International Journal of Engineering & Technology, 7 (1.9) (2018) 173-177

International Journal of Engineering & Technology

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

Research Paper

Data analytic framework for crime sector using open grid rule generation algorithm

M. Sharmila begum [1] *, A.George [2]

[1] Research Scholar, Department of Computer Science and Engineering, Periyar Maniammai Institute of Science and Technology
[2] Professor, School of Humanities Science and Management, Periyar Maniammai Institute of Science and Technology
**Corresponding author E-mail: sharmilagaji@gmail.com*

Abstract

Data analytics (DA) is the process of exploring datasets in order to illustrate conclusions about the information they contain. It is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information. A database contains both structured and semi-structured data. The semi-structured data are from different sources. The system is provided with an open grid rule generation for analyzing the data from whole data container. According to this concept, the analysis is much absolute rather than other, data mining technique. The main objective of the proposed study is to provide data having better and significant perspective.

Keywords: *Data Access; Data Analytics; Data Filtering; Mapping Data; Open Grid Rule Generation (OGRG) Algorithm.*

1. Introduction

Data Analytics has become one of the most discussed topics among researchers. The term Big Data is used to refer the enormous amount of datasets. Compared to the traditional datasets, big datasets comprised a set of unstructured data which requires more significant real-time investigation [1]. The concept of big data helps us to understand the abstraction and in-depth understanding behind various hidden values. Data analytics is a technology-enabled strategy for gaining richer, deeper, and more accurate insights into the details of the customers, partners and the businesses and hence ultimately gaining competitive advantage. By processing a steady stream of real-time data, organizations can make time-sensitive decisions faster than ever before, monitor emerging trends, course-correct rapidly and jump on new business opportunities [2].

1.1. Structured data

Structured data is the data included in relational database system. Being structured and highly organized, it can be managed by SQL and its multiple variations developed by IBM, ADO.net, ODBC in RDBMS environments, Due to explicit semantics and structure, efficient search is possible for focused content by simple and straightforward search engines [3].

1.2. Semi structured data

Semi structured information is one kind of organized information which is not having the information display structure, Further it will accommodate a formal or unbending structure. This semi structured information does not require a composition definition as it is fairly discretionary and contains labels or different markers to isolate semantic components and authorize pecking orders of records fields inside the information .To change over the semi structured information to structured information, customary information mining systems or normal preparing dialect is relevant [4].

1.3. Data analysing

It is observed that nowadays, various organizations are consuming large amount of essential information which can be useful in various other fields such as monitoring of an objects activity, sensor deployment, tracking of data etc. A severe flood associated with data is termed as Big Data, which yields to the challenging situation on the present infrastructure of Data Storage management and the Statistical estimation of data, Analogous situations arises when an organization wants to explore its data from its personal websites for analyzing the customer's feedbacks, customized services towards a product [5].

As a result, the decisions makers would convey their conclusions grounded on the analysis of extracted data or those data which carry some value or weight age [6]. Data analytics is also mapped between unrelated attributes of datasets which can be obtained from machine learning, database systems and statistics. In a variety of scientific fields, grid had been developed with storage, processing, and availability of data. A grid is a collection of distributed computing resources available over a local or wide-area network that appears to an end user or application as one large virtual computing system. It can be seen that some specific technologies and implementation are required for the cloud for providing infrastructure, platform and software sources [7].

The collection of digital information in terms of structured and non-structured data known as big data is rapidly developing. Big data is definitely a phenomenon with direct impact on quality of life. Applications of big data can be found in mobile cloud computer systems, such as divisions of purchase transactions social networks, teacher commentary, e-science and healthcare systems. Analysing data in order to summarize them and look for patterns is an important part of every evaluation. Strategies for the analysis of the data and how the data will be synthesized should be decided at the evaluation design stage. By processing a steady stream of real-time data, organizations can make time-sensitive decisions faster than ever before, monitor emerging trends, course-correct rapidly and jump on new business opportunities [8].

