Visualizing the performance of Datawarehouse using Microsoft Power BI tool

Uha Rani Gunupuru

Department of Computer Science

University of North Carolina at Charlotte Charlotte, USA

ugunupur@uncc.edu

***Abstract* - Graphical representation of any data makes it easier to understand even for a naïve user. Data warehouse is a huge store of data accumulated from various sources within a company. Sometimes, it becomes difficult to identify and resolve the problems of data warehouses as data from various sources is stored in it for analysis of the data and make future predictions. Visualizations bring forward the unknown facts of the data and enables the decision makers see the current situation and make proper decisions. Hence, to identify errored data and get insights about performance of the warehouse on daily basis, we generate a report using Microsoft Power BI tool.**

Keywords- Data Warehouse; Data Visualization; Power BI

I. INTRODUCTION

Microsoft Power BI is a Microsoft product that is used for analysis and generating reports for complex data. It offers a free desktop version where the users can design their reports, dashboards, etc. and publish their reports on "powerbi.com". It also has two paid versions called the Power BI Pro version and the Power BI Premium version. The first one improves the data refreshes, provides more data storage, etc. than the Power Bi Free version and is cost effective when compared to the Premium version of the Power BI. However, with Power BI Pro version, if the user wants to share reports or dashboards with others, the others also should have a "Pro" license. To overcome this, the Microsoft has released a fully upgraded version called the "Premium" version where the organization with the Power BI Premium version has a super-powered server running in their Power BI environment.

Data Warehousing and Power BI

For analysing the data of an organization that has data coming from different departments like the Human Resources, Finance, Customer, etc. and indenting all these data together in a uniform manner, there are three processes called the "ETL (Extract, Transform and Load) process" is performed on the data and then stored in a huge database called the data warehouse that when connected with the Power BI allows users to generate reports and give a deeper view of the data. In this paper, we have used the organizational data of "XPO Logistics", a UK based supply chain company.

The first process of the "ETL process" is the extraction of data. Since we have data from different departments of the organization, we have different sources like the oracle database, salesforce, flat files, etc. as each department may use different methods of storing the data. The process of extracting data from the various sources available and staging them into a temporary area is called "Extraction of data". Usually, XPO performs a few pre-requisite steps using Software System Integration Services (SSIS) before storing the data in the data warehouse.

The second process or the process of "Transformation" of data is done. Usually in XPO, this is done when data from various sources is extracted and stored in a "Landing" area. This transformed data is then moved into a temporary staging location called the "Archive" area.

The third process of the ETL is "Loading", loading the data into the warehouse. This is done once the data in the staging area or the Archive area is stable with the transformations made on the data.

II. TECHNICAL OVERVIEW

The data reporting tools used for designing dashboards may vary from one organization to the other based on the size of the organization, their budget and the profit made by the projects. At XPO Logistics, we have used a Microsoft tool called the Microsoft Power BI (Business Intelligence) tool as it one of the open source tools for the desktop version and cost effective for "Pro" license to the users. The "Pro" license usually costs 10.00 USD for a user.

Like the former tools like tableau, info gram and other tools in the field of Data Visualization, the Power BI tool also enables users to plot line graphs, clustered graphs, bar graphs, etc. for data visualizations. The Microsoft continuously releases updates for Power BI that adds customised features to the existing ones. In this paper, we try to analyse the performance of the data ware house by designing a dashboard that allows the users to view if all the data is loaded into the warehouse from the sources, which parts of data is missing, how many records are missing, how is the variance changing over time to identify if there is a leak in the system.

The data warehouse at XPO Logistics is maintained by executing the SSIS Jobs that perform ETL process and store the final output in the warehouse in a log table. We connect to this log table with the help connectivity feature of the Power BI. This connectivity allows the user to view data directly from the warehouse.

