ISSN No. 0976-5697

Volume 7, No. 6(Special Issue), November 2016

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A BUSINESS INTELLIGENCE BASED NOVEL FRAMEWORK FOR BIG DATA ANALYSIS

Chandanjit Kaur
Student (M.Tech),IKG PTU,
catherine261992@gmail.com

Er. Prabhdeep Singh
Asst. Professor, IKG PTU,
ssingh.prabhdeep@gmail.com

Dr. Sandeep Singh Kang
Professor, IKG PTU,
hodcse@globlinstitutes.org

Abstract- *Due to the different connotations (ideas) of Big Data, there are different definitions of Big Data. In today's technical era, the advent of technology leads to an explosive increase in digital data (like posts on social media, transaction records, GPS signals, digital images and videos etc.). The Data can be structured, unstructured (like XML) and semi-structured. It is easy to manipulate structured or traditional data that centralized in nature and present in Gigabytes, Terabytes whereas the Big Data which is distributed in nature and present in Petabytes (10^{15} bytes), Exabytes (10^{18} bytes), Zettabytes (10^{21} bytes or 10^{12} GB), Yottabytes (10^{24} bytes) raise the data warehousing cost. Big Data refers to enormous datasets which act as a key to many business operations. Therefore, there is a need to utilize such enormous data that are formidable (difficult) to gather, manipulate and processed by traditional information technology and to ensure its efficient storage. This paper demonstrates a close-up view of Big Data visual analytics as well as technology adapts to deal with its problems like data deluge (a large amount of data come at the same time).*

Keywords: *Big Data, Business Intelligence, Big Data Analysis, Visual analysis, Data cleaning.*

I. INTRODUCTION

This is a data-driven era, in which data is pursuing to accumulate and is being accumulated at increasing rates. Every day quintillion bytes of data generated every day. This massive growth of information is mostly contained unstructured data and overall complex.

In the era of information, the growth of different fields and sectors has made the large growth of digital data which arises many challenges like inconsistently capturing of the massive amount of data, its analysis, storage and visualization which results in impending or slow down the processing of data. Big Data helps in improving the knowledge of business intelligence.

The collection of data is done via various sources like sensors used for various purposes, social networking sites, internet surfing, log mining, transactions from e-commerce, GPS communication [1]. The continual use of social networking site or e-commerce generates billions of data. This data is either in the form of photos or videos which are unstructured data[2].

Also, Big Data changes the way of doing work in different domains due to which the size of data across the world doubles every 1.2 years[3]. Due to this reason, there is needed to analyze the Big data implementation and execute it accurately.

Many organizations analyze the large chunk of data and extract required information by surmounting software tools so that they can perform various actions in a short time[4].

Big Data played its role in a number of fields. Everyone took akeen interest in this and they all are aware of its usage and know how to take benefit from it and enhance their performance. Big Data is kept on expanding from year to year. Does anyone know the reason for this? The accumulation of information each and every year through various sources like Facebook, Twitter, Instagram, Google, and much more could be the main reason of its expansion [5]. According to Gartner, there will be billion or trillion of devices on this planet by the next coming four to five years [6].

To address challenges of Big Data various research programs is launched by the government agencies[7]. 3 main principles of Big Data are velocity, variety, and volume. It is difficult to explore the patterns and overall view of the data architecture with the help of these principles.

Figure 1: Forms of Big Data

Therefore, various mechanisms of Big Data are used for better understanding of large datasets, i.e.visual analytics for Big Data[1]. The term Visual analytics mean survey at large scale and handling of information. Its aim is to detect hidden patterns anddisplay it to the user. It takes less time to insight and makes direct interaction with information. It is efficient in presenting the required information in a large amount of data as well as also derives complex analysis[7].

An increase in the use of social network causes the continuous accumulation of data in enormous amount. The data of these social networks are complex, noisy and written in informal languages[8]. This is the main problem for analyzation of social media data [2]. Therefore, it is necessary to manage such kind of data.

In this paper, we analyze the dataset. The analysis is the crucial phase of any research which helps in generating the values from a large volume of data so that decision making is done[9]. Analysis of Big Data which is large in size can be done so that the user can rapidly extract values[10].

Our approach is to use R statistical software to visually analyze the Big Data. The proposed method contains 3 steps which are described in section 2. Section 3 contains the information about R statistical software. The visual analysis of Big Data is described in section 4 and in the last section contains conclusion of this paper.

II. RELATED SURVEY

Big Data changes the life of various industries, professionals and researchers[5]. It has significance for various fields and sectors like at the national level, industries upgradation, science, and interdisciplinary research, and also helps the people to predict future[11]. I search a number of reputed journals, articles, web sources and much more to obtain information about various factors of Big Data. After selecting particular research papers related to my work,then Igenerate a report. We conclude the results in the form of a graph as shown in thefollowingdiagram. This graph contains the information about various searches on different factors of Big Data from the year 2012 to 2016.