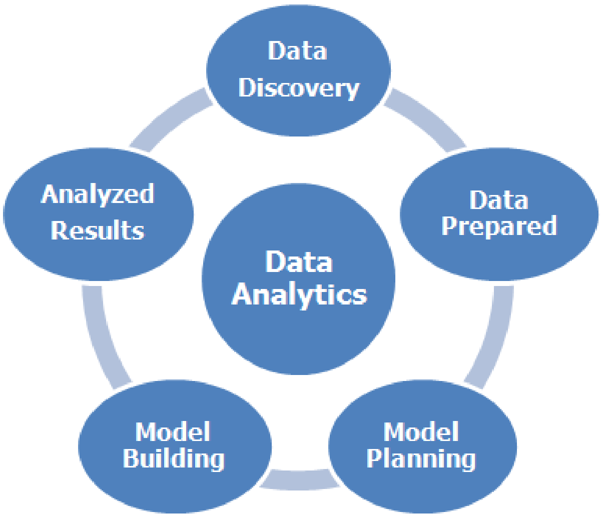


Fig. 1: Life Cycle of Data Analytics.

2. Related work

Chris Clifton, (2002) proposed To continue this proposed work, some survey is made on the organization of semi structured data which is recognized as one of the major uncertain problems in the information industry and data mining paradigm [16]. New data processing systems make the computing grid work by managing and pushing the data out to individual nodes, sending instructions to the networked servers to work in parallel, collecting individual results and then reassembling them to produce meaningful results. Processing the data where it resides is faster and more efficient then before analyzing transporting it to a centralized system. R. Buse, (2012) noted It will be in the form of computerized information that does not have a data model and hence are not used by data mining [15]. [14] LaValle S, (2011) proposed the task of managing semi-structured data signifies possibly the major data management opportunity which subsequently results in managing relational data. Semi-structured data constitutes about 70% of the data collected or stored in larger organizations which are difficult to access, use or retrieve. Most existing tools generally deal with a single text corpus, or individually handle different corpus. [13] discovered these tools may not give a full picture of ongoing events on social media. Topic Panorama was recently proposed to allow researchers to simultaneously analyze and correlate the topics of different corpora simultaneously. Barbierato E, (2014) noted Topic Panorama is highly interactive and assists users in interacting with matched topic graphs at different granularity levels. However, Topic Panorama only handles small-scale graphs and simultaneously visualizes several corpora.

Millard (2013) proposed Big Data allow researchers to decode human DNA in minutes, predict where terrorists plan to attack, determine which gene is mostly likely to be responsible for certain diseases and, of course, which ads you are most likely to respond to on Facebook. The business cases for leveraging Big Data are compelling. N. Diakopoulos, (2010) discovered The instance, Netflix mined its subscriber data to put the essential ingredients together for its recent hit House of Cards, and subscriber data also prompted the company to bring Arrested Development back from the dead. T. Menzies, (2013) discovered Hashing technique was applied to evaluate the analytics process. This method satisfies volume and velocity, whereby better performance was observed for 7285documents. For evaluation, these documents were split into 72% of training data 28% test data, respectively. The actual sizes of these two datasets were 7285 and 18,846. This method is feasible for different sorts of data like optical characters, speech audio, and document scripts. Sparse hashing technique is not suitable for velocity, since it considers fixed data size. Moreover, it does not provide anything related to data accuracy [9].

2.1. Problem Identification

Most of the existing system provides only analyzing the data which is unstructured and there is no query complexity. The unstructured/semi structured data are being continuously comes from various sources like satellite images, sensor readings, email messages, social media, web logs, survey results, audio, videos etc. Hence our proposed open grid rule generating algorithm provides the data into absolute filtered data. Before filtering, the data are mapped and reduced to a level, which is more efficient in giving approximate required data [10].

3. Proposed methodology

The data analysis concept of Big Data gives analytical methods which can be applied to analyze traditional datasets which includes analytical architecture and software requirement for exploration of big data [11]. Open grid rule generation is one of the most essential stages of the big data value chain where the main objective is to extract the meaningful information and providing suggestions and decisions. Different types of possible and gravitational values can be produced through the several stages of analysis in different fields [12].