Just like any product that goes through the phases of software development, the dashboard generation also used “agile methodology”. The first step in any product development is the requirements gathering. Here the main purpose of the dashboard is discussed as the requirements of the dashboard. Once, the requirements are gathered, we have the designing of the product. In this process of designing a dashboard, firstly, we have done many brainstorming sessions for the dashboard where all the ideas are put forward together to be able to give user a better view of data. On analysing the pros and cons of the various design ideas, we have taken a final design that gives the performance of a warehouse just like a child’s report card is used to analyse their performance by the parents. The third and the most important step is the development phase where the product is brought to existence. The next phase is the testing phase where the data on the reports is tallied with the data in the warehouse. The final phase is the Deployment and maintenance of the product. This is the process of deploying the product to the customers/clients and is maintained by the development team through continuous monitoring of the product for any bugs. During the maintenance process, the client might add in new features to the product. When this happens, a new cycle is then started from the requirements gathering phase to the maintenance phase.

III. IMPLEMENTATION

The above discussed system is implemented using agile methodology and the following are the activities done in each phase of the dashboard development.

A. *REQUIREMENTS GATHERING*

The main purpose of the data warehouse performance analysis dashboard is to analyse how the records are loaded into the warehouse from the log table when SSIS jobs are executed to perform the load transfer from source to the warehouse. The main requirements of the dashboard are to analyse the following.

1) *Warehouse health:* If the record transfer from the sources into the warehouse is anything

a. greater than 97 percent then we say the health of the warehouse is A+
b. between 90 to 97 percent then we say it is A,
c. between 85 to 89 percent then B+
d. else B

2) *Identify percentage of tables with missing records:* evaluating the percent of tables from the warehouse have missing records.
3) *Identify the missing records:* From total percent of missing records tables, identifying the number of records each table is missing.
4) *Identifying the change in variance over time:* plotting a graph to visualize how the number of missing records is varying in the ware house.
5) *Filters:* creating filters for different tables in the warehouse to make the search for the tables more precise and easier.

B. *Designing*

After a lot of brain storming, the following visualizations are identified to be suitable for the given requirements.

1. Card visualizations for displaying the performance of the warehouse (ex: A+, A, etc.)
2. Pie chart to identify the number of tables having missing records and the number of tables having successful load.
3. Table visualization for displaying the tables with missing records and their number.
4. Line graph for showing the variation of the load over the time.
5. Check boxes for filtering the tables based on their types, transfer mode, etc.

With these visualizations, the dashboard is set to start with the development.

C. *Development:*

To start with the development of the dashboard, first we must look at the data. The data in this case is retrieved with a SQL statement. At XPO, the business intelligence tool used for designing dashboards, generating reports is the Microsoft Power BI. We can connect to the data from the Power BI tool with the help of “get data” feature of the tool that allows the users to connect data in the database, cubes, cloud, etc. Here, we use the SQL Query connectivity to connect to our data that is in huge database.

After a proper connection is established with the data warehouse, the data is loaded into the tool in the form of a spreadsheet that allows the users to preview the data before even starting with the visualizations.

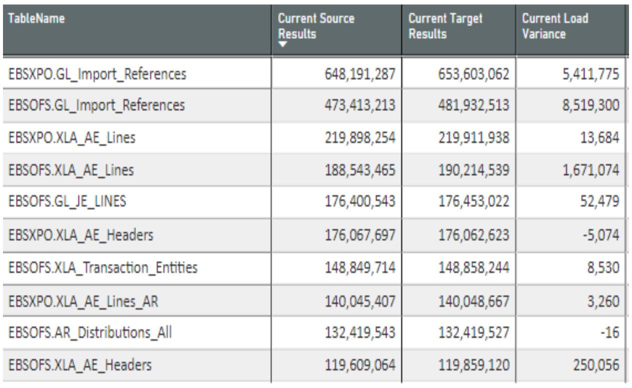
Like we create relationships between multiple tables in the relational database systems for extracting related data from various tables, the Power BI tool also offers a feature to create relationships between the tables in the tool. Here, since we are only one single log table of the warehouse, we do not have to use this functionality.

