International Journal of Engineering & Technology, 7 (4.5) (2018) 596-600

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Optimization and analysis of information using business intelligence techniques and reporting using dashboards

Ishwarappa Kalbandi [1] *, P. P. Hakarnikar [1], Mohana [2], H. P. Khandagale [3]

[1] Department of Computer Engineering, Dr. D. Y. Patil Institute of Engineering, Management & Research Akurdi, Pune, Maharashtra-411044
[2] Department of Telecommunication Engineering, R. V. College of Engineering Bangalore, Karnataka-560059
[3] Department of Technology Shivaji University Kolhapur, Maharashtra-416003
**Corresponding author E-mail:*

Abstract

In Today's world data is collected at an unpredictable scale from various application areas. Prior to the arrival of Big Data, all the data that was generated was handled manually. With data being produced in the range of terabytes today, that is impossible. To make the situation worse, almost 80% of the data generated by organizations is unstructured. This means that it cannot be understood in its avail- able format. It is very difficult and risky to make decisions just based on such crude data. In order to make quick, yet correct decisions, the generated data has to be optimized. This Paper discusses to create an end-to-end system to optimize approximately 6 million records of unstructured data provided as .txt files, which is in the form of strings and numbers into understandable or structured data. The next step is to analyse the structured data in order to make calculations on the given dataset. Finally, the analysed data will be represented in the form of dashboards, which are tabular reports or charts. In this Paper, unstructured data in the form of .txt files will be transformed into structured data in the form of tables through the SQL stored procedures in SQL Server Management Studio (SSMS). Along with the data, four other tables called dimensions will be created and then all five tables will then be integrated using SQL Server Integrated Ser- vices. Then an Online Analytical Processing (OLAP) cube is built over this data with product, customer, currency and time as its dimen- sions using the SQL Server Analysis Services (SSAS). At last this analysed data is then reported through dashboards through SQL Server Reporting Services (SSRS).The results of the analysed data is viewed in the form of reports and charts. These reports are customizable and a variety of operations can be performed on them as required by an organization. Since these reports are short and informative, they will be easy to understand and will provide for easier and correct decision making.

Keywords: *Big Data Analysis; BIS; Data Warehouse; OLAP; Optimization.*

1. Introduction

Information integration is active and challenging research area, despite significant progress made in the recent years. The volume of data available online and in electronicform has grown exponentially over the recent years, increasingthe significance of information integration for any organizational growth.Today 80% of the data generated by the organizations is primarily in the form of text. The data stored in these files may be unstructured or semi- structured. This unstructured data cannot be used for analyzing the data as they do not have a predefined format and cannot gain proper insights or any key business driving elements. Thus, this data has to be converted into structured data in order to analyze and then gain the insights from the analyzed data

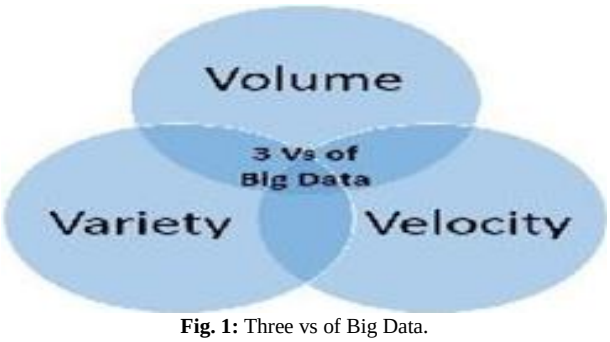

Fig. 1: Three vs of Big Data.

The primary purpose of the Business Intelligence (BI) solutions are due to the three main challenges in Big Data as shown in Fig .1 which are

Variety: The data is being generated by electronic systems in the unstructured, semi-structured or even a mixture of both formats, which makes the analysis of crude data difficult.

Velocity: The data is being generated at an unbelievable rate. The data was being generated only batch wise whereas now the data is getting streamed along with the batch wise data. This makes the generation of data unpredictable.

Volume: Nowadays the data generation has grown exponentially, Facebook alone generates almost 10TB of data every day.
In order to overcome these challenges the data generated has to be optimized through BI solutions, which will ensure that only the relevant information is being analyzed and provided to the right people at the right time.