3.1. Open grid rule generation algorithm

Big data analytics is typically performed using specialized software tools and applications for predictive analytics, data mining, text mining, and forecasting and data optimization. Collectively these processes are handled with open grid rule generating which shows highly integrated functions of high-performance analytics. This enables an organization to process extremely large volumes of data that a business has collected to determine which data is relevant and can be analyzed to drive better business decisions.

a) Creating the data set with a new template for importing the data from the database which contains both the structured and semi-structured data. When the data are imported it shows in the raw form, what is to be used, within the new data set template .datx (data analytic tool extension).

b) The data are set into its framework after importing for the purpose of mapping, there is a raw data loaded in the framework. Hence, it is taken for the purpose of reducing the attributes.

c) The raw data are reduced for the process of mapping that is assigning the attributes list into the absolute form. Since the mapped data are processed in the database.

d) Data filtered based on threshold value i.e support and confidence level to reduce the redundancy and duplication of data. The history of data and log files can be retrieved.

A flexible and reconfigured grid along with the big data pre processing enhancement, mapping and reducing, data filtering, data-parallelization schemes can be more effective approaches for extracting more meaningful knowledge from the given data sets.

The following figure 2 shows the process of data analytics with the open grid rule generation Since the process is more effective, it shows the data to a higher level with high efficiency [17]. All the raw data are cleared and can be retrieved whenever the user needs.

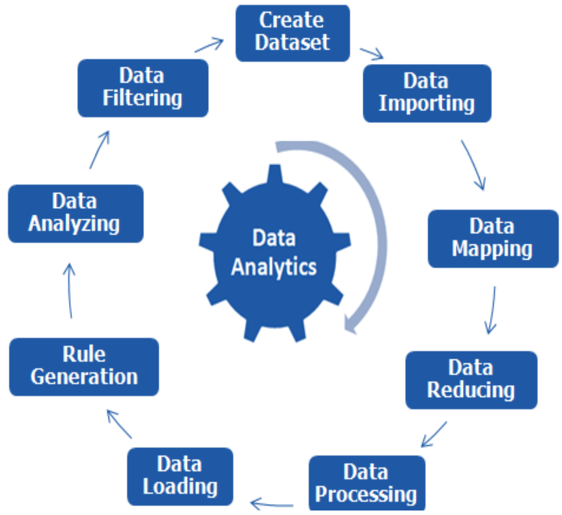


Fig. 2: Analytics Process in Framework.

3.2. Algorithm

```
Arraylist Lk ← New Arraylist
Regex R ← New Regex
For I= 0 → L.Count
String [] Subl1 ← R.Split(L[I].Tostring.
For J= I+1 → L.Count
String [] Subl2 ← R.Split (L [J].Tostring.
// Comparing Two Items
String Temp → L [J].Tostring
//Store The Two Key Sets.
For M=0 → Sub L1. Length
Boolean Subl1mlnsubl2 = False
For N=0 → Sub L2.Length
If (Subl1 [M] ← Subl2 [N])
Subl2mln Subl2= True
If Subl1mlnsubl2 == False
Temp=Temp +”,”+Subl1 [M]
String [] Subtemp → R.Split(Temp)
If (Subtemp.Length ← Subl1.Length + 1
Bools Is Exists = False
For M=0 → Lk.Count
Bool Is Contained = True
For N =0 → Subtemp.Length
If (! Lk [M].Tostring ().Contains (Subtemp [N])) Is Contained = False.
If (Iscontained ==True) Is Exists=True
If (Exists == False)
Lk Add (Temp)
Return Lk.
```

3.3. Efficiency calculation

Preliminaries

m_i – initial time.

m_d – destination time

a_t – average time

p_t – approximate time

m_t – mean time

OGRG Algorithm process execution

```
Mean time $m_t = m_i + m_d/a_t$
$a_t \leftarrow$ avg time
$a_t \leftarrow m_t [m_i + m_d]$
// avg time for the attribute
$p_t \leftarrow a_t [ m_t/2]$
$[m_t \rightarrow m_i + m_d/a_t]$
$a_t [m_i + m_d/a_t]/2$
$p_t \rightarrow a_t [m_i + m_d]$
// approximate value for those crime data
$P_t$ is unequal to $m_t$
$m_t \leftarrow m_i + m_d/a_t$
// here, the mean time value is identified.
$P_t \leftarrow a_t [m_t/2]$
$P_t$ not equal to $m_t$
// but the approximate value is similar with attribute mean time.
```

Every attribute has the specific mean time as the user specifies. The initial time for each attribute may vary due to its possibility, such in case the mean time for attribute changes, but it has only some difference with the average time provided for those attributes. In such a case, the approximate time for each attribute is valued by their mean time and its initial time [18].