Once the relationships between the tables is established, we start creating our visualizations. In this process of building the dashboard, we may have to create new measures that help us with the necessary evaluations. For any visualization to work, we must drag and drop columns from the column pane of the tool to the visualizations pane of it. This is one added advantage of using Power BI tool as it is a simple drag and drop of columns and measures instead of writing some code for it to make it work.

Table Visualization:

We start with the table visualization since we have all the necessary columns for the table visualization available from the data that we have extracted. Dragging and dropping columns, arranging them are key aspects of any visualization. Here, in table visualization, for a user to identify which table has missing records, we place the table names in the first column, followed by the number of records in the source, number of records in the warehouse then followed by the number of missing records.

However, the data that we are extracting using the SQL query is the data over a time. But, if the user wants to view only the number of records that were missing with the previous day load, then we must create a new measure to view this data. Usually measures in Power BI tool are created using DAX (Data Analysis Expressions). The DAX expressions are like functions in excel. Using DAX, we created two new measures for identifying previous day's load and the number of records that are loaded into the warehouse on that load. Then, finding the number of records missing is just a subtraction between the two measures.



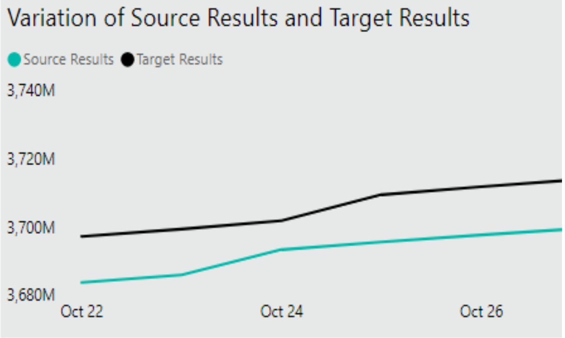
TableName	Current Source Results	Current Target Results	Current Load Variance
EBSXPO.GL_Import_References	648,191,287	653,603,062	5,411,775
EBSOFS.GL_Import_References	473,413,213	481,932,513	8,519,300
EBSXPO.XLA_AE_Lines	219,898,254	219,911,938	13,684
EBSOFS.XLA_AE_Lines	188,543,465	190,214,539	1,671,074
EBSOFS.GL_JE_LINES	176,400,543	176,453,022	52,479
EBSXPO.XLA_AE_Headers	176,067,697	176,062,623	-5,074
EBSOFS.XLA_Transaction_Entities	148,849,714	148,858,244	8,530
EBSXPO.XLA_AE_Lines_AR	140,045,407	140,048,667	3,260
EBSOFS.AR_Distributions_All	132,419,543	132,419,527	-16
EBSOFS.XLA_AE_Headers	119,609,064	119,859,120	250,056

Adding the new measures to the table visualization and arranging them fulfils the two basic requirements[A] of building the dashboard.

Line Graph:

The line graph is used to show how a parameter is changing. Here, in this dashboard, we use line graph to show how the number of missing records is changing over time. We don't have to create any new measures for this visualization as we have the number of records that are there in the source and the number of records that there in the target or the warehouse. Using the two columns, we draw a line graph and observe the gap between the two lines and how is it varying over time. If the gap is increasing over the time then, the number of records missing everyday is increasing and if the gap is decreasing then the number of missing records is going down. However, if both the lines overlap on each other then, there are no missing records for that table.

Now, to identify the variation for each table, we must make the line graph interactive with the table visualization that is created. The on and off the interactivity between the visualizations is done with the help of "interactions" option in the format ribbon that appear when clicked on a visualization.