2. Methodology

To organizations data or information is a commodity. To become useful information, data must be put into a specific busi- ness context. This means that data is one of the important factors on which the organization can make a significant progress in a short period of time. Not all the data or information generated by an organization or even related to that organization is important to all the people within the organization. The information makes sense or it is relevant to only the people who are involved in the decision making for the progress of the organization. Thus the delivery to the right people at the right time, right information within the organization is the key factor in achieving such a signif- icant progress.
In today's world almost 80% of the data generated is in the form of text. This data generated in the form of text is the main contribution for the generation of unstructured data. Due to the digital world data explosion is an inevitable trend. Without information, today's businesses can't function, Information must be crafted and made available to employees, customers, suppliers, partners and consumers in the forms they want it at the moments they must have it.To become useful or meaningful information, data must be put into a specific business context. the delivery to the right people at the right time, right information. Thus optimization of the in- formation is an essential for the business to be a success.

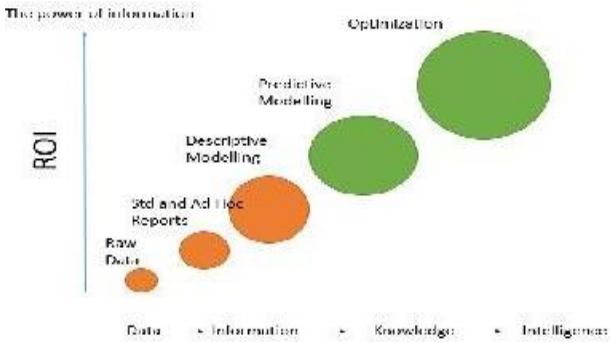

Fig. 2: Importance of Data Analytics.

The importance of the data analytics is graphically shown in Fig. 2. The figure shows the Return on Investment (ROI) against different levels of the analyzed data. The ROI based on the raw data or the unstructured data is very minimal. As the data keeps on changing its form the ROI keeps on escalating. The ROI will be at the peak when the underlying data is optimized to the fullest extent. Initially the data doesn't make much of a difference in decision making, but generating Ad Hoc reports on top of that data would lead to information. Similarly the data would give in much information such that the decision making can be done with ease.
The requirement of the Paper is mainly to gain insights
into the sales data like sales volume etc. for products of a Global Consumer Packed Goods (CPG) major and identify the location which made the lowest sales for improving marketing in that region and also to identify the best sales over a required customer, product and the time selection. The company currently tracks all its data through its Enterprise Resource Planning (ERP) solutions which serves the purpose of data gathering but does not allow the scope for analysis. A system has to be developed that focuses on the analysis of the data generated from the ERP system of the company. The system that has to be developed must be self- sustained and easy to use solution.

The ERP generated data for the invoices would be given in the form of an excel file. The customer and also the product files will be provided in the form of excel files. These different sources have to be integrated into one place and then the fact/invoice data has to be analyzed against the products and customers for a specif- ic time period.
The main aims of the Paper are:

- Optimize the crude data generated by the ERP systems.
- Integrate the received data into one database.
- Analyze the integrated and structured data using BI tools.
- Report the analyzed data through dashboards.

The data generated by the ERP systems for the invoices are given in the form of an excel file. The same holds true for the product and customer information

Fig. 3: Methodology.

The overview of the methodology is provided in the Fig.3. As the above figure depicts the data in the excel files will be first loaded into a data warehouse through the Extract Trans- form Load (ETL) tool. Here the text files will be picked up by the ETL tool, then the data in these files may be transformed accord- ing to the specifications and then loaded into the data warehouse or sometimes the data is directly dumped into the database as it is found in the file. Such transformed data is then cleansed or opti- mized or converted into a structured format and then stored in a different table. This is done at step 3, which is the SQL part. This kind of optimization is usually done through the stored procedures. Then the structured data will be analyzes by building an Online Analytical Processing (OLAP) cube. The main purpose of build- ing an OLAP cube is for the ease of access of data and then and also to improve performance of the system. Such analyzed data will be then reported back in the form of dashboards.
The entire end-to-end system will be built making use of
the Microsoft SQL services. The ETL tool that will be used for the purpose of data integration is SQL Server Integration Services (SSIS). SQL Server Analysis Services (SSAS) is used to build the OLAP cube for analyzing the data. Finally the reports and the dashboards that will be generated on top of the analyzed data will be built using the SQL Server Reporting Services (SSRS).
The Below Fig.4 shows the steps involved in extracting source files, cleansing them, performing aggregations on them and finally reporting them. The first step is to create a database to store the given data in. Then, tables are created and then populated using stored procedures which are executed in SSIS packages. Follow- ing that, the data is cleansed and then analysed in an OLAP cube. The analysed data is finally represented in tabular fashion or in the form of charts using SSRS.

