are some prevalent tools to visualize and process data, such as Matlab, Powersim, and Stella. However, these tools cannot handle big data. For data management, scientists are pursuing a new generation of Database Management Systems (DBMS) to replace traditional Relational Database Management Systems (RDBMS), because RDBMS is good at data management, but does not perform well on raw data and time series data. Furthermore, most of the new tools, such as Hadoop, cannot fulfil scientists' needs of data management. This paper introduces a web-based application created by us to import, visualize, and process big data. The system offers two methods for users to import data—obtain data from foreign repositories and upload users' files. We used SciDB for data management, because SciDB is designed for big data and scientific use. We used D3.js and Dygraphs libraries for data visualization, which enable users visualize millions of points without experiencing lag.

Keywords-SciDB; Data Visualization; Data Discovery; Data Processing; Big Data

1 Introduction

Big data and its management is prevalent and significant for both businessmen and scientists. The digital era brings us many opportunities and also many problems. Almost every device generates data at all times and there are "gold mines" buried in these datasets. For example, the Facebook warehouse stores upwards of 300 petabytes with a daily incoming rate around 600 terabytes [1]. There are 300 hours of video materials uploaded to YouTube every minute [2]. However, it is hard to manage and analyze big data. Researchers held many conferences to resolve these hard big data problems, such as Extremely Large Databases and Data Management [3]. Dr. Michael Stonebraker and his team created SciDB to manage big data. The initial version of SciDB was published in 2008 and the latest version is 14.12, which was published on January 09, 2015. After seven years of testing and development, SciDB is an excellent database and recommended by many scientists, such as authors of [4] pointed that one of most import key to create big data analytics software is to use a NoSQL database, like SciDB.

Most database management systems lack interactivity [5]. We need to build a wrapper to interact with databases. Data visualization is one of these interactions and it is a significant step in data processing. It is much easier to find the trend and compare results from visualization results than searching raw data.

Our system is named EDP&V, which is short for Environment for Datasets Processing and Visualization. We aim to create a system that can discover, process, and visualize large datasets. The system will allow users to select available datasets from different repositories and change the datasets format into SciDB, if the system finds the datasets format is not in SciDB format. Then users can choose different methods to visualize the data, such as bar charts and line charts. Also, users can analyze selected datasets within the system. They can get minimum, maximum, and average values of the datasets. We plan to enable users to do some complex data analysis with some models in the future. If users have some questions about how to use the system, they can view Help Manual. For further instructions, users can email us. We used SciDB to create the backend, Html and JavaScript to create the frontend, D3 and Dygraphs libraries to visualize and interact with datasets. The reason that we did not choose other new tools (such as Hadoop) to manage data, is that these tools cannot fulfil scientists' needs [6]. Now, the latest version of EDP&V can visualize the datasets chosen by users and the datasets discovery and processing are still in development.

The paper, in the remaining part, is organized as follows: Section II introduces the problems in building a web-based application for big data; Section III presents our proposed system named EDP&V; Section IV includes the screenshots of EDP&V and introduces different parts of the system; finally, Section V presents three similar tools and discusses their advantages and disadvantages.

2 Problems

There are several challenges to build a web-based application for big data.

First, we need to choose a good database management system. There are many mature and popular Relational Database Management Systems (RDBMS), such as MySQL, Postgres, and the Oracle Database. However, these RDBMSs are not suitable to large scale and high-concurrency applications [7]. NoSQL database management systems, such as SciDB and MongoDB, are designed to manage big data. The

978-1-943436-01-9 / copyright ISCA, SEDE 2015
October 12-14, 2015, San Diego, California, USA

