Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 947 - 952
Received 20 January 2021; Accepted 08 February 2021.

ETL - A Solution to Challenging Issues in Big Data Analytics

Gendlal M Vaidya
Research Scholar,
Department of Information Technology, Yeshvantrao Chawan College of Engineering.
Nagpur, India
gendlalvaidya@gmail.com

Dr. Manali M. Kshirsagar
Principal,
Rajiv Gandhi College of Engineeringg & Research, Nagpur, India
manali_kshirsagar@yahoo.com

ABSTRACT

Now a days, tremendous data generation is going on across the globe. Most of the social networking sites are rich source of data. Other ecommerce sites, government as well as private sectors, healthcare organizations, cloud networks, different servers etc are contributing towards collections of huge volume of data. The collected data may be in structured or unstructured formats. Various formats of data are also the adding features in it. While processing, the complete process, it is pipelined in such a way that, obtained final resultant data can give some important and meaningful conclusions to work on it. Some analytical results are again useful for decision making. But in many cases, the obtained results may vary in some values, graphs and figures. This happens due to use of some unrealistic tools available for the processing of big data. In this paper, we suggested some proper ETL tools and solutions for the processing of big data, which may results into getting suitable analytics and conclusions from the data.

Keywords— Big data, ETL, Hadoop, ETL tools, Talend

I. INTRODUCTION

The term big data is obviously huge in quantity of data. It may contain various formats within it. In today's life, every software firms, IT companies, public and private sectors are collecting lots of data from their available sources. Once the data is stored into storage devices then they may face several challenges like data management, data security and many other data processing related issues. Once they started with pure and clean data, then further complications can be avoided. But think what, if the proper neat data is not processed then, the generating results can not conclude to proper further decisions and therefore the ETL (extraction, transform and load) method plays a very important role in the data processing technique. Data warehousing is a platform where the data is integrated. The starting ETL process is not as easy and low priced. It all depends on the design of the data warehouse. Firstly the data extracted from various sources and after this, cleaned and standardized data stored in data warehouse. Business intelligence tool are used for the decision making purposes[1]. Figure 1 shows the cloud based data which is managed by several storage management techniques where virtual machines can work very stringly to process the heavy data sets available on cloud and its clusters.

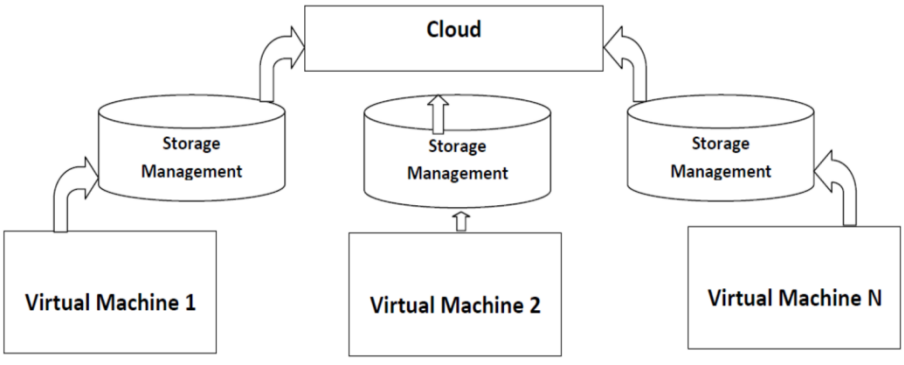


Figure 1: Cloud Storage Management

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 947 - 952
Received 20 January 2021; Accepted 08 February 2021.

Information stored in a batch format like fusion DB are handled by typical ETL and it needs to be processed individually to handle spatial information stored in a stream form. Most of the typical spatial information available at social networking services like facebook, twitter are actively beneficial to all the researchers and scientists. Such type of spatial information is increasing in day to day life and its being utilized by several applications and software industries for analysis and analytics purposes. A powerful ETL is having very close relevance to the quality of the spatial data. In short, the spatial information having more meaningful information its quality grows up automatically. If researchers are not confident about the inpt of the spatial data then the resultant output will be some wards unreliable and meaningless, and thus resulting into wastage of efforts, cost and time. While analyzing the spatial information of stream type then the ETL must be that much powerful and capable to process in stream form. This study has designed and implemented ETL engine by adding some improved functionalities of extraction, transform and load[8]. Several existing designs are compared and proposed the design for ETL engine are some added features o this study.[2]

Day by day number of smart cities are growing across the globe. UK government undertaking various smart city projects, announced the London Plan and situated Smart London Board. Similarly, countries like US, China and France are continually trying to improve the public facilities like traffic management and issues related to environment to increase the count of the smart cities. The existing software that can handle such spatial data remains very demanding and it may generate more revenue to develop the nation. This study focused on the processing libraries of spatial information using Apache NiFi platform. It is easily accessible and provedes interface on graph based[3].