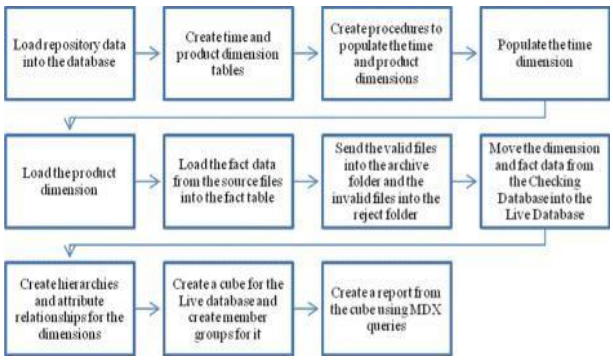

Fig. 4:. Block Diagram to Show General Sequence of Steps.

3. Design

As mentioned in the methodology, a step by step progress is done in order to achieve the aim as shown above. The Figure 3.1 represents the proposed design in order to achieve the Paper re- quirements and the objectives mentioned above. This figure also represents the BI landscape or the BI life cycle of a Paper. It first starts with the cleansing of the data and exists till the reporting of the analyzed data.
As the figure depicts the first step in achieving the objectives is to store the combined data in a warehouse, which is also known as Online Transactional Processing (OLTP). The next step would be to create the OLAP cube for analyzing with the benefits of ease of access of the data and also improving the performance of the whole system. The last step would be to create the reports based on the analyzed data. The reports can sometime be generated di- rectly on top of the OLTP systems without the existence of the OLAP cube.

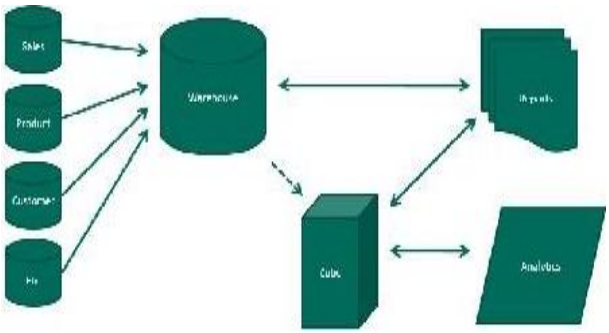

Fig. 5:. Design of an End-to-End System.

3.1. SSAS Cube

An SSAS cube is developed in order to analyse the information or the data. In SSAS the information that has to be analysed is called the fact and against which this information is analysed are called dimensions. The proposed solution has four dimensions namely, Calendar (time), Customer (Customer), Currency and Product. The time, product and customer dimensions were transferred to the POS_Live database from the POS_Checking database while the currency dimension was created in the POS_Live database itself since its data was not needed in any earlier stage.There are two fact tables - mart.tbl_fact_pos and mart.target. The values that needs to be analysed within the fact table are called measures and the collection of these measures is called a measuregroup.The advantage of building a cube is that it stores the data in a columnar fashion, whereas the SQL database stores the data row wise. This helps in the easy access of the data and also helps in the perfor- mance. The first step in creating the SSAS solution is to connect to a SQL database to fetch the data. This can be done by right clicking on the Data Source in the solution explorer and adding a new data source. In this Paper, the connection will be to the cut- lass database in the local server.

3.2. Cube deployment and processing

After the SSAS dimensions and the cube is built the SSAS solu- tion has to be deployed and the processed. The term Deploy means to bring in the structure of the cube in place. The term Process means to bring in the data into the deployed cube.The processing options for the dimensions are different from those of the processing options for the fact data. The two mainly used processing options for the dimensions are listed below:
Process Full – This option is selected whenever there is a struc- tural change in the dimensions. The structural change can represent the changes in the attributes or in the hierarchy, etc.
Process Update – This option is selected only when there is a change in the underlying data, but the structure remains the same.