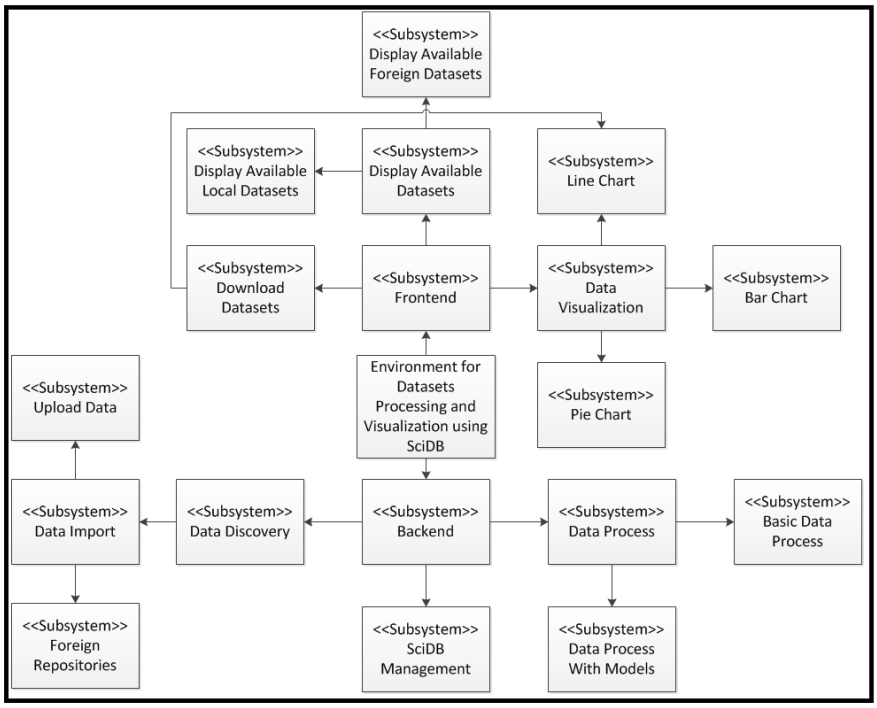

Figure 1. System-level Diagram

reasons we chose SciDB instead of other NoSQL databases were that SciDB focused on scientific use and SciDB aimed to manage 100 PB [8], which is the largest scale for an open source database in our experience.

Second, we need to visualize datasets stored in SciDB. There are some mature and popular data visualization tools, such as Matlab, R, and D3. Matlab is widely used in the scientific field, however there is not a glue tool to connect Matlab and SciDB in our knowledge. Dr. Michael Stonebraker and his team did create SciDB-R, which integrated R with SciDB. However our goal is to create a web-based application and it is hard to integrate R with a web-based application. Finally, we found a JavaScript library, named D3, which is used to produce interactive and dynamic data visualizations in the frontend of a web-based application [9].

3 Proposed System

3.1 System-level Diagram

Figure 1 presents the system-level diagram. The whole system has two main parts—frontend and backend.

The Backend is mainly in charge of database management, process data, and datasets discovery. We put data processing in the backend not frontend because it is faster to process datasets with Python than JavaScript. Data management is a significant section of the whole system and it is in the Backend. Data management includes two parts: Data Discovery and SciDB Management. Users can import data by uploading files or input a foreign repository URL. If the system receives a URL, it will try to connect to the foreign repository and display available files and folders. Users can import files into SciDB directly from the foreign repository.

The Frontend is mainly in charge of data processing results presentation and data visualization. With the help of D3 library and Dygraphs library, data visualization is fast and data interaction is intuitive and convenient. EDP&V offers three methods to visualize data—line charts, bar charts, and pie charts. Users can visualize raw datasets and data process results. The reason that there is an arrow from "Download Datasets" to "Line Chart" is that EDP&V enables users to choose an area from line charts and download the chosen data only.

3.2 System Use Scenario

We believe it is easier for others to understand how EDP&V works with the help of some scenarios. Therefore, we list two scenarios below:

First scenario: User A is a fish scientist. He wants to predict the fish population density of a river. User A explores EDP&V and the system displays all the available data stored in the EDP&V server. However, there is no suitable data and User A does not want to search foreign repositories. Therefore, he uploads a csv file to EDP&V. User A finds that there is a good fish model offered by EDP&V and he or she uses his upload file as an input for this model. EDP&V visualizes the model outputs (fish population density) with a line chart. User A does not want to download all the outputs, so he or she chooses an area from the line chart and just downloads the chosen data.

Second scenario: User B is a climate scientist. The user wants to understand the trends of air temperature in Nevada.

