International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Extracting and Transforming Heterogeneous Data from XML files for Big Data

Tanuja Das, Ramesh Saha, Goutam Saha

Abstract: ***Digital technology is fast changing in the recent years and with this change, the number of data systems, sources, and formats has also increased exponentially. So the process of extracting data from these multiple source systems and transforming it to suit for various analytics processes is gaining importance at an alarming rate. In order to handle Big Data, the process of transformation is quite challenging, as data generation is a continuous process. In this paper, we extract data from various heterogeneous sources from the web and try to transform it into a form which is vastly used in data warehousing so that it caters to the analytical needs of the machine learning community.***

Keywords**: *Big data, data transformation, data warehousing, ETL.***

I. INTRODUCTION

Big data is data that comes from various heterogeneous sources and is generated in such a high pace that the traditional database systems fails to handle it. The data generated is voluminous, has great momentum and does not conform to our traditional database designs [1]. Characteristics of big data involve Volume, Variety, and Velocity, i.e. the three V's [2]. Gartner [2] defined Big Data as "Big data is high volume-velocity-variety repositories of knowledge which necessitates worthwhile and progressive framework of mining knowledge in order to attain improved perception and making better decisions." In order to tackle big data to generate knowledge, the data must be transformed into a form that is suitable for the analytics process.

Due to generation of huge amounts of data from multiple sources [3] which includes the Internet, social media, digital sensors, etc., it becomes imperative to integrate these data in order to fit into a data warehouse. Illustrating the case of Web pages which are basically strings of texts, data extraction or scraping the page can be performed by searching for presence of certain keywords. The scraping of the web pages has tremendous significance to Big Data [4]. Enhancing the data warehouses with current data from the web is a gaining importance with the increase in the number of internet users. Maintaining the quality of data in the data warehouse is crucial as most of the outcomes of research are dependent on the data stored in the data ware-house.

To ensure effective usage of data, the Resource Description Framework (RDF) [10] was developed by the World Wide Web Consortium (W3C) which accumulates data from various sources in numerous formats and transforms it to a common format. The essence of the idea behind this framework was to enable interchange of data even if their fundamental schemas are not the same. In order to meet the Big Data needs, Malik et. al [11] tried to transform the data in such a manner that the information loss is minimized. The core idea of this process is to maintain data and metadata such the relationship between them remains strong without increasing the complexity.

The "pay-as-you-go" approach [12] handles heterogeneous data by building a metrics-driven environment instead of the schema-first feature of conventional data integration approach. The approach provides a number of fundamental services in order to enhance the semantic relationships among the data in a repetitive manner [13]. The technique is quite a step-up over the prevailing data management systems as it takes into account the various constituents of a dataspace management system and utilize these measures to identify the various contexts, priorities and techniques of the data spaces [14][15]. However as there is much apprehension for these state-of-the-art systems by the user community, a number of challenges are still to be addressed.

The ETL process (Extract-Transform-Load) [6] provides a framework for integrating data from various sources and storing them into a data warehouse for further analysis. But due to the changing and complex nature of the generated data in the present scenario, the traditional ETL process becomes inefficient. To cope up with this, a large number of data warehousing projects like the Oracle Warehouse Builder (OWB) [7], IBM InfoSphere DataStage [8], Microsoft SQL Server Integration Services [9], etc. redesigned the traditional ETL framework for modeling the data ware-house. Also to tackle the 3 Vs of big data (volume, variety and velocity) the tradition-al ETL process has been urged to switch to ELT (Extract, Load and Transform) on Hadoop [16]. ELT on Hadoop grants quite a versatile framework for data processing. Though the ELT on Hadoop has been there for a while, it has not been widely adopted as change from conventional ETL tools to ELT on Hadoop is quite a huge adaptation.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
Tanuja Das*, Department of Information Technology, Gauhati University, Guwahati, India. Email: tanujadas55@gmail.com
Ramesh Saha, Department of Information Technology, Gauhati University, Guwahati, India. Email: ramesh1saha@gmail.com
Goutam Saha, Department of Information Technology, North Eastern Hill University, Shillong, India. Email: dr.goutamsaha@gmail.com

ETL tools are expected to stay in the industry based on their adaptation to handle query execution concurrently [17]. Also its ability handle large amount of data makes it the foundation of data warehouse systems.