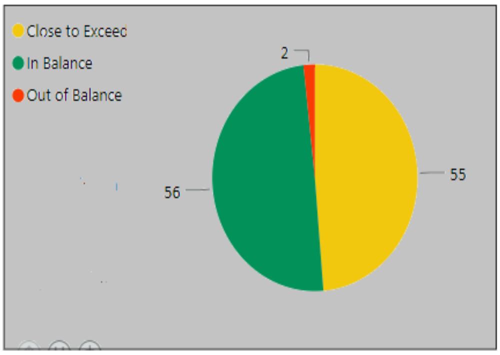
We can edit the interactions between the visualizations by clicking on the "Edit Interactions" on the format ribbon of the tool. On clicking the "Edit Interaction", two options appear on the visualizations. These two options allow the users to either filter the data of one visualization when the other is filtered or to block the data filter on one visualization when the others are filtered.

Pie Chart:

Though pie charts are the most confusing visualizations and the least used due to the lack of clarity on the data that they display, they can be used for small and simple data that does not need much understanding. Before the development of the dashboard is started, a threshold has been identified to classify the tables that have missing records into "in balance", "out of balance", and "close to exceed". This threshold is initially set to 1 percent that can be changed based on how the executives would like to view the dashboard. If any table has missing records greater than 1 percent of its source number of records, then we classify it to be a "out of balance" table. However, if the table has missing records but, it is less than 1 percent then we classify it to "close to exceed" and if there are no missing records then the table belongs to "in balance" category.

To depict the number of tables from the warehouse are "in balance", "close to exceed" and "out of balance", we use pie chart. However, this is a little complicated as new measures for identifying the threshold, classifying them into categories must be created.

The colours used for depicting the visualizations is very important. These help the user to identify the positives and negatives of a task easily. Usually "red" colour is used to depict something "bad" or "danger", "green" is used for "good" or "in best state". These are the two main colour codes that are followed in designing anything. Similarly, here, we have used "red" for "out of balance", "yellow" for "close to exceed", "green" for "in balance".



Just like line graph's interactivity is set with the interactivity of the table visualization, the interactivity of the pie chart is also set. However, one important thing to identify is that the pie chart should interact with the table and line visualization but must remain constant when clicked on a row in the table visualization. This is done by blocking the table visualization's interactivity with the pie chart.

Filters:

The Power BI offers a visualization that can be applied on classified data. Suppose, there is a data that is classified based on certain value in the data then we can use this feature that classifies the data to filter the data. In Power BI we have "Slicer" visualization that helps in filtering the data. This slicer can be modified into a check box filter, a list in a dropdown filter, button where the user can select the categories of data that they would like to view or a date filter that enables users to drag and leave the tip of the slicer at their required date.

Here, for the performance analysis of the data warehouse dashboard we are using three columns to filter the entire data on the dashboard. One is the "table name" column, where the users can directly select the table for which they would want to analyse. Two, the type of transfer i.e. from "Archive to Warehouse" or "Source to Archive". Third, is the schema of the tables. Since XPO Logistics have acquired a lot of companies in the past, they have classified the data based on the company they have acquired.

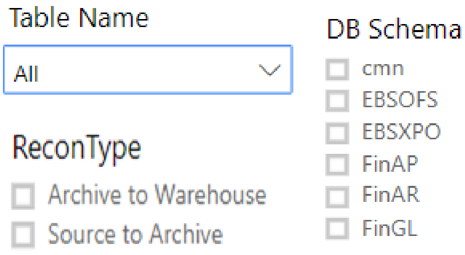


Table Name

All

ReconType

□ Archive to Warehouse

□ Source to Archive

DB Schema

□ cmn

□ EBSOFS

□ EBSXPO

□ FinAP

□ FinAR

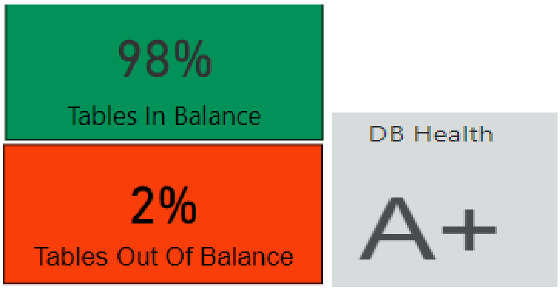
□ FinGL

With the slicer visualization of the Power BI, the developer can set the selection limit the options that the users can select in the filter. i.e. the single selects or the select all options could be set on the data.