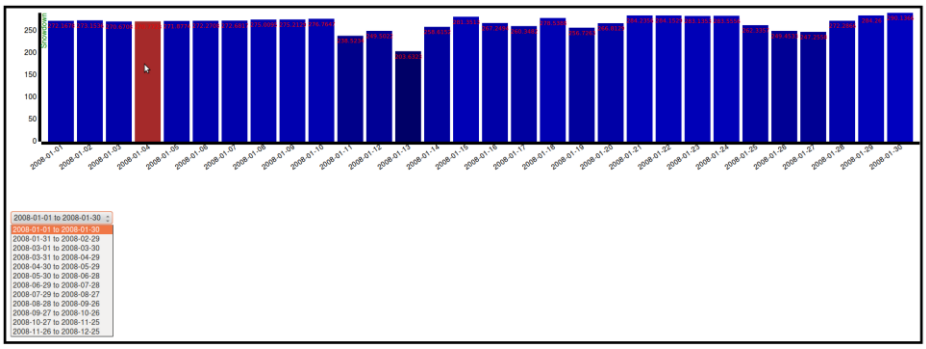

Figure 2. Bar Chart of Snow Down

User B explores EDP&V and the system presents all the data stored in EDP&V server. However there is no climate data stored in EDP&V. User B knows that website C offers qualified climate data. Therefore, User B inputs the URL of website C's FTP server into EDP&V. EDP&V displays all folders and files in website C's FTP server. User B chooses one of the air temperature files to import into the EDP&V server database (SciDB) and EDP&V offers three methods to visualize the air temperature data—line chart, bar chart, and pie chart for User B.

4 Results and Evaluation

The latest version of EDP&V mainly focuses on data visualization. There are three visualization methods in our system: 1) bar chart; 2) line chart; and 3) pie chart. Users can interact with these charts with the help of D3 library and Dygraphs library. EDP&V enables users to discover data only when the foreign repository is a FTP server. There is only one fish model integrated with EDP&V, which means there is still a lot of work to do to make the subsystem "Data Process with models" functional. We also introduce an interesting tool to present contact information in this section.

4.1 Bar Chart

Figure 2 is a screenshot of a bar chart in our system. It displays a dataset of snow down in the database. There is a label at the top of each bar, which presents the exact value of that bar. The date value is marked in the x-axis and snow down is marked in the y-axis. Each bar has different colors and the color is decided by the value of a bar (bigger number with lighter color). The bar will turn red when the user's cursor hovers over it.

There is a drop down list below the bar chart. Each option in that list is the name of a dataset in the database. In Figure 2, the option is a window of date, because we named each dataset of snow down with its date. When users choose a dataset from the drop down list, the bar chart will be updated. Because we used D3 library, the update process is smooth and the update animation will be displayed. This drop down list is created dynamically. It means if the administer deletes or adds datasets in the database, the drop down list will update automatically.

4.2 Line Chart

Sometimes users prefer line charts, because line charts can display data changing trends better than bar chart. Figure 3 shows a line chart screenshot of our system.

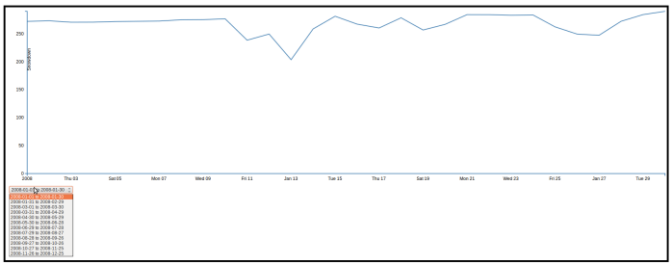

Figure 3. Line Chart of Snow Down

The drop down list below the line chart is created with the same technique introduced in the bar chart section. Therefore, the drop down list will be updated automatically, if administers change the datasets stored in the database.

4.3 Pie Chart

EDP&V also enable users to visualize datasets with pie chart. For example, Figure 4 displays a dataset of snow down in the database. The left pie chart displays the snow down value of each day. The size of each arc is decided by the value of snow down. A high value is assigned with a large arc. When users click on the pie chart, the value of each arc will be updated with the date as the right pie chart presents. If users click on the pie chart again, the value will change back to snow down. All the update processes are really smooth with the help of D3 library.