One such example is Alooma [18] which is capable of acquiring, transforming and storing data from a huge number of transactions over a large range of data sources and streams. Such adaptations contribute to exciting new opportunities, especially in the domain of Machine Learning.

In this paper, we collect data from numerous heterogeneous sources from the web and try to transform them into the one of the most dominant file formats of a data warehouse, viz, the CSV (comma separated value) form.

II. BACKGROUND

Although the term ETL depicts a simple three-step process, in reality the process includes a number of intermediate steps. As a result of the existence of considerable number of new technologies and the overlapping among the different stages, it faces a lot of challenges. The process may not succeed due to numerous reasons like omitted data values or omitted extracts. So it becomes important that the ETL process takes the necessary steps for optimizing such challenges.

A properly constructed ETL process assures accurate transformation of the data into the required configuration. Organizations have already started to upgrade to the latest technology by generating their own mechanism to extract and transform real time data. In this section, first we discuss the working of the basic ETL process followed by how this process has been adapted for the Big data environment.

A. The Extract, Transform , Load process

In a nutshell, the ETL can be described as a mechanism that comprises of the following steps:

- Extracting data from various sources.
- Transformation of the data according to the need.
- Loading the data to the respective warehouse.

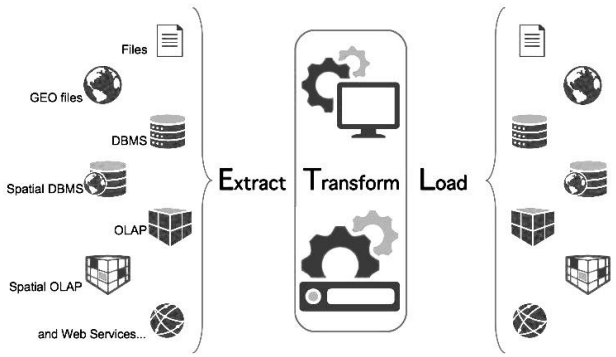

Figure 1: The ETL process [29]

Extract: The first and the most important step of the ETL are extracting data accurately from the respective sources. There are several systems with various formats from which data are assembled in the data warehouses. Along with the traditional data source frameworks like relational databases [19], there exists a variety of other formats like Information Management System (IMS) [20], Virtual Storage Access Method (VSAM) [21] or Indexed Sequential Access Method (ISAM) [22], etc. At times when no medium for storing the data is needed, ETL can be implemented by extracting the streaming data and loading it directly to the destination database [23]. Thus, the main objective of the extraction step is to bring the data to a common format such that it is desirable for the next step.

Transform: The next step in the ETL process involves implementing a sequence of rules to the data extracted from the previous step in order to make the data appropriate for the end goal. In order to accommodate the various requirements of the database in focus, transformation of the data is required [24]. Some of the rules are elaborately given in [25] and [26] mainly focussing on decomposition mechanisms and declarative mechanisms respectively. The decomposition mechanism uses triple model to signify the numerous heterogeneous data standards while the declarative mechanism aims to utilize the meta-data for the same. Validating the data after performing these transformations is a very crucial step as this data is to be supplied to the next step depending on the schema [27].

Load: The final step of the ETL process is to load the data, generally into the data warehouse. The technique differs extensively based on the needs of the various organizations. Some data warehouses may replace previous data with aggregate data or may append new data in historicized form [28]. The global quality of data from the ETL process is also dependent on the constraints described in the database schema. So any changes in the schema are to be handled appropriately in order to maintain the stability of the system. Also the way in which the system is to be recovered in case of a failure is an important aspect here.