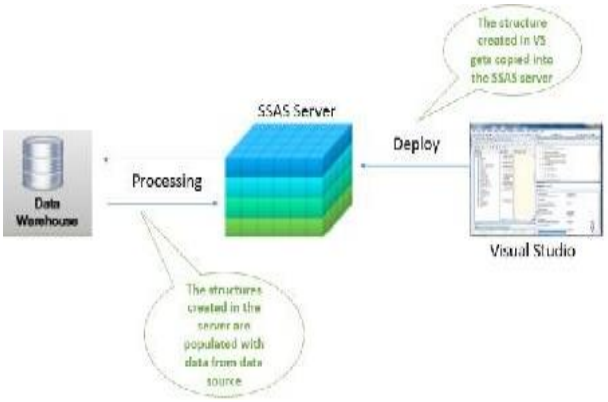

Fig. 6: SSAS Cube Deployment and Processing.

Fig. 6 shows the difference between the deployment and processing.The most frequently used processing options for the cube are:
Process Data – This option processes all the fact data by pulling in the latest information from the underlying table or the view.
Process Index – This option processes all the designed aggregation in the cube solution.
Process Full – This option processes both the data and the aggregations.
If the processing option for the cube is selected as Process Full, then the time during which the cube is processing, one cannot access the cube. But this can be overcome by doing a Process Data followed by Process Index as implemented in this Paper. By opt- ing this method, the cube can be accessed with the older data, until the new data is processed.

4. Results

Detailed information about different types of reports and charts is given along with information about datasets and parameters. Design and preview of the reports and charts are also shown. The results in this Paper are the dashboards or the reports that will be generated on top of the analysed data. The data that is analysed using SSAS can generally be viewed in two forms. The two forms are Excel Pivot Tables and Reports or Dashboards.
The analysed data can be viewed in excel in the form of pivot table. Excel can connect to the analysis services server and can fetch the data. One can connect to the SSAS server and database by going to the DATA pane in excel and selecting the option From Other Sources and under that selecting the From Analysis Services option. After selecting the server name and the database name has to be provided for establishing the connection and browsing the data in the form of pivot tables. However, in this Paper reports and dashboards are built to view the data and help the clients in taking their decisions easier and faster.SQL Server Reporting Services (SSRS) tool provided by Microsoft is used to build the reports and the dashboards. SSRS is chosen because, it enables the users to quickly generate the dashboards or the reports. It also allows the designer to consider the very minute detail in the report and cus- tomise it to the very extent.

4.1. Data sources

Data sources are the repository sources of data from which the reports can pull in the information and the numbers. The data source is a database within a data warehouse. There are two different kinds of data sources. They are Embedded Data Source and Shared Data Source.
An Embedded data source means that, the data source if specific to one report and no other report has access to that data source. Shared Data source is a kind of data source which allows all the reports under that SSRS solution to access the data from that data source. In this Paper a shared data source is used. It is a good practice to

have a shared data source for the entire solution, instead of having different data sources for each report.
In general one particular SSRS solution should have one data source, but can have different reports within one solution. And all those reports will access the data from one data source. As shown in Figure 3.1, there is an option whether to make the data source a shared one or not. If it is checked, it becomes a shared data source else an embedded data source.

4.2. Reports

In SSRS there are many kinds of reports through which the information can be delivered. The different types of reports available in SSRS are Tabular report, Matrix report, Charts, Sub-reports, Drill-through reports, etc. In this Paper Matrix and Table reports are built and used as the main report and the charts as the sub reports. A matrix is a kind of report, with one series on the X-axis and the other to the Y-axis.The series on these axes are dimensions. It can be one single dimension or a combination of them as hierarchies. The data to be filled, against the cross selection of the X-axis di- mension and the Y-axis dimension will be a measure from the fact table
In this the product hierarchy is on the X-axis. The higher levels of hierarchy for the product dimension are used as parameters to filter out the data as per the user's demands. So, the user can view products only for certain sector and/or certain categories as per his/her wishes. Also, since the parameters are based on a hie- rarchy, they are cascaded, i.e., the selections available for a para- meter of a lower level of hierarchy depends on the selections made for the higher level of the hierarchy. Thus, if a particular sector is chosen as a parameter, then only the segments under that sector should be available as choices for the segment parameter.
The Y-axis has the time hierarchy, with 5 drilldown levels. These levels are Year, Half Year, Quarter, Month, Week. The measure to be filled in, can be any one of the measures in the measure group. They are Retail Value, Retail Volume, YTD Value, YTD Volume.