This study proposed a big data architecture that combines together variety of database models based on the bank's status and related data or information. They build real-time and non-real time data, structured, poly-structured and unstructured data and the processing stems with transactional databases, Massive Parallel Processing databases, frameworks of Hadoop, Spark and Storm. The entire ETL process was based on the enhanced performance features of Massive Parallel Processing system. Their study finally validates the performance of Oracle and Massive Parallel Processing databases on one of the factor of 5V's model i.e velocity and compression with five data operating modes. They also verified the effectiveness of the entire database by using Massive Parallel Processing system[4].

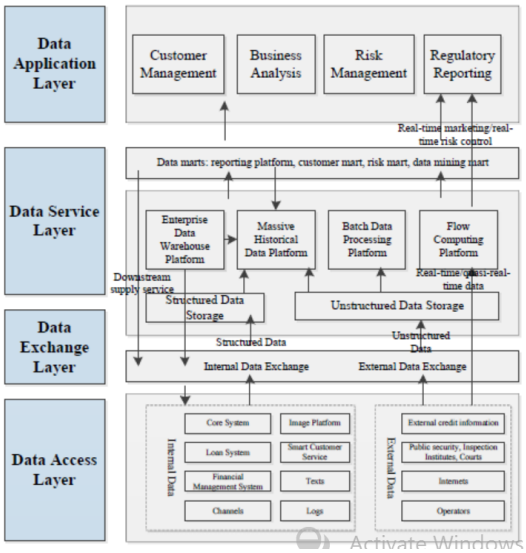


Figure 2: Data Architecture in Banking Systems[4]

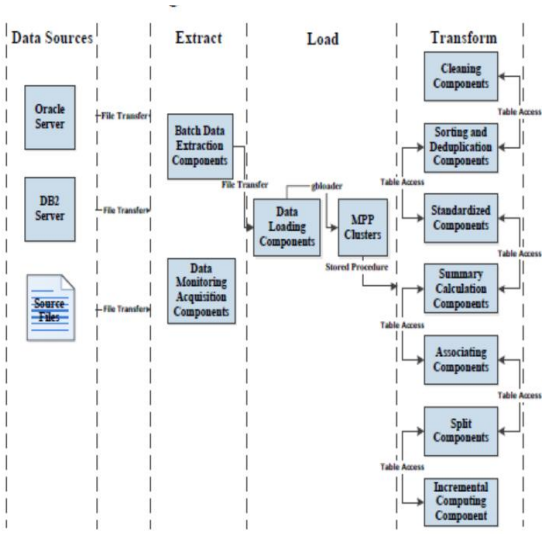


Figure 3 : Process of ETL[4]

According to [5], various national and international universities are in demands due t o the new courses based on the data science or machine learning. The students interests are getting increasing day by day towards theses courses and hence there is increase in the quantity of students who are enrolling to existing as well as new programs in science and technology areas. More the candidates, more will be the information and hence creates multiple data sources for various purpose of data analytics. The study is focused on the academic data analysis. Authors study is related to various related database that are working on t he academic data. They compared and studied various database and proposed their own architecture for the processing of the academic information. They analyzed students assessment data by courses, computing scores and etc. Their work contributed more for the decision making in the academic institutions and the related universities. In data science domain, data is categorized into 5V's Model which are volume, velocity, variety, veracity and value.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 947 - 952
Received 20 January 2021; Accepted 08 February 2021.