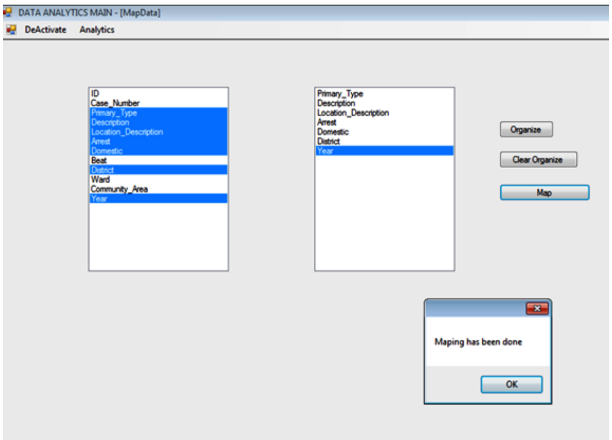
The rough estimation of the investigation procedure all through the essential sort and its weight. By the Process, the data put in the database and the crude data are cleared from the data compartment and it demonstrates the overseed information delivered for diminishing the rough data from the unstructured shape and makes it into the sorted out data with the capable course of action with the way toward sifting [19,10,12,22]. Data analytics is a technology-enabled strategy for gaining richer, deeper, and more accurate insights into customers, partners and the business and ultimately gaining competitive advantage. In the present procedure, the data are inspected through a such essential arrangement and the separated data are needy upon 55%.Here the examination over through 95% induced with high compelling and sifted information are more estimated.

4. Results and discussions

Fig.3. states the information examination throughout the year and its primary type of the crime dataset. By the open grid rule generation process, the information in the database are broken down up to the superior level. The crude information are cleared from the information compartment and the manage is produced for decreasing the crude information from the unstructured shape and makes it into the organized information with the proficient arrangement with the rule of mapping. In the current strategy, the information is examined through a such basic configuration and the broke down information are dependent upon 60%. Here the examination over through 95% surmised with high efficiency.

4.1. Mapping data

The mapping is the process of assigning the data which is reduced in a level which can be retrieved from it whenever the user needs. The following figure shows the process of mapping the dataset.



ID
Case_Number
Primary_Type
Description
Location_Description
Arrest
Domestic
Beat
District
Ward
Community_Area
Year

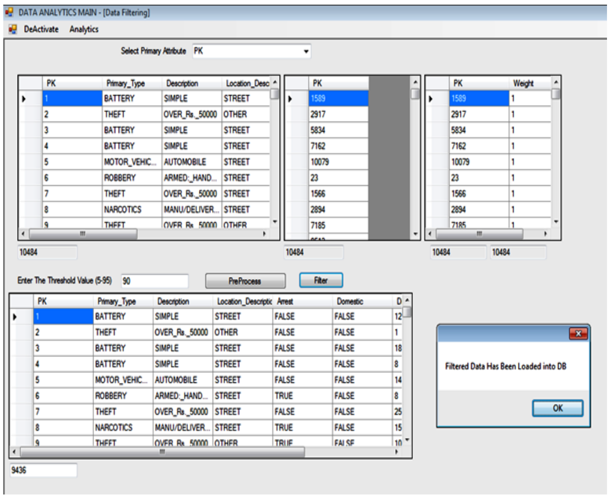
Primary_Type
Description
Location_Description
Arrest
Domestic
District
Year

Maping has been done

Fig. 3: Data Mapping.