4.2.1. Dataset

Datasets are the subsets of the data source from which the report will get the information and the numbers that has to be filled in the report. These datasets has to be retrieved from the data source by writing MDX queries in the dataset panel. An MDX query can be written as follows:
SELECT <Dimensions> ON ROWS,<Measures> ON COLUMNS
FROM Database_name
WHERE <Optional filtering clause>
In this query all the names, filters and databases will be corresponding to the SSAS server. Even the datasets are divided into two types. They are Embedded dataset and Shared dataset.
Embedded dataset is one which is specific to that report, whereas the shared dataset will be a common dataset for all the reports. It is a good practice to have embedded datasets rather than having shared datasets. Embedded datasets are used in this Paper. The datasets can be made dynamic with the help of report parameters. The changes in these values will change the underlying datasets for the required purpose. Each parameter that is created in a report will be associated with its own dataset. Thus not conflicting the values for the other parameters.
Parameters are created in order to add the filters for the reports. These parameters are used in the WHERE clause of an MDX query. The same holds good for the parameter cascading, i.e. one parameter will be used in the WHERE clause of another parame- ter's MDX query for filtering out the records.

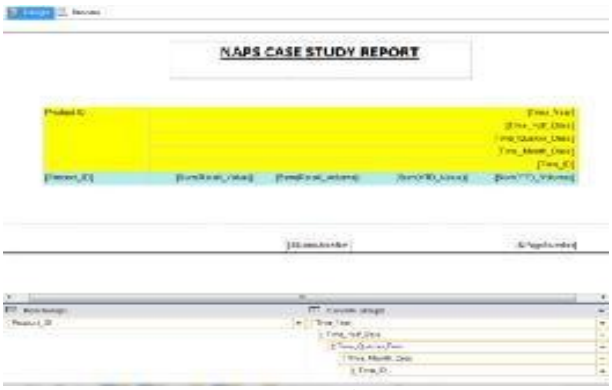

Fig. 7: A). Design Window for Matrix Report.

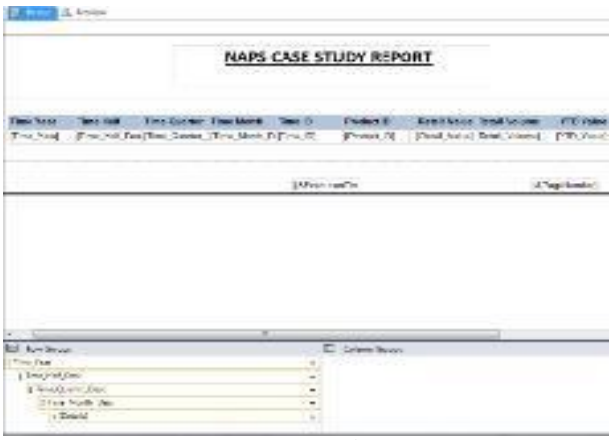

Fig. 7: B).Design Window for Table Report.

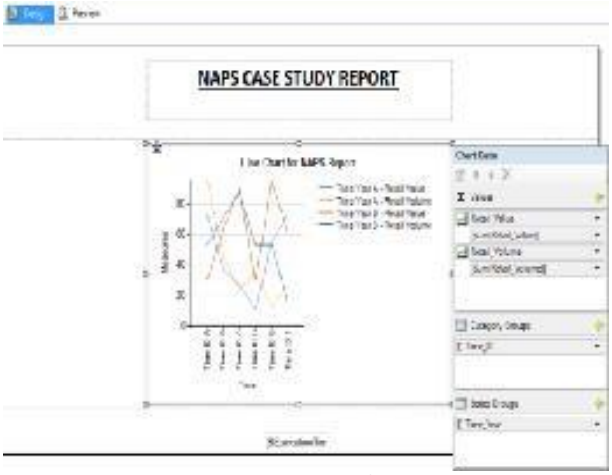

Fig. 8: Design Window for Charts.

