Available online at http://www.mecs-press.net/ijem

DataViz Model: A Novel Approach towards Big Data Analytics and Visualization

Rohit More[a], R H Goudar[b]

[a] *Department computer Network Engineering VTU, Belagavi 590018, India*
[b] *Department computer Network Engineering VTU, Belagavi 590018, India*

Abstract

Big Data is the collection of large data sets which contains large amount of data. There are different areas which are generating huge data, this data may be present in the form of semi-structured or unstructured and to get useful information from such raw data there is need of data analysis. Due to Big Data's excessive volume, variety, and velocity it is very difficult to store and process huge data. The process of extracting the information from such raw data is called Big Data Analytics. Big data Analytics processes data gives result in the form of structured data. Again this data is huge size and very difficult to understand since it is present in the form of CSV or excel or simple text files. So for effective decision making and to understand the information quickly the data need to be visualized as human mind understands images and graphs better and faster than text data. In this paper a model called Data Visualization (Viz) is designed which integrates big data analytics and the data visualization. This model first takes the data from various sources and then processes it and converts it into structured form, if want this data can be stored to RDBMS. Finally the text result can be visualized with the help of Visualization module of the DataViz. Here text result is represented in the form of charts and graphs.

1. Introduction

Big data is collection of large data sets which contains the large amount of data that may be present in the form of semi structure or unstructured. This is very difficult to process using on hand database tools or traditional processing applications. Big data is nothing but the large portion of unstructured or semi structured data available online and in real time manner. Sensors, network devices, medical equipments, and many other

* Corresponding author.
E-mail address:

are the sources generating large amount of data which is difficult to handle using the conventional processing approach.

Due to Big Data's properties such as volume, variety, and velocity it is very difficult to store and process and extracting the useful information from it. The process of extracting the information from such raw data is called *Big Data Analytics*. There are different tools and systems are available to analyze huge data sets and extract the useful information and generating the required result in the form of reports or text documents, which is again does not prove to be good approach when it comes to be used as a tool for decision making. The Data outputted from Analytic tools is in form of CSV, Excel or some text files, which contains huge text data and also become very difficult for understanding and also becomes difficult for decision making. So there is need of alternative way to represent the data in accurate and precise way. This problem can be solved by visual representation of data. So the data visualization is the one of the way to solve the above problem of data complexity. So to give better way to analytics the data visualization is the best approach.

2. Motivation

All when it comes to quick decision making the generated data has to be visualized along with the data analysis. There are different areas generating huge volume of data in the form of semi-structured or unstructured form. This need to be converted to structured form before storing to the databases and also need to be analyzed to get useful information out of it. Following are the areas which are in need of data analysis and also the data visualization for effective decision making.

Visualization of social network: Social network or sites are main stream of generating large amount of the data, the use of these sites are increasing day by day in promotions of different industrial products and used for generation of the feedback. Again this data is huge in volume and present in unstructured form. To get accurate and specific result from such data, the data need to be analyzed and visualized.

Visualization of Healthcare data: The healthcare data basically consists of data generated from four different fields such as claims and cost data, pharmaceutical and research & development (R&D) data, clinical data, and patient behavior and sentiment data. These all data present in huge amount and present in either unstructured or semi structured form. For effective decision making it is necessary to present this data into visual format such as graphs and charts. Here to get specific and accurate result big data analytics and data visualization can be integrated, where data will be first analyzed then obtained structured text result is represented in the form of graphs and charts.

Visualization of government Data: The government services, public sectors are known for generating huge volume of data thus big data analytics provides opportunities to analyze such huge data. There are many fields in government which are generating huge amount of data and one of the fields is tax. where huge data is generating this data is related to people who have paid income tax and the people who have not paid the income tax all records will be maintained as text report that has to be analyzed. To make effective decision out off this income tax data only data analysis cannot help, data need to be visualized so that quick decision can be made about the income tax has been paid or not.

Data visualization is the visual representation of data. Here obtained result from the data analytics is visualized in the form of graphs and charts, as human mind understands the images, graphs and charts much better and faster than the text reports. The data visualization approach helps a lot in decision making. So data visualization plays important role in big data analytics to provide qualitative and comparative results to the users. As there are huge datasets available and the generated information from analytics are not attractive and effective to support decision making. To achieve better decision making capabilities and attractive reports from the analytics, Data visualization can be integrated with big data analytics.