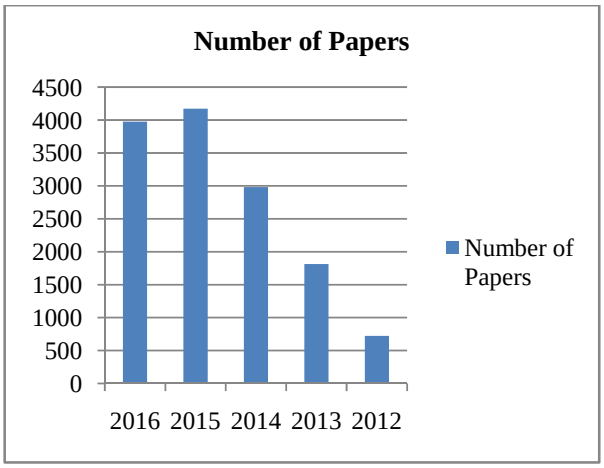


Figure 2: Number of searches in different years on various factors of Big Data

III. PROPOSED METHOD

The proposed method contains 3 steps- Data Capturing, Division of Data and Data Scrutiny. The detail of these steps described in the rest of the section.

Figure 3: Proposed Methodology

A. Data Capturing

Initially, we search a number of sources to get a dataset and then we find the dataset named as a superstar. It is a dataset of the online shopping portal that contains information about the customers as well as the products and its various details like row and order id, order priority, order quantity, sales, discount, shipping mode and cost, profit, product name and category, and much more.

B. Division of Data

This step contains thedivision of data into smaller chunks by using various sets of rules and operations and so that we obtain such kind of data that can be easily managed and clarify.

C. Data Analysis

In this step information obtained from first and second step is analyzed and its graphical representation is done with the help of R statistical tools which helps in interacting with data. It has a graphical facility for analysis and generates the result. It is open source software used for statistical computing and graphics. It is also a programming language with built-in functions. The user can insert functions as a part written in other languages. With the help of this sophisticated graph is generated. It offers multidimensional data with multi - panel charts and also offers 3D surfaces.

D. Objectives

Here we define the main objectives of our proposed work.

a) Find the Product Category which is mostly liked.
b) In a particular year, which shipping mode is most expensive?
c) When ordering priority is high which shipping method is preferred?
d) Which Product provides maximum profit?.
e) Find the maximum sale in a particular year.

IV. DATASET

It is a complete set of data which contain tables along with different fields and shows the relation between the tables. It also contains the information which is gathered during research or survey. It is the information arranged as a stream

978-93-85670-72-5 © 2016 (RTCSIT)

of bytes in the record. The organization of each record is same as well as uniform throughout a dataset.

Number of entries: Dataset we used for research contain nearly 8,57,96,432 entries which contain various fields as defined in the first step.

Extension: Their exit number of extensions like .txt, .Cs, .XLs, .Xml, .ppt, .PPS, .Dat and so on but from them, we give preference to .Csv extension which stands for comma separated values. It helps the data to be saved in a table or relational or structured format.

V. IMPLEMENTATION

In this paper, we visually analyze the Big Data with R statistical software and we substantiate its results.As we know in today's modern world e-commerce providesa complete range of benefits to different users.A number of enterprises show their interest in doing business online due to increasing demand for purchasing products online. Users can achieve anything online.

In the proposed method we initially fetch data and perform various setsof operations over it to obtain easily interpreted data. In the last, we analyze the information obtained and generate reports in the form of a graph.

VI. RESULT

This section contains the results which are obtained by performing operations using R statistical tool to obtain the visual analysis of the dataset in the form of graphs.

Result 1: In this, we find the product category which is mostly liked. The dataset contains three product categories as shown below. After applying operations and performing calculations, we obtain the result as shown in the below figure.

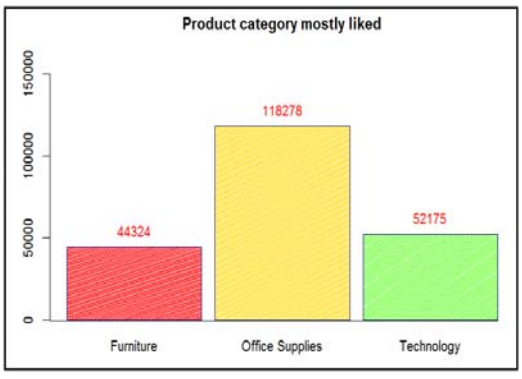


Figure 4: Result of objective (a)

Result 2: In this, we show the maximum cost used for shipping cost in different years. Basically, our dataset contains data nearly 5 years from 2011 to 2015.