Card Visualization:

Card visualization is just a simple text that appears on a card. We can set the background colour, font colour, font style for a card like the other visualizations in Power BI. Here, we have used cards to display the performance of the warehouse i.e. "A+", "A", etc., percentage of tables that are within the threshold and those that are exceeding the threshold.

To identify the percent of tables that are within the threshold, we create a measure that evaluates the percentage based of the entire warehouse. Based on the measure that is created, it is easy to determine whether the performance of the warehouse is "A+", "A", "B+" or "B". We use DAX to create a column to identify whether a table is within the threshold or exceeding the threshold.



98%

Tables In Balance

2%

Tables Out Of Balance

DB Health

A+

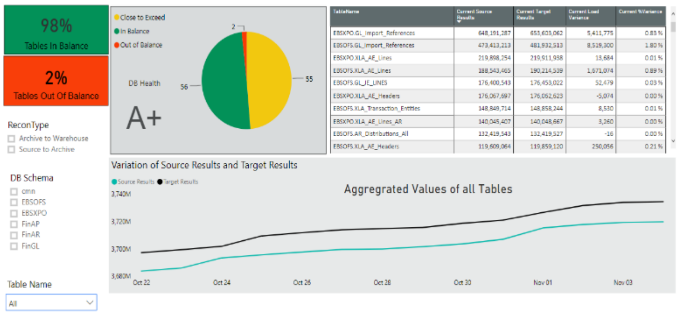
D. Testing:

The dashboard is then testing if the values shown are same as the values in the warehouse. In this case, we check the number of tables that are out balance in the warehouse by using a simple SQL statement that counts the number of tables whose threshold is greater than 1 percent. We also tested for the variance in the line graph manually. Once these tests are done and the dashboard does not fail in any of the tests, then the dashboard is published on the Power BI workspace.

The published app can be accessed in the form of mobile app if the dashboard is published to the Power BI app as well.

E. Deployment and Maintenance:

After thorough testing, the dashboard is published to the Power BI workspace. The final dashboard is as follows:



TableName	Current Source Results	Current Target Results	Current Load Variance	Current %Variance
EBSXPO.GL_Import_References	648,191,287	653,603,062	5,411,775	0.83 %
EBSOFS.GL_Import_References	473,413,213	481,932,513	8,519,300	1.80 %
EBSXPO.XLA_AE_Lines	219,898,254	219,911,938	13,684	0.01 %
EBSOFS.XLA_AE_Lines	188,543,465	190,214,539	1,671,074	0.89 %
EBSOFS.GL_JE_LINES	176,400,543	176,453,022	52,479	0.03 %
EBSXPO.XLA_AE_Headers	176,067,697	176,062,623	-5,074	0.00 %
EBSOFS.XLA_Transaction_Entities	148,849,714	148,858,244	8,530	0.01 %
EBSXPO.XLA_AE_Lines_AR	140,045,407	140,048,667	3,260	0.00 %
EBSOFS.AR_Distributions_All	132,419,543	132,419,527	-16	0.00 %
EBSOFS.XLA_AE_Headers	119,609,064	119,859,120	250,056	0.21 %

During the process of maintaining the dashboard there were new insights that the clients have questioned if we are able to draw those insights from the existing dashboard. The insights they have asked for are as follows:

1. Moving average
2. If the same dashboard could produce the results for different environment.
3. Also, to identify when the data is refreshed last time.