Table 1: Five V's Of Big Data

Category	Description
Volume	It indicates the size of thedata which is available in huge quantity. From social media and other e-commerce site several Exabytes of data collected on daily basis.
Velocity	It indicates the speed of the data generation. As the years are passing the velocity of data is also getting increasing.
Variety	It tends to different formats of data. Many types of formats like .csv, .doc, .jpeg, .xls, .doc, .png etc are the available formats. These all contributes towards variety of data.
Veracity	It indicates accuracy and reliability of data. More correct and reliable data generates useful information.
Value	Value derived from processing of data remains very useful for the decision making and other work.

All above studies indicates shows very close relations to all the 5V's model of big data. Each study focused new challenges in ETL techniques. Some of them designed some framework for the processing of big data. In frameworks, once the data entered it remains capable to process whole data and generates useful information. Some study executes ETL on some open source tools available for preprocessing. Such studies or work have to do all the necessary operations on several platforms and hence consumes more time. To reduce the time consumption on each platforms, it requires all the useful platforms should be that much reliable to do the allotted task within a short spam of time.

II. METHODOLOGY

Eftim Zdravevski et al[6], proposed an architecture which is based on cloud and used for the efficient ETL process of Big Data. Two phases of ETL i.e extraction and transformation performed by Spark framework and finally used the distributed load agents for the loading purpose[9]. The proposed ETL used the processing resources of the cluster slaves, instead of using the database server. Amazon AWS provides cluster of Hadoop with short time intervals. Authors proposed algorithm based on cluster size optimization, ETL process completes within the required time found to be cost effective. Their proposed approach helps to reduce the rows significantly. The work was carried out on system generated log files. Such process helps to minimize the development time, data complexity and corresponding tools.
Extraction, transform and load (ETL) is a process in which raw data can be turned into some valuable and important information which can be used for business intelligent. Major function of any ETL tools are as:
(E) Extraction of data from various data sources
(T) Data transformation in order toget the suitable data model like data warehouses
(L) Data loading into target database or any destination

A. Options for ETL tools

ETL tools must be flexible and quick enough to match with all the basic requirements of the important data. Most of the organizations and industries prefers such ETL tools[10]. There are many options for ETL tools. Mostly chosen by considering the format of data, sources of data and complexity factors available in the data. Some of t he options for ETL tools are as;

- Open Source tool: These tools are more widely used by scientists and researchers. These typed tools are capable to handle data of several formats.
- Cloud based tools: These tools works on unstructured or various formats of data and makes data readily available and are more flexible. Because of its flexibility nature, it is widely used.
- Legacy or Incumbent tool: These tools basically works on structured data and generally provides core data; Legacy tools are less flexible in nature[7].

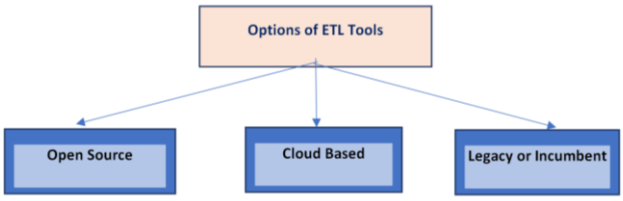


Figure 4: Options for ETL Tools

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 947 - 952
Received 20 January 2021; Accepted 08 February 2021.

Most commonly used ETL tools are Xplenty, AWS Glue, , Talend, Stitch, Informatica PowerCenter and Oracle Data Integrator. We are using the Talend, one of the open source data integrator tool which is moreefficient to gnerate the desired output. Basically, it provides an option to make the design of the available information (source file) which then processed (by applying filters) and produces the reduced information (destination file) which can be used for further analytics purpose and hence useful for decision making.

Figure 5: Data Integrator toll (Talend)