4.2. Data Filtering

The result Fig 4 shows that the filtered data which is more absolute. This process is handled by clearing the raw data from the data base.



PK	Primary_Type	Description	Location_Desc
1	BATTERY	SIMPLE	STREET
2	THEFT	OVER_Rs._50000	OTHER
3	BATTERY	SIMPLE	STREET
4	BATTERY	SIMPLE	STREET
5	MOTOR_VEHIC...	AUTOMOBILE	STREET
6	ROBBERY	ARMED:_HAND...	STREET
7	THEFT	OVER_Rs._50000	STREET
8	NARCOTICS	MANU/DELIVER...	STREET
9	THEFT	OVER_Rs._50000	OTHER

10484

PK
1589
2917
5834
7162
10079
23
1566
2894
7185

10484

PK	Weight
1589	1
2917	1
5834	1
7162	1
10079	1
23	1
1566	1
2894	1
7185	1

10484 10484

Enter The Threshold Value (5-95) 90

PK	Primary_Type	Description	Location_Descriptic	Arrest	Domestic	D
1	BATTERY	SIMPLE	STREET	FALSE	FALSE	12
2	THEFT	OVER_Rs._50000	OTHER	FALSE	FALSE	1
3	BATTERY	SIMPLE	STREET	FALSE	FALSE	18
4	BATTERY	SIMPLE	STREET	FALSE	FALSE	8
5	MOTOR_VEHIC...	AUTOMOBILE	STREET	FALSE	FALSE	14
6	ROBBERY	ARMED:_HAND...	STREET	TRUE	FALSE	8
7	THEFT	OVER_Rs._50000	STREET	FALSE	FALSE	25
8	NARCOTICS	MANU/DELIVER...	STREET	TRUE	FALSE	15
9	THEFT	OVER_Rs._50000	OTHER	TRUE	FALSE	10

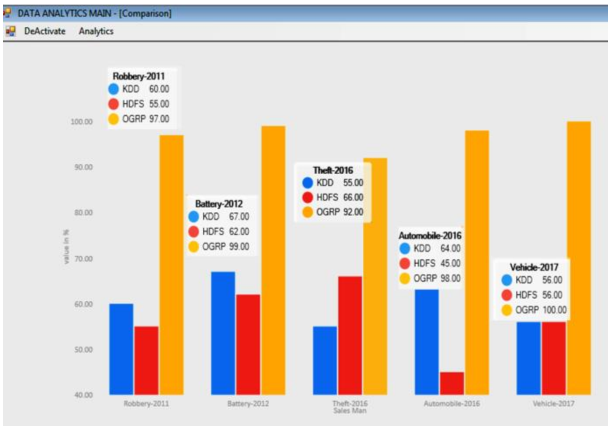
9436

Filtered Data Has Been Loaded into DB

Fig. 4: Data Filtering.

Table 1: Legends KDD Process: Knowledge Discovery Databases, HDFS Process: Hadoop Distributed File System, Open Grid Rule Generation

S.No	Primary Type	Year	KDD Process	HDFS Process	OGRG Process
1	Robbery	2011	60%	55%	97%
2	Battery Theft	2012	67%	62%	99%
3	Assault	2016	55%	66%	92%
4	Motor Vehicle Theft	2016	64%	45%	98%
5	Burglary	2017	56%	56%	100%

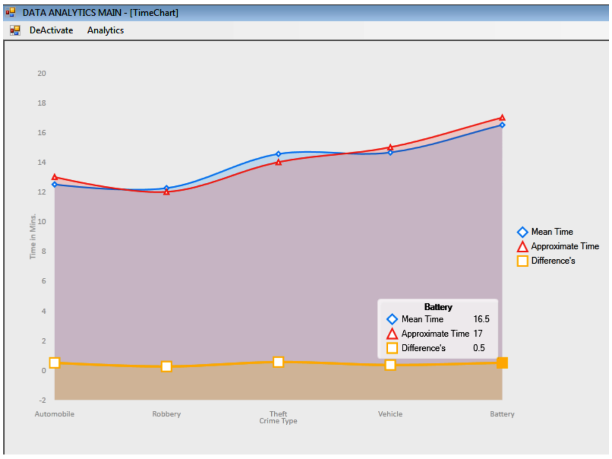


Graph. 1: Efficiency Comparison.