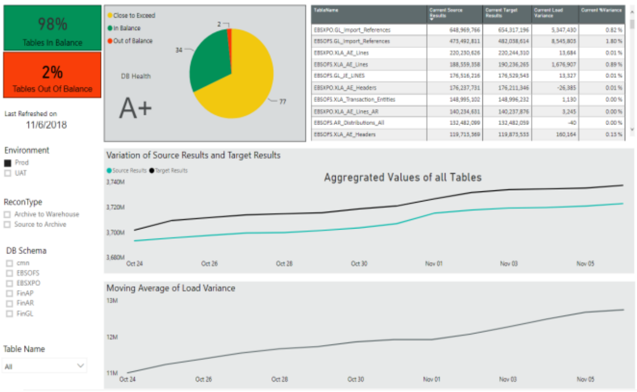
IV. DISCUSSION AND DASHBOARD ANALYSIS

On analysing the designed dashboard, the mentioned [E] new requirements are added. To be able to fulfil these insights, new visualizations must be added. When it was identified that new visualizations are to be added, a new version of the dashboard design has started taking the existing one as the base.

This new version of the dashboard helps users to understand when the data was last refreshed in the warehouse and the visualizations on the dashboard are based on that last refreshed data. We use a simple card visualization to implement this by finding the maximum refresh date from the data using DAX and placing that measure in the field for the card Visualization.

The other requirement of checking the system for leaks using the moving average is done with the help of a line graph. This graph helps the users identify if the number of records missing in the warehouse for a table is increasing daily or if it is decreasing or remaining constant. These variations in the graph acts like a red alert regarding the missing records. This helps them in handling the situation before getting it worse.

After these visualizations are added, the final dashboard is as follows.



TableName	Current Source Results	Current Target Results	Current Load Variance	Current %Variance
EBSXPO.GL_Import_References	648,969,766	654,317,196	5,347,430	0.82 %
EBSOFS.GL_Import_References	473,492,811	482,038,614	8,545,803	1.80 %
EBSXPO.XLA_AE_Lines	220,230,626	220,244,310	13,684	0.01 %
EBSOFS.XLA_AE_Lines	188,559,358	190,236,265	1,676,907	0.89 %
EBSOFS.GL_JE_LINES	176,516,216	176,529,543	13,327	0.01 %
EBSXPO.XLA_AE_Headers	176,237,731	176,211,346	-26,385	0.01 %
EBSOFS.XLA_Transaction_Entities	148,995,102	148,996,232	1,130	0.00 %
EBSXPO.XLA_AE_Lines_AR	140,234,631	140,237,876	3,245	0.00 %
EBSOFS.AR_Distributions_All	132,482,099	132,482,059	-40	0.00 %
EBSOFS.XLA_AE_Headers	119,715,369	119,875,533	160,164	0.13 %

V. CONCLUSION

Power BI is a powerful reporting tool that can connect to data in simple text/csv files, excel files, web (xml, json files, etc.), databases (SQL, Oracle, Microsoft Access, etc.), data cubes, cloud, etc. The reports designed in Power BI can be made more powerful with the help of strong DAX queries. Also, applying machine learning algorithms gives better visibility to the decision makers of the organization to view the predicted revenue or sales of the organization. The Power BI provides an option for the developers to design their dashboards using "R" language.

Like weather forecast, we can forecast the important features of an organization that the executives would want the estimations for. The Power BI primary tools include Power Query to extract and transform the data; Power Pivot to model and analyze the data; and Power View and Map to visualize the data modeled and analyzed. The Power BI Designer consolidates what were separate tools into an all-in-one application and removes dependencies to Excel or Office. PowerBI.com (sometimes called Power BI Site or Service) is used for sharing datasets, reports and dashboards. Microsoft Power BI supports data generated by the user, agile data analysis with self-service BI analytics which is managed in the cloud for collaboration and sharing with other Power BI users. This is different than a data warehouse, as Power BI permits:

1. User generated, agile extract inquiry and analysis (i.e. create calculated columns, change measures and dimensions on the fly and without IT involvement and rigid ETL tools); Data warehouses normally require data to be uploaded (via ETL programs) to the warehouse, then generate cubes, and then analyze, thereby making BI more rigid, IT dependent and difficult to iterate in short periods
2. Self-service BI with Natural Language Processing (NLP); and
3. Data can be shared from a central location as it is a graphical Designer tool with online publishing.