B. ETL for Big Data

Various methods have been adapted and are still being adapted according to the growing need of the Big data environment. The form in which Big data arrives may be textual, multimedia, etc., which may be related or unrelated. Textual data forms a considerable part of the data generated by any organization, may it be structured or unstructured, like from e-mail, corporate documents, web pages, social media data, etc [30]. And, as this era has ventured into the world of internet technologies, multimedia big data has also come to the picture [31].

The data which arrives from all the above mentioned sources is very crucial in making real time predictions for various problems. Generally, the target of any organization is to transform the data so that it becomes consistent and can be appended with the existing data, can be transferred to other systems, or can be used to sum up information in the data [32]. Thus it is very important that supervision is taken on the data when it is being transformed so that it can cope up with the complexity of the data coming from all the distributed real time sources continuously.

Retrieval Number: B3438129219/2019©BEIESP
DOI: 10.35940/ijeat.B3438.129219
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

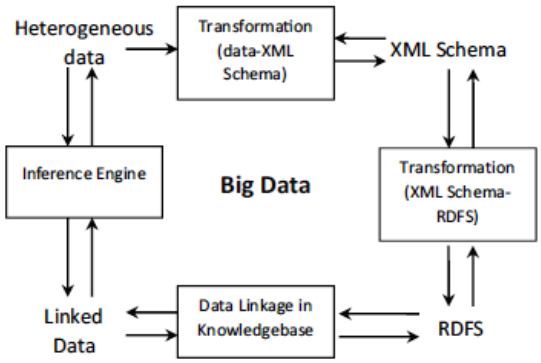

Figure 2: Transformation of Big Data from heterogeneous data in addition with an inference engine [11]

In [11], an enhancement has been proposed in the XML and RDF to conquer these problems. In the process as shown in Figure 2, Big Data is converted to DTD of a XML document. After this, the DTD is converted to RDFS.

III. EXPERIMENTAL PROCEDURE

In this paper, we try to implement the ETL procedure in context of Big data. The idea exploited here is the Semantic Web technologies [33] which is the most used technique exploited for meaningful description of data from a number of heterogeneous sources. For this, we extract data from the web and then try to transform them into a structured dataset, more specifically into csv format. The basic steps of the mechanism include:

- First, we locate the data source. In this process we do so by locating the HTML source based on the title of the page.
- Then, we try to figure out the structure of the data and the appropriate transformations it requires. Here we do that by converting the page DOM (document object model) object [Figure 3]. We then identify the tables from the model.
- The final step includes the mappings needed to define how the respective fields are to be handled. Now from the tables identified in the above step, we parse it to find the row elements of the table. As the table is somewhat unstructured in the web page consisting of images, urls, line breaks, icons and many more, data cleaning is also done along with parsing of the DOM.

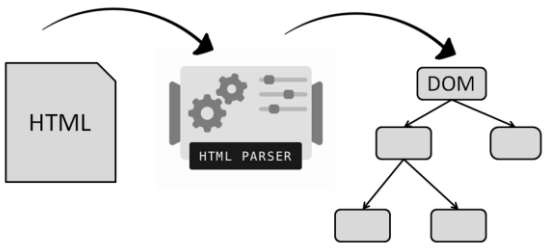

Figure 3: Transformation of heterogeneous data from web to DOM

For example considering a table, taken from an HTML document [34] :

```
<TABLE>
<ROWS>
<TR>
<TD>Rapunzel</TD>
<TD>Elsa</TD>
</TR>
<TR>
<TD>Tangled</TD>
<TD>Frozen</TD>
</TR>
</ROWS>
</TABLE>
```

The parsing of the fields is shown in Figure 4.