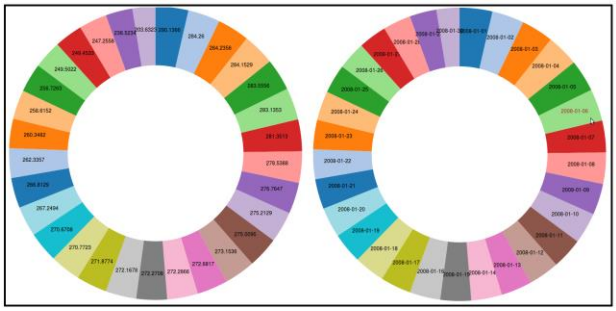

Figure 4. Pie Chart of Snow Down

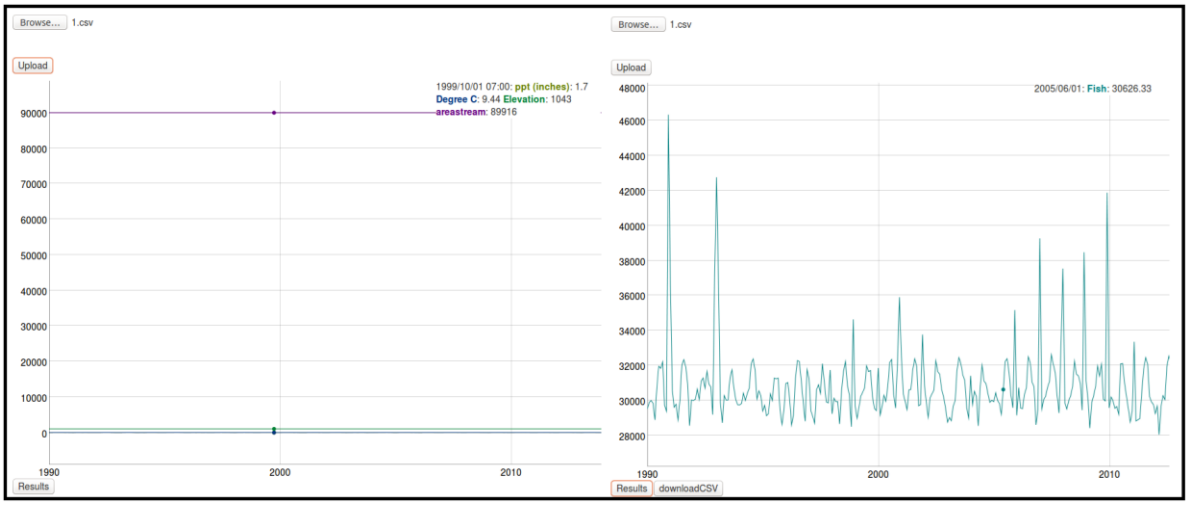

Figure 5. Fish Model Multi-runs

4.4 Data Process with Models

There is a fish model integrated with EDP&V now. There are two ways to run the models. Users can choose inputs by using sliders and run the model once. Figure 6 displays the user interface of this fish model. The model is designed for Webb Creek and Sweetwater Creek. When users change temperatures of the creeks, the color of these creeks will change. When users change precipitation, the width of these creeks will change (the bigger the precipitation, the wider the creeks). When users click the result button, a result will be displayed.

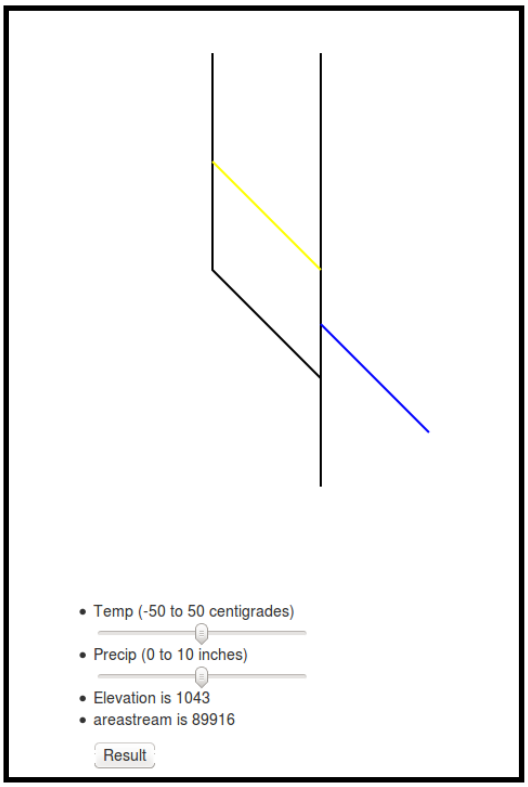

Figure 6. Fish Model Sliders Input

Users can also upload a csv file to EDP&V servers to run this fish model for multiple times. Figure 5 displays the user interfaces. Users can click the "Browse" button to upload a input file and the upload file will be visualized with a line chart as the left part of figure 6 presents. When users click the "Result" button, the line chart will be updated to display model results as the right part of Figure 5 displays and a "DownloadCSV" button will appear near the "Result" button. Users can click it to download the model outputs (a csv file).

If users want to download parts of the outputs, they can choose part of line chart as Figure 7 left part presents. The line chart will zoom in and when users click the "DownloadCSV" the system will offer a csv file containing the chosen part of data only.