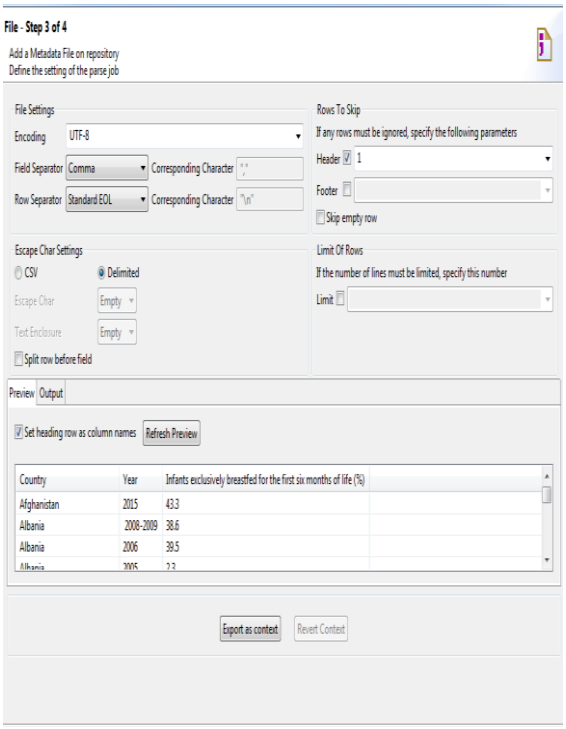


Figure 6: Preprocessing in Talend-Part-1

Figure 5 showns the GUI of Talend ETL tool. It shows the basic information about the tool and Figure 6 shows the preprocessing of the tool. Jere we have to set the necessary requirements from the tool.

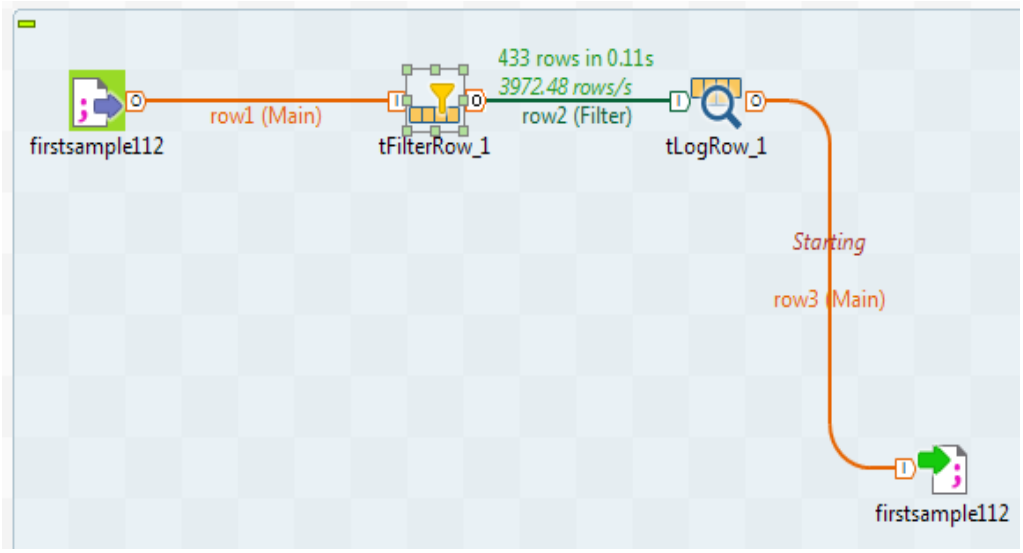
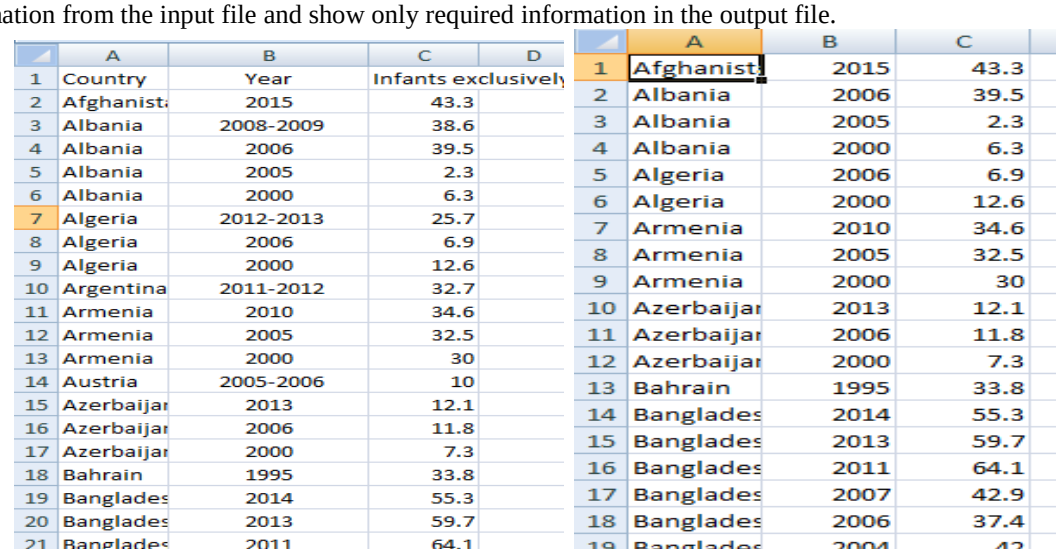


Figure 7: Applying Filters to input data part-2

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 947 - 952
Received 20 January 2021; Accepted 08 February 2021.