The above graph compares the produced open grid rule generation algorithm with Data Mining KDD process and Hadoop HDFS algorithm. It is clear that the proposed system has produced high efficiency information with the analytics process; there is a huge difference between the existing and the proposed algorithm.

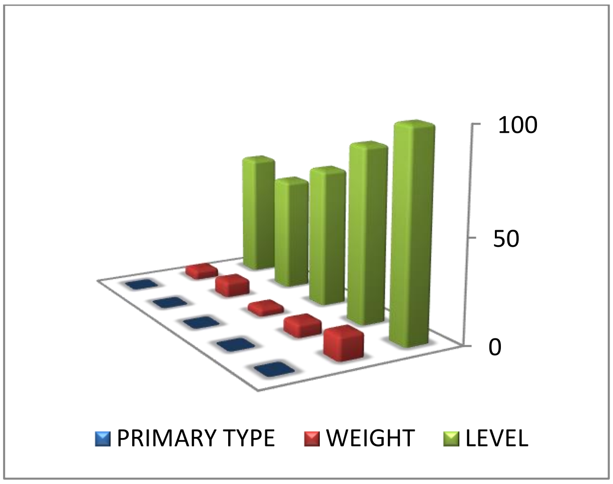
Table 2: The Time Comparison of the Attributes by OGRG

S. No	Attribute	Mean time	Approx value	Difference
1	Robbery	12.5	13	0.1
2	Battery Theft	12.25	12	0.4
3	Assault	14.55	14	0.6
4	Motor Vehicle Theft	14.65	15	0.5
5	Burglary	16.5	17	0.3



Graph. 2: Efficiency Comparison.

The above graph states the time estimation of crime detection by open grid rule generation by showing the mean value for the attributes and our process shows the better approximate value.



Graph. 3: Attribute Comparison Chart.

The above graph estimates the analyzing process of the attributes by their weightage. Data analytics is a technology-enabled strategy for gaining richer, deeper, and more accurate insights about customers, partners and business environment. This process is handled by clearing the raw data from the data base. When a data mapping process is handled it is reviewed that data has been reduced and there is no data redundancy

4.3. Comparison between KDD, HDFS with OGRG

The KDD process is efficient for the user, but it has some issues due to the undefined approximate value, this process may change on behalf of these attribute list. The big data analytics would be executed in distributed processing across several servers (“nodes”), utilizing the paradigm of parallel computing and ‘divide and process’ approach. Likewise, models and techniques—such as data mining and statistical approaches, algorithms, visualization techniques—need to take into account the characteristics of big data analytics.

The HDFS process is also similar to previous process, but has some changes that this is a file system; it has the database storage, since the data retrieving has some difficulties due to its time recruitment. It does not show the mean time taken and have not shown the approximate deviation for the attribute which the user creates of their own. Since our proposed system has shown the better efficiency on the attributes. It is clear that our proposed system has produced high efficiency information with the analytics process; there is an huge difference between the existing and our proposed algorithm.

5. Conclusion

Big data analytics has the potential to transform the way which shows sophisticated technologies to gain insight from their clinical and other data repositories and make informed decisions. The analytics system was handled before by using many principles, but it does not give the clearance, Open Grid Rule Generation algorithm gives. It is proved that this proposed system has produce more efficient result than comparing KDD and HDFS process. The data filtered is with high efficiency and there is no redundancy. The raw data can be also retrieved whenever the user needs, the data is more efficient up to 95% of approximate.

References

[1] Maluf, A. David, Bell, G. David, Knight, Chris,Tran, Peter, La, Tracy, Lin, Jenessa, McDermott,Bill, Pell, Barney,"NASA-XDB-IPG: ExtensibleDatabase - Information Grid" Global Grid Forum8, 2003.

[2] S. Kaisler, F. Armour, and J. A. Espinosa, "Introduction to big data:Challenges, opportunities, and realities minitrack," in 2014 47th Hawaii International Conference on, pp. 728-728, 2014.