In this paper a system called *DataViz model* (section 3) is designed which integrates big data analytics and the data visualization. In-depth working of DataViz Model is explained in the following section.

3. Proposed System

In this section we are going to see about the design and working of the DataViz model which basically integrates the data analytics and the data visualization. The fig.1 shows the brief overview of DataViz model whereas fig.2 shows internal structure of DataViz model.
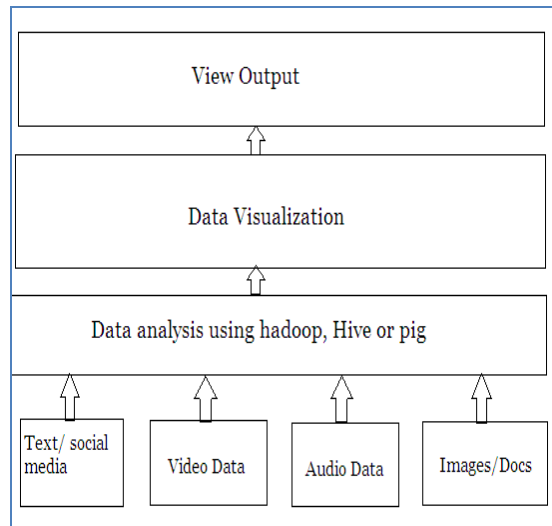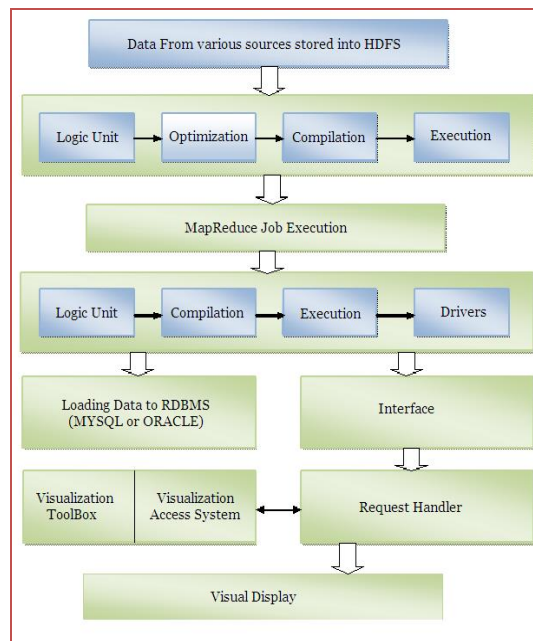

Fig.1.DataViz

Fig.2. Internal Structure of DataViz Model

Fig.1. shows the DataViz model which involves several steps such as data analysis and the data visualization. Here in DataViz model data from different areas is collected then first it is analyzed using different analytical tools so that raw data can be converted to structured data these tools can generate data in the form of CSV, excel or simple text file. Which is again huge data which is difficult to understand so data visualization helps to understand data? Here DataViz model allows data visualization of analyzed data. This data can visualize in the form of graphs and charts.

Fig.2. shows the complete working of DataViz model. The data from various sources is stored into Distributed file system such HDFS. The data may present in the form of unstructured or semi structured that need to be converted to structured form using any big data analytical tools such as Hadoop, Apache Pig, and Apache Hive. This data further can be stored into RDBMS such as MYSQL or Oracle Database. In above system fig.2 the data is first converted to structured form. For this data, schema will be defined so that data can be stored in raw and columns which can further visualized for effective decision making.

3.1. Big Data Analysis

This Step involves processing raw data to convert data to structured format and defining schema to structured data then if required storing the data into RDBMS. Big data analytics is the process of analyzing the large scale data available online which is unstructured or semi- structured that need to be transform to structured data to get the useful information which can help in decision making. The result obtained from the analytics is huge in size which can be again become difficult to understand therefore data visualization can help to represent data in much simplified way. The above proposed model integrates data analytics and the visualization so that the analyzed data can be represented in the form of graphs and charts.

Step 1: *Parsing Data Files to Structured form*

Here data from HDFS is loaded to Analytical tool for processing the raw data.

i.   *Logical Unit:* A required queries or code is written to convert the raw data to structured form.
ii. *Optimization:* Logical Plan is passed, which carries logical optimizations such as projections or pushdown.
iii. *Compilation:* compiles the optimization logic and converts it into series of MapReduce jobs.
iv. *Execution Unit:* here the MapReduce Jobs are submitted to analytical tool such Hadoop in sorted order.
v.   *Jobs Execution:* Finally MapReduce Jobs are executed on Hadoop and gives required result.