Figure 7 shows the filters which are applied to the input file. Here, the filters, removes the unwanted information from the input file and show only required information in the output file.



	A	B	C	D
1	Country	Year	Infants exclusively	
2	Afghanist	2015	43.3	
3	Albania	2008-2009	38.6	
4	Albania	2006	39.5	
5	Albania	2005	2.3	
6	Albania	2000	6.3	
7	Algeria	2012-2013	25.7	
8	Algeria	2006	6.9	
9	Algeria	2000	12.6	
10	Argentina	2011-2012	32.7	
11	Armenia	2010	34.6	
12	Armenia	2005	32.5	
13	Armenia	2000	30	
14	Austria	2005-2006	10	
15	Azerbaijar	2013	12.1	
16	Azerbaijar	2006	11.8	
17	Azerbaijar	2000	7.3	
18	Bahrain	1995	33.8	
19	Banglades	2014	55.3	
20	Banglades	2013	59.7	
21	Banglades	2011	64.1	

	A	B	C
1	Afghanist	2015	43.3
2	Albania	2006	39.5
3	Albania	2005	2.3
4	Albania	2000	6.3
5	Algeria	2006	6.9
6	Algeria	2000	12.6
7	Armenia	2010	34.6
8	Armenia	2005	32.5
9	Armenia	2000	30
10	Azerbaijar	2013	12.1
11	Azerbaijar	2006	11.8
12	Azerbaijar	2000	7.3
13	Bahrain	1995	33.8
14	Banglades	2014	55.3
15	Banglades	2013	59.7
16	Banglades	2011	64.1
17	Banglades	2007	42.9
18	Banglades	2006	37.4
19	Banglades	2004	42

Figure 8: Source and destination files

Figure 8 shows the input file, contains the information, where some rows does not give the specific ideas about the infants exclusively whereas the output file is having more clear and precise information. This is what we can get from the ETL tools exactly.

III. RESULTS AND DISCUSSION

(Experiment 1)

Table 2: Experimentation Summary on WHO datasets

Sr No	Category	File Credential	Size in KB	Reduction in %
1	Input File	WHOSIS_0000066.csv	12 KB (Original)	100 %
2	Output File	WH1.csv	9 KB (Reduced)	25 %
3	Output File	WH3.csv	1 KB (Reduced)	88.89 %

Experiment performed on sample dataset provided by World Health Organizations (WHO) where original size of an input file was 12 KB. After applying ETL technique and data integrator tool the original size reduced upto 25% whereas subsequent reduction arises to 11.11% of t he original size.

(Experiment 2)

Table 3: Experimentation Summary on Consumer Complaints datasets

Sr No	Category	File Credential	Size in KB	Reduction in %
1	Input File	Consumer Complaints.csv	650068 KB (Original)	100 %
2	Output File-1	abc.csv	442244 KB (Reduced)	21.97 %
3	Output File-2	Abc2.csv	70089 KB (Reduced)	94.16 %