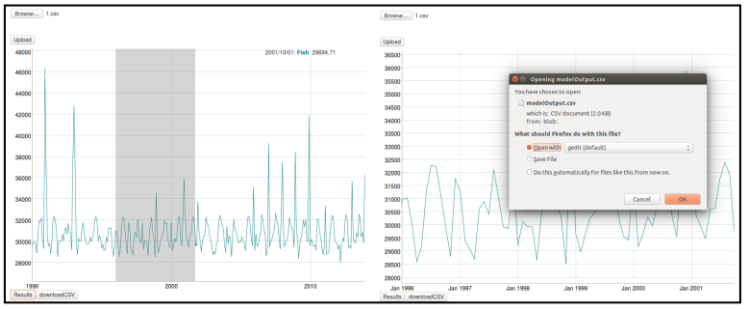

Figure 7. Download Chosen Model Outputs

4.5 Data Discovery

The latest version of EDP&V enables users to import csv files from foreign FTP servers. Users need to offer valid URLs and EDP&V will try to connect to the FTP server. If the connection is established successfully, the system will display all the files and folders in the FTP server.

Figure 8. Data Discovery Screenshot

Figure 8 displays the interfaces of EDP&V data discovery part. When users click a folder, the system will display folders and files inside that folder. When users click on a file, the system will download the chosen file and import it into the EDP&V server. Figure 8 displays the screenshot when EDP&V connects to a FTP server. Users can click the "Upper level" button to return to the parent of the current working directory.

4.6 Contact Information

We found most websites display contact information in a straightforward way. We want to offer our users some new experiences. Therefore, we used the technique from [10]. Figure 9 shows a screenshot of the contact information display tool.

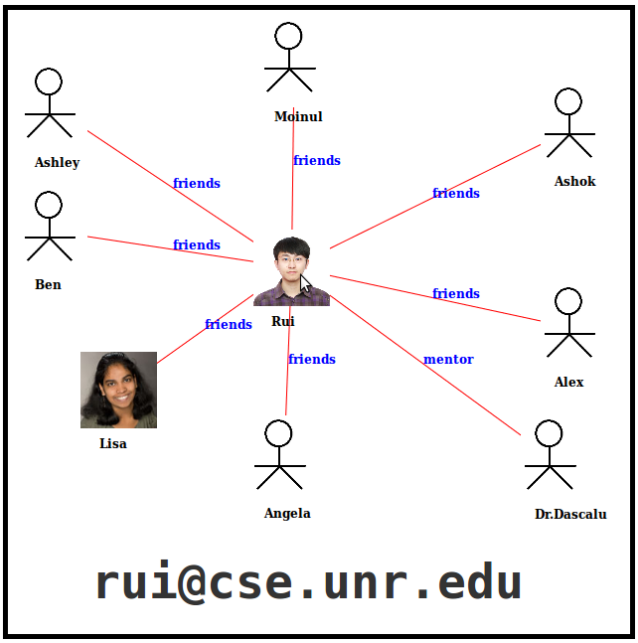

Figure 9. Contact Information and Relations

Each actor is connected with each other with a line. Users can drag the photo of a person to a certain place and the photo will stay when users release their mouse. When users double click a certain photo, the force of each line connected to the photo will be released. These lines will act like rubber bands. When a user's mouse hovers over a photo, the tool will display his or her email address. Also, the color of the lines connected to the photo will change into red and there will be blue words appearing on each line, which describe the relations between the chosen person and other persons.

5 Related Work

There are some similar projects to EDP&V, such as VISTED [11], InstantAtlas [12], and Dygraphs [13]. They are all good software programs used to visualize datasets and have their own advantages and some disadvantages.

VISTED is developed by Ph.D. student, Likhitha Ravi at University of Nevada, Reno. Climate researchers can identify the trends, outliers, and patterns with the help of visualization results generated by VISTED. The advantages of this tool is that it is easy to use and users can interact with some of the visualization results, such as line chart. However, users can just visualize and download datasets generated by NCAR/WRF climate model [14], which is from 1980/01-2009/12 NCEP/NCAR reanalysis and CCSM3 and 2040/01-2069/12 for CCSM3 based on the A2 Scenario. In contrast to it, EDP&V aims to enable users to discover datasets by themselves, which means users can load datasets into databases from different websites and repositories.