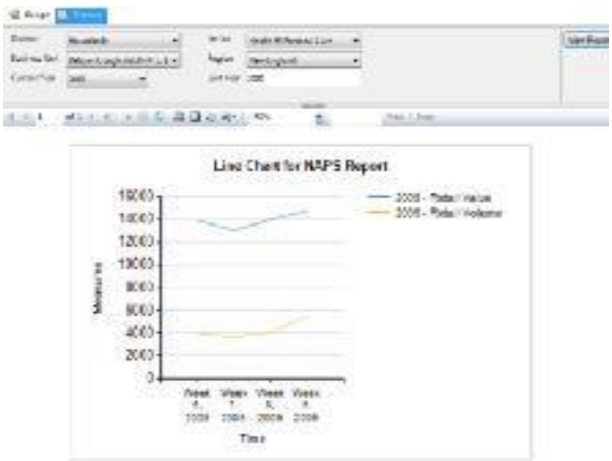

Fig. 9: Line Chart Preview.

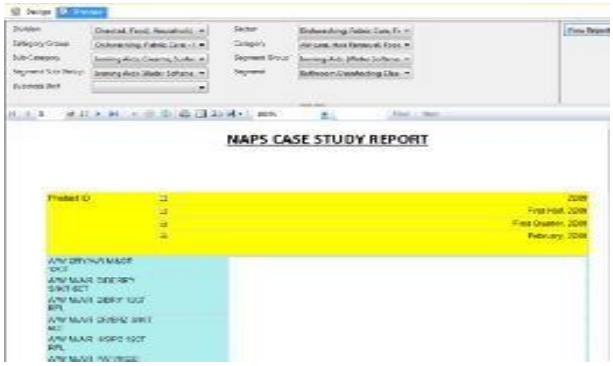

Fig. 10: A) Matrix Preview.

Fig. 10: B).Table Preview.

The Fig 7(a) shows the Design window of the matrix report and Fig 7(b) shows the same for the table report. The design window of the charts is shown in the Fig 8.
The preview of a chart is shown in the Fig.9. Previews for the matrix and table are shown in Fig 10(a) and Fig 10(b) respectively.

5. Conclusion

The paper provides an end to end development system of a Business IntelligenceSolution. The generated reports provides the information about the source data such as the YTD Value and the YTD Volume for a given time period which can be in years, half years, quarters, months or weeks. The transformation of the unstructured or the semi-structured data (present in the text files, excel files or in any other sources) into the structured data, which is stored in a data warehouse, increased the coherence of the data and enabled data to be extracted from it to generate the reports. Also, to increase the efficiency of the processing and deployment of the data, partitions were used so that only the data that was needed to be viewed would be processed. In addition to partitions, aggregations were used to increase the query performance of the processed data by 25%. Thus when the structured data is analysed, it provides the easier access to the data and also improves the per- formance of the system. This structured data is then put into the form of reports or dashboards, helps the decision making process easier and accurate.

References

[1] Agung W. Setiawan, NedyaUtami, Tati R. Mengko and AdiIn- drayanto, “Implementation of Electronic Medical Record in Com- munity Health Center Towards Medical Big Data Analytics Appli- cation”, 2014 IEEE International Conference on Electrical Engi- neering and Computer Science, 24-25 November 2014, Bali, Indo- nesia.
[2] Sara B. Elagib, Aisha-Hassan A. Hashim and R. F. Olanrewaju, “CDR Analysis using Big Data Technology”, International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering, 2015.
[3] Kun Wang, Yun Shao, Lei Shu, Chunsheng Zhu, and Yan Zhang, “Mobile Big Data Fault-Tolerant Processing for eHealth Networks”, IEEE Network, January/February 2016.
[4] Alfred Daniel, Anand Paul and Awais Ahmad, “Near Real-Time Big Data Analysis on Vehicular Networks”, 2015 International Conference on Soft-Computing and Network Security (ICSNS - 2015), Feb. 25 – 27, 2015, Coimbatore, INDIA.
[5] Ishwarappa, Anuradha J, “ A Brief Introduction on Big data 5Vs Characterstics and Hadoop Technology”,2015 International Conference on Intelligent Computing, Communication & Convergence (ICCC-2015),procedia computer science pp319-324.
[6] Aravindkumar D Gumtaj, H. V. Ravish Aradhya, Mohana, Gouri S Katageri “GPS and GSM Based Database Systems for User Access” International Association of Scientific Innovation and Re- search-IASIR, International Journal of Software and Web Sciences (IJSWS), pp.24-28, March-May 2015.
[7] Tong Wu, “ETL Function Realization of Data Warehouse System-Based on SSIS Platform”, IEEE 2010.
[8] Microsoft SQLCAT Team, “SQLCAT’s Guide to: BI and Ana- lytics”, e-book, 2013 Edition.