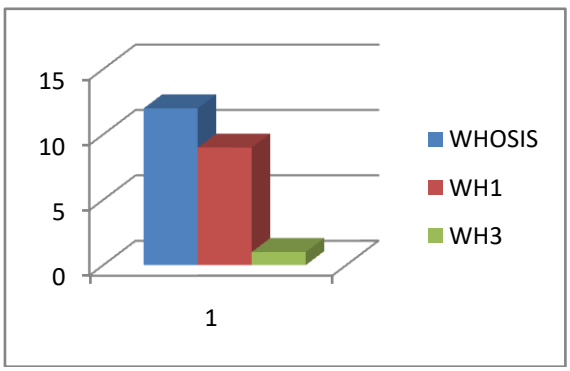


Figure 9: Analysis of Experiment 1

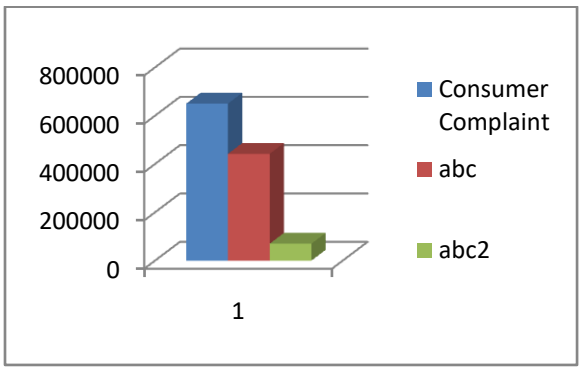


Figure 10: Analysis of Experiment 2

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 947 - 952
Received 20 January 2021; Accepted 08 February 2021.

IV. CONCLUSION

In this paper we discussed some ETL related work, different tools and their working strategies used to overcome some issues. We performed the ETL operation on Talend, which is an open source tool, used for the reduction in file size which results to get more useful information for decision making purpose. The reduced file size contribute to avoid wastage of memory and time as well. Hence, it is the most cost saving and effective technique which researchers can apply while working on huge datasets. Experiment results (Table 1 and Table 2) shows the minimized file size which occurred due to actual filtering the rows. This approach also reduces the data complexity and beneficial for the technologies which are depended upon the resultant data. The further studies will be based on this concept and further tools like Hadoop, MapReduce and related will be used. For the better working of such tools, suitable algorithm will be designed with pruning feature and further it will be implemented.

REFERENCES

[1] Papa Senghane Diouf and et al, “Variety of data in the ETL processes in the cloud:state of the art”, published in IEEE International Conference on Innovative Research and Development (ICIRD), dated 11-12 May 2018,Bangkok Thailand, page no 1 to 5
[2] Sang-su Kim and Kwaun-Sik Song, " Implementation of a Distributed Processing Engine for Spatial Big-Data Processing Based on Batch and Stream"pubished in IEEE xplore 2017, 1196-1198
[3] Sang-Su Kim, Wang-Ro Lee, Jun-Hui Go " A Study on Utilization of Spatial Information in Heterogeneous System Based on Apache NiFi" published in 2019 IEEE xplore page 1117 - 1119
[4] Shenglan Ma, Hong Xiao, Botong Xu, Ran Tao, Fangkai Xie, Daicai Zeng, Tongsen Wang, " Bank Big Data Architecture based on Massive Parallel Processing Database" published in 12018 15th International Symposium on Pervasive Systems, Algorithms and Networks, 93-99
[5] Gant Gaw Wutt Mhon and Nang Saing Moon Kham "ETL Preprocessing with Multiple Data Sources for Academic Data Analysis" published in 2020 IEEE Conference on Computer Applications(ICCA), page No 1-5
[6] Eftim Zdravevski, Petre Lameski, Ace Dimitrievski, Marek Grzegorowski, Cas Apanowicz, " Cluster-size optimization within a cloud-based ETL framework for Big Data" published in 2019 IEEE International Conference on Big Data (Big Data), 3754 - 3763
[7] https://www.talend.com/resources/etl-tools/
[8] Hebert Silva, Tania Basso, Regina Moraes, Donatello Elia, Sandro Fiore, "A Re-identification Risk-based Anonymization Framework for Data Analytics Platforms", published in 2018 14th European Dependable Computing Conference, Page No 101-107
[9] S.N.Mohd Isa, S.A.Abdul Shukor,N.A.Rahim, I.Maarof, Z.R.Yahya, A.Zakaria1, A.H. Abdullah1, R.Wong4, "A Review of Data Structure and Filtering in Handling 3D Big Point Cloud Data for Building Preservation", published in 2018 IEEE Conference on Systems, Process and Control (ICSPC 2018), 14–15 December 2018, Melaka, Malaysia, Page No 141-147
[10] Dimitrios Sisiaridis, Olivier Markowitch, "Reducing Data Complexity in Feature Extraction and Feature Selection for Big Data Security Analytics", published in 2018 1st International Conference on Data Intelligence and Security, Page No 43-48