InstantAtlas helps researchers and businessmen visualize datasets and analyze them with different charts [12]. The basic idea of this tool is that it offers some templates to users and users can customize these templates in a dashboard. Users need to choose different datasets to process. The InstantAtlas official website offers some datasets based on the chosen area and template. Users can also upload datasets by themselves. Based on these datasets, users can add different charts, such as line chart and bar chart. InstantAtlas is a good dataset visualization tool. It is not complicated to use and there are some tutorials on the official websites. However, this tool is not free. Only, when users upload the datasets fit that the chosen template, it will work.

Dygraphs enables users to explore and interpret dense data sets [13]. Dygraphs is a high-level library, which means it is easier for users to customize their visualization results than most of the other tools. However, users need to write their own JavaScript programs, as Figure 10 presents. This could be a difficult task for people that do not know how to program. Dygraphs is easier to use than D3 library, but there are more limits to customization than D3 library. Dygraphs aims to process big datasets and it can plot millions of points without experiencing lag [13].

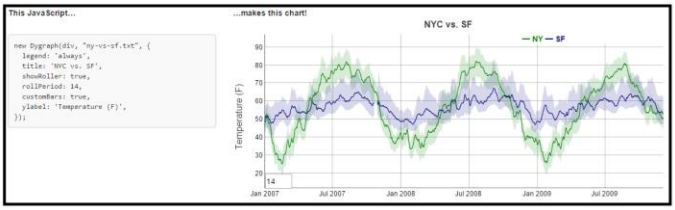

Figure 10. An Example of Dygraphs Visualization [13]

There are some good tutorials on its official website and it is a totally free library.

Compares to these tools, EDP&V is still not mature. However, we learned many good points from these tools. To be better, EDP&V should make datasets discovery easy for users and EDP&V should process datasets with some complex methods other than drawing graphs. For example, users can choose to process datasets with models offered by EDP&V.

6 Conclusion and Future Work

People have realized the precious value of data and they need good tools to manage and process different datasets. However, traditional tools are not capable of handling big data. Therefore, we need to build some new tools.

In this paper, we introduced our system, named EDP&V. It is a web-based application for big data, which has three parts: 1) dataset discovery, which means users can load datasets into SciDB from different websites and repositories; 2) dataset visualization, which means users are able to visualize datasets with different methods, such as bar charts and line charts; and 3) dataset process, which means users can process datasets with some basic methods, such as obtaining maximum value, and they can also process the chosen datasets with some complex methods, such as processing datasets with a system offered model. We have finished the datasets visualization part of EDP&V and plan to finish the other two parts in the future.

References

[1] Scaling the Facebook data warehouse to 300 PB (accessed 5/4/2015), https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/

[2] Statistic—YouTube (accessed 5/4/2015), https://www.youtube.com/yt/press/statistics.html

[3] Stonebraker, M., Becla, J., DeWitt, D. J., Lim, K. T., Maier, D., Ratzesberger, O., and Zdonik, S. B. (2009). Requirements for Science Data Bases and SciDB. In *Proceedings of Fourth Biennial Conference on Innovative Data Systems* , Vol. 7, pp. 173-184.

[4] Otero, C. and Peter, A. (2014). Research Directions for Engineering Big Data Analytics Software. Intelligent Systems, IEEE, 30 (1), pp 13-19.

[5] Icingir, H. T. (2013). Visualization of Semantic Windows with SciDB Integration. Department of Computer Science, Brown University.

[6] Stonebraker, M. (2012). What Does' Big Data Mean? *Communications of the ACM, BLOG@ ACM*.

[7] Han, J., Haihong, E., Le, G., & Du, J. (2011, October). Survey on NoSQL database. In *Proceedings of the 6th International Conference on Pervasive Computing and Applications (ICPCA), 2011* IEEE, pp. 363-366.

[8] Product: SciDB –A Science-Oriented DBMS At 100 Petabytes (last acessed 5/11/2015), http://highscalability.com/blog/2010/4/29/product-scidb-a-science-oriented-dbms-at-100-petabytes.html

[9] Data-Driven Documents (last acceseed 5/11/2015), http://d3js.org/

[10] The Force-Directed Relationship Diagram with D3.js (last accessed 5/12/2015), http://www.ourd3js.com/wordpress/?p=606

[11] Home Page – VISTED (last accessd 5/11/2015), http://sensor.nevada.edu/VISTED/

[12] Interactive mapping software | InstantAtlas (last acessed 5/12/2015), http://www.instantatlas.com/

[13] Dygraphs.com (last accessed 5/12/2015), http://dygraphs.com/

[14] Modeling Output (last accessed 5/12/2015), http://sensor.nevada.edu/NCCP/Downloads/Modeling%20Output.aspx