Step 2: *Defining schema for the structured data*

Here structured data is represented in the form of Rows and Column.

i.   *Logical Unit:* Queries are fired to define table and storing the data to that table.
ii. *Compilation:* Compile the fired queries.
iii. *Execution unit:* converts the compiled code into series of MapReduce jobs and finally these jobs are executed on Hadoop.
iv. *Drivers*: Drivers help in communication with external databases.

3.2. Visualization of Big Data

The analyzed data can either stored into RDMS or it can be further sent for the visualization. Data visualization is process of representing data in the form of graphs and chart. In DataViz model (Fig.2) visualization module does the work of visual analytics.

Here data present in the text form is processed using the available visualization tools that convert text result

in the form of charts and graphs. These graphs or charts can dynamically changes as per user convenience in the form of bar chart, pie chart or line graph etc. The result can be further elaborated with effective graphical capabilities.

There are various steps involved in the process of visualization.

i. *Interface:* It is also called as Intelligent Interface which can receive the data from users or above analytical system and gives it to the request handler.
ii. *Request Handler:* This request handler takes the data and forwards data for visualization to visualization Access system.
iii. *Visualization Access system*: Gives access to different Tools used for the Visualization.
iv. *Visualization Toolbox:* Consists of visualization tools and resources required for the visualization.

4. Related Work

All In this section we are going to look at the work done by different authors on the big data analytic and the use of Data Visualization and its use in effective decision making. With help of Data visualization how the result of analytics can be represented in the form of visual effect such as graphs and charts.

Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler[2] here authors descibed the scope of visual analytics and what all the things that visual analytics consists off such as information analytics , geospatial analytics, knowledge discovery etc. Also described the visual analytic process, how the visual analytic work, Also mentioned about the application challenges of the visual analytics. "Physics and astronomy, business, security, disaster and emergency management, Healthcare and medical information" all these areas requires visual analytics, Also presented the Technical challenges of visual analytics.

Pravin Chopade, Justin Zhan, Kaushik Roy, Kenneth Flurchick[3] authors designed system called "X-SimViz" which allows user to do dynamic big data analytics and visualization. This system is basically used for the "real time large scale network data". Also explained about the "challenges and the complexity" of the big data analytics and data visualization.

Yingjian Qi, Xinyan Yu, Guoliang Shi, Ying Li [4], in this paper, how the text visualization can help to visualize the geographical information is explained, also described "text visualization basic framework" that can help "micro- blog data visualization under text visualization " and studies the characteristics and social significance of text visualization.

Qunchao Fu, Wanheng Liu, Tengfei Xue, Heng Gu, Siyue Zhang, Cong Wang [5] This paper presents the data processing method called binned aggregation for the visualization of big data and data reduction methods for visual summaries of data, And how binned aggregation can be used for the visualization.

The above discussion tells about the work done by different authors in the field of big data analytics and the visualization. And also explain about the challenges, scope and characteristics of visualization. Some authors described the methods for the visualization. And some are explained about the methods of big data analytics but they lack in describing the methods that combines both analytics and visualization together. Hence above model (section 2) can help in solving the problem of data analytics and the visualization.

5. Conclusion

All in this paper DataViz model is designed which integrates Big Data analytics and the Data visualization where in raw datasets can be analyzed and the text report will be represented in the form of graphs and charts. If this model can implemented then it can give better way to analytics and effective decision making. This model can give specific and precise output in the form of visual effects. Thus this model can be the better way to big data analysis.

References

[1] Ishan Sharma, Rajeev Tiwari and Abhineet Anand. "Open Source Big Data Analytics Technique", International Conference on Data Engineering and Communication Technology Volume 468 of the seriesAdvances in Intelligent Systems and Computing, pp 593-602, 2016.

[2] Daniel A. Keim, Florian Mansmann, J¨orn   Schneidewind, Jim Thomas, and Hartmut Ziegler "Visual Analytics: Scope and Challenges", Visual Data Mining, pp 76-90, 2008.