[3] Motoi Iwashita, Ken Nishimatsu, Shinsuke Shimogawa. Semantic Analysis Method for Unstructured Data in Telecom Services.IEEE13. IEEE 13th International Conference on Data Mining Workshops; p.789-795, 2008.

[4] Ohlhorst F: Big Data Analytics: Turning Big Data into Big Money. USA: JohnWiley & Sons; 2012.

[5] LaValle S, Lesser E, Shockley R, Hopkins MS, Kruschwitz N: Big data,analytics and the path from insights to value. MIT Sloan Manag Rev,52:20–32, 2011.

[6] Gärtner M, Rauber A, Berger H, Bridging structured and unstructured data via hybrid semantic search and interactive ontology-enhanced query formulation. Knowl Inf Syst 1–32, 2013.

[7] Barbierato E, Gribaudo M, Iacono M, Performance evaluation of NoSQL big-data applications using multi-formalism models. Future Gener Comput Syst37:345–353, 2014.

[8] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, “Diamonds in the rough: Social media visual analytics for journalistic inquiry,” in Proceedings of IEEE Conference on Visual Analytics Science and Technology, pp. 115–122, 2010.

[9] Rajesh, M., and J. M. Gnanasekar. "Path Observation Based Physical Routing Protocol for Wireless Ad Hoc Networks." Wireless Personal Communications 97, no. 1: 1267-1289. (2017)

[10] Chris Clifton, Murat kartarcioglu, Jaideep vaidya,Xiaodong Lin and Michael Y.Zhu, Tools for Privacy Preserving Distributed Data mining, SIGKDD Eplor. Vol 4, Issue 2, 2002.

[11] Mansuri I.R. Sarawagi S. “Integrating Unstructured Data into Relational Databases” Data Engineering. ICDE '06. Proceedings of the 22nd International Conference, IIT Bombay 2006

[12] Dafang Zhuang , Wen Yuan, Jiyuan Liu, Dongsheng Qiu, Tao Ming, „The Unstructured Data Sharing System for Natural Resources and Environment Science Data of the Chinese Academy of Science‟ in Data Science Journal, Volume 6, Supplement, 20 October 2007.

[13] T. Menzies, "Beyond data mining; towards idea engineering," inProceedings of the 9th International Conference on Predictive Models in Software Engineering, 2013, pp. 1-6.

[14] hu, h., wen, y., chua, t-s., li, "toward scalable systems for big data analytics: a technology tutorial," access, ieee , vol.2, no., pp.652-687, 2014.

[15] Chun-Wei Tsai,Chin-Feng Lai,Han-Chieh Chao, Athanasios V. Vasilakos, Big data analytics: a survey, ‚Journal of Big data, Springer, December 2015

[16] V. Srilakshmi, V.Lakshmi Chetana , T.P.Ann Thabitha ,A Study on Big Data Technologies, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 6, June 2016

[17] D. Assunçãoa, Rodrigo N. Calheiros b ,Big Data computing and clouds: Trends and future directions Marcos, Elseiver, 27 August 2014

[18] Samiya Khan, Kashish AraShakil, MansafAlam, “Cloud Based Big Data Analytics: A Survey of Current Research and Future Directions”, Journal of Contemporary Psychotherapy, 2015.

[19] Chang RM, Kauffman RJ, Kwon Y (2014) Understanding the paradigm shift to computational social science in the presence of big data. Decis Support Syst 63:67–80.

[20] Agrawal, D., Das, S., El Abbadi, A.: Big data and cloud computing: current state and future opportunities. In: Proceedings of the 14th International Conference on Extending Database Technology (EDBT/ICDT’11), pp. 530–533 (2011)

[21] Chen, C.L.P., Zhang, C. Y.: Data-intensive applications, challenges, techniques andtechnologies: a survey on big data. Inf. Sci. 275, 314–347 (2014)

[22] Zikopoulos PC, DeRoos D, Parasuraman K, Deutsch T, Corrigan D, Giles J: Harness the Power of Big Data. McGraw-Hill: The IBM Big Data Platform; 2013.