We summarize the advantages and the limitations of the Power BI as follows:

1. The ease of use of the Power BI tool. Initially, the users don't require any knowledge about how to use the tool. However, to generate powerful reports, they must have the knowledge of DAX scripting, power queries, etc.
2. The self-service feature with Natural Language Processing gives answers to the users' questions in the Q&A section of the tool.
3. The pace of innovation is great as updates for the visualizations or for the tool in generally are released monthly.
4. Apart from the basic visualizations that the tool provides, the users can add "custom" visuals from the Power BI community.
5. The in-memory analytics engine and columnar database are the technologies underneath the Power BI that supports tabular datastore structures used by Power Pivot. This achieves a balance between performance and ease of use.

Apart from the advantages mentioned above, we have issues that users are generally observing with the Power BI tool.

1. It is impossible to create entity specific dashboards as the Power BI does not use accept or pass user, account or other entity parameters.
2. Using the real time connections to get the live updates of the data has severe issues with the basic functionalities of the tool. For example, Power BI access to a single data source, voids the Edit View and eliminates key capabilities such as the Q&A and Quick Insights functions. Real-time connections to data sources other than SSAS also eliminate key behaviors such as DAX formulas.
3. Sharing of dashboards becomes difficult when the users email are not in sync with their office 365 emails.
4. One of the most important issue that is faced by the users of Power BI is the memory issue. Power BI does not accept files larger than 250 MB. Also, there is a limit of 1GB memory for a dataset.
5. This Microsoft solution is normally used to extend — not replace — other reporting tools. In most cases, it will not replace the enterprise data warehouse. For most companies, it is likely that their enterprise data warehouse tools will continue to be used for high volume data processing reports which do not change much, while Power BI may be used for one time,

progressive or more frequently changing analysis on smaller data sets.

Thus, we conclude that Power BI might not be a tool that replaces other reporting tools in the market but, for quick and faster generation of reports with simple data is easier with Power BI when compared to other tools as it is like the Microsoft Excel.

VI. ACKNOWLEDGEMENT

I would like to thank my advisor, Dr. Dazhao Cheng for his time and guidance without which this report would not have been possible.

VII. REFERNCES

[1] Kincaid R. Signallens: Focus+ context applied to electronic time series. IEEE Transactions on Visualization and Computer Graphics.

[2] Yang J, Fan J, Hubball D, Gao Y, Luo H, Ribarsky W, Ward M. Semantic image browser: Bridging information visualization with automated intelligent image analysis. InVisual Analytics Science And Technology, 2006 IEEE Symposium On 2006 Oct 31 (pp. 191-198).

[3] Yi JS, ah Kang Y, Stasko J. Toward a deeper understanding of the role of interaction in information visualization. IEEE transactions on visualization and computer graphics. 2007 Nov;13(6):1224-31.

[4] Baur, Dominikus, et al. "The streams of our lives: Visualizing listening histories in context." Visualization and Computer Graphics, IEEE Transactions on 16.6 (2010): 1119-1128.

[5] 32222Thudt A, Hinrichs U, Carpendale S. The bohemian bookshelf: supporting serendipitous book discoveries through information visualization. InProceedings of the SIGCHI Conference on Human Factors in Computing Systems 2012 May 5 (pp. 1461-1470).

[6] Stasko J, Görg C, Liu Z. Jigsaw: supporting investigative analysis through interactive visualization. Information visualization. 2008 Jun;7(2):118-32.

[7] Jonathan C Roberts: Sketching designs for information visualization the five sheet designs (FDS) approach.