[3] Pravin Chopade, Justin Zhan, Kaushik Roy, Kenneth Flurchick. "Real-Time Large-Scale Big Data Networks Analytics and Visualization Architecture", 2015 12th International Conference & Expo on Emerging Technologies for a Smarter World (CEWIT), pp 1-6, 2015.

[4] Yingjian Qi, Xinyan Yu, Guoliang Shi, Ying Li. "Visualization In Media Big Data Analysis", 2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS), pp 571-574, 2015.

[5] Qunchao Fu, Wanheng Liu, Tengfei Xue, Heng Gu, Siyue Zhang, Cong Wang. "A Big Data Processing Methods for Visualization", 2014 IEEE 3rd International Conference on Cloud Computing and Intelligence Systems, pp 571-575, 2014.

[6] Zipeng Liu, Zhenhuang Wang, Siming Chen, Zuchao Wang, Zhengjie Miao, Xiaoru Yuan. "A Platform for Collaborative Visual Analysis on Streaming Messages", 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), pp 375-376, 2014.

[7] Mark Shackleton, Fadi El-Moussa, Robert Rowlingson, Alex Healing, John Crowther, Joshua Daniel, Theo Dimitrakos, and Ali Sajjad. "Deploying Visual Analytics Through a Multi-cloud Service Store with Encrypted Big Data", On the Move to Meaningful Internet Systems: OTM 2016 Conferences Volume 10033 of the series Lecture Notes in Computer Science , pp 883-889, 2016.

[8] James Davey, Florian Mansmann, J¨orn Kohlhammer, and Daniel Keim "Visual Analytics: Towards Intelligent Interactive Internet and Security Solutions", The Future Internet Assembly FIA 2012, pp 93-104, 2012.

[9] Divanshu Gupta, Avinash Sharma, Narayanan Unny, and Geetha Manjunath. "Graphical Analysis and Visualization of Big Data in Business Domains", Big Data Analytics Volume 8883 of the series Lecture Notes in Computer Science, pp 53-56, 2014.

[10] Nascif A. Abousalh-Neto, Sumeyye Kazgan "Big data exploration through visual analytics", 2012 IEEE Conference on visual Analytics Science and Technology(VAST), pp 285 – 286, 2012.

[11] Sachchidanand singh, Nirmal Singh " Big Data analytics", 2012 International Conference on Communication, information & computing Technology(ICCICT), pp 1 – 4, 2012.

[12] David Jonker, scott Langevin, peterSchretlen, Casey Canfield "Agile visual analytics for banking cyber big data", 2012 IEEE Conference on visual Analytics Science and Technology (VAST), pp 299 – 300, 2012.

[13] Shinnosuke Takeda, Aimi kobayashi, Hiroaki Kobayashi, Saori Okubo, Kazuo Misue "Irregular Trend Finder: Visualization tool for analyzing time-series big data", 2012 IEEE Conference on visual Analytics Science and Technology(VAST) , 305 – 306, 2012.

[14] Daniel Keim, Huamin Qu, Kwan-Liu Ma "Big Data Visualization" IEEE computer Graphics and Application, volume 33, pp 20 – 21, 2013.

[15] Bapin Bihari Jayasingh, M. R. Patra, D Bhanu Mahesh "Security issues and challenges of big data analytics and visualization" 2016 2nd International Conference on Contemporary Computing and Inforatics(IC3I), pp 204 – 208, 2016.

Authors' Profiles

Rohit More is currently a research scholar in department of computer network engineering, VTU, Belagavi. His area of interest in Big Data Analytics. He has published papers in International Conferences and Journals.

Dr. R H Goudar, currently working as an Associate Professor, Dept. of Computer Network Engineering, Visvesvaraya Technological University, Belagavi. He has 13 years of Teaching Experience at Professional Institutes across India. He worked as a faculty Member at International Institute of Information Technology, Pune for 4 years and at Indian National Satellite Master Control Facility, Hassan, India. He published over 145 papers in International Journals, Book Chapters and Conferences of High Repute. He has guided over 140 M.Tech Dissertation and 04 ongoing Ph.D Students. He has received various awards like Outstanding Faculty Award, Research Performance Award, Young Faculty Award from VIFA, and Young Research Scientist Award from VGST Karnataka. He has received research grants from AICTE, UCOST and VGST, Karnataka. He has received over 751 citations for the work in subjects of Interest includes Semantic Web, Cloud, Big Data, Network Security and Wireless Sensor Networks.