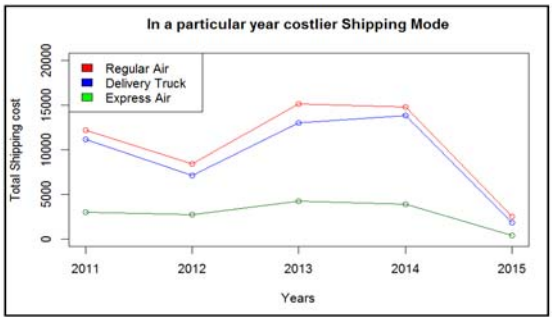


Figure 5: Result of objective (b)

Result 3: This result shows the preferred shipping mode when order priority is high. It concludes that Regular Air Shipping mode is highly preferred and the delivery truck has a lesser priority.

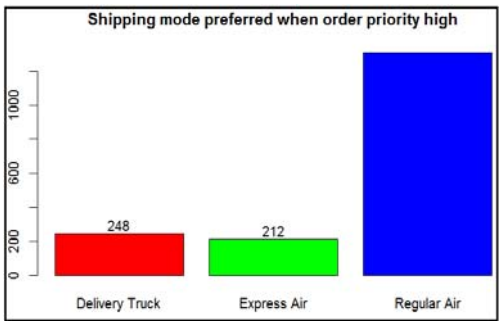


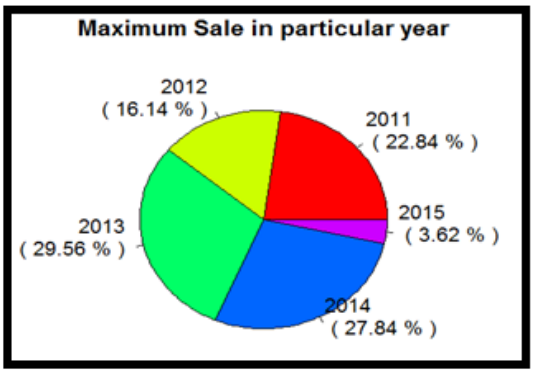
Figure 6: Result of objective (c)

Result 4: This result shows the product which gains a maximum profit of Rs. 272200.69/-. It doesn't contain the graphical value because we have to find the only single product.

```
#Product with highest profit
c()
] "Videoconferencing Unit  is the Product
| with highest profit and its profit: 27220.69"
```

Figure 7: Result of objective (d)

Result 5: It shows the total or maximum in each year with the help of the pie chart and it also shows what percent sale was done in each year.



978-93-85670-72-5 © 2016 (RTCSIT)

Figure 8: Result of objective (e)

7. CONCLUSION

Factors	Traditional Approach	Proposed Approach
Decision Making	✗	✓
Visual Analytics	✗	✓
Large Amount of Data	✗	✓
Storage	✗	✓

Table 1: Factors involvedin this Research

An increase in the growth of information in different fieldsraises many challenges. Big Data is growing at an immense and its technologies are accepted over a large area by different sectors and fields to unveil the hidden values of enormous data.In this paper,we visually analyze the dataset of online retailing, which contain information about the customer and product id, product quantity,product category and much more. According to the above framework, we initially capture the data from web sources, then perform parsing and data scrutiny with the help of R statistical software and generate results as shown in section VI.

8.REFERENCES

[1] Young-Ho Park Aziz Nasridinov, "Visual Analytics for Big Data using R," in *2013 IEEE Third International Conference on Cloud and Green Computing*, Karlsruhe, 2013, pp. 564 - 565.

[2] Srinath Srinivasa Vasudha Bhatnagar, "Big Data analytics," in *Second International Conference, BDA 2013*, Mysore, India, 2013.

[3] Chun-Yang Zhang C.L. Philip Chen, "Data-intensive applications, challenges, techniques," *Elsevier*, 2014.

[4] M. Haider A. Gandomi, "Beyond the hype: Big data concepts, methods, and analytics," *Big Data Research*, vol. 35, no. 2, pp. 137-144, April 2015.

[5] Marco Greco and Michele Grimaldi Andrea De Mauro, in *International Conference on Integrated Information, AIP Conference Proceedings* , Madrid, Spain, 2014.

[6] (2014) Gartner. [Online]. http://www.gartner.com/newsroom/id/2684616

[7] Huamin Qu, Kwan-Liu Ma Daniel Keim, "Big-Data Visualization," 2013.

[8] Aleksandr Ometov , Yevgeni Koucheryavy and Thomas Olsson Ekaterina Olshannikova, "Visualizing Big Data with augmented and virtual reality: challenges and research agenda," *Journal of Big Data*, 2015.

[9] D. Culler, K. Pister, and G. Sukhatme D. Estrin, "Connecting the physical world with pervasive networks," *IEEE*, vol. 1, no. 1, pp. 59-69, 2002.

[10] ShiwenMao, Yin Zhang, Victor C.M. Leung Min Chen, "Big Data Related Technologies, Challenges and future prospects," *Springer*, 2014.

[11] Benjamin W. Waha, Xueqi Cheng, Yuanzhuo Wang Xiaolong Jin, "Significance and Challenges of Big Data Research," *Big Data Research*, vol. 2, no. 2, pp. 59-64, Feburary 2015.