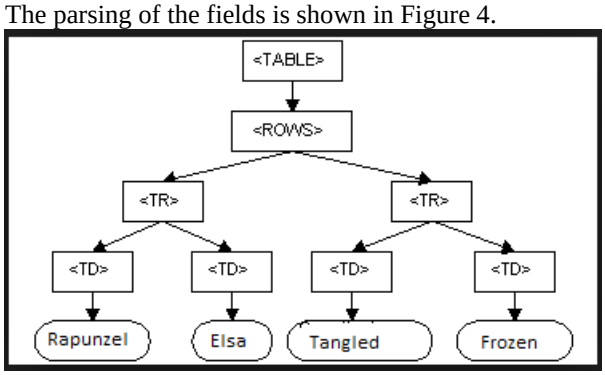

Figure 4: The parsing of rows in a DOM [34]

IV. RESULTS AND DISCUSSION

Here we transform a set of XML collections based on Wikipedia which are very relevant for a large variety of XML IR/Machine Learning tasks. The dataset we focus in this work is the wikipedia server [35]. Wikipedia is basically a multi-lingual online encyclopedia which is built on open collaboration through a wiki-based content editing system [35]. The files along with their URLs which we have taken for consideration are given in Table I.

Table I. The datasets used from Wikipedia server

Sl.No.	Title of the page	URL
1.	Religion in India [35]	https://en.wikipedia.org/wiki/Religion_in_India
2.	Demographics Of India [36]	https://en.wikipedia.org/wiki/Demographics_of_India
3.	List of states in India by past population [37]	https://en.wikipedia.org/wiki/List_of_states_in_India_by_past_population

We have successfully transformed the tables available in the above pages as shown in Table I into CSV format. The snapshot of the format is as shown in Figure 5-7.

```
pc34@pc34-20-b102in:~$ /usr/bin/python3 "/home/pc34/laptop materials/1 dataTransformation/transform4.py"
                           0                  1                  2  \
0               Religiousgroup  Population % 1951  Population % 1961
1                     Hinduism              84.1%             83.45%
2                        Islam               9.8%             10.69%
3                 Christianity              2.30%              2.44%
4                      Sikhism              1.79%              1.79%
5                     Buddhism              0.74%              0.74%
6                      Jainism              0.46%              0.46%
7               Zoroastrianism              0.13%              0.09%
8  Others/Religion not specified            0.43%              0.43%

                   3                  4                  5                  6  \
0  Population % 1971  Population % 1981  Population % 1991  Population % 2001
1             82.73%             82.30%             81.53%             80.46%
2             11.21%             11.75%             12.61%             13.43%
3              2.60%              2.44%              2.32%              2.34%
4              1.89%              1.92%              1.94%              1.87%
5              0.70%              0.70%              0.77%              0.77%
6              0.48%              0.47%              0.40%              0.41%
7              0.09%              0.09%              0.08%              0.06%
8              0.41%              0.42%              0.44%              0.72%

                       7
0  Population % 2011[59]
1                 79.80%
2                 14.23%
3                  2.30%
4                  1.72%
5                  0.70%
6                  0.37%
7            not counted
8                   0.9%
```

Figure 5: The transformed data for the dataset "Religion in India" in CSV format

```
guest-faculty-2@guestfaculty2-OptiPlex-3060:~/tanuja/dataTransformation$ /usr/bin/python3 /home/guest-faculty-2/transform4.py
                          Demographics of India                             Demographics of India.1
0   Map showing the population density of each dis...  Map showing the population density of each dis...
1                                      Population                  1,324,171,354 (2016 est.)[1]
2                                         Density                 382 people per.sq.km (2011 est.)
3                                     Growth rate                            1.19% (2016) (96th)
4                                      Birth rate         19.3 births/1,000 population (2016 est.)
5                                      Death rate          7.3 deaths/1,000 population (2016 est.)
6                                 Life expectancy                        68.89 years (2009 est.)
7                                          â¢ male                        67.46 years (2009 est.)
8                                        â¢ female                        72.61 years (2009 est.)
9                                  Fertility rate         2.2 children born/woman (2016 est.)[2]
10                          Infant mortality rate  41 deaths/1,000 live births (2016 est.)[citati...
11                                  Age structure                                   Age structure
12                                     0â14 years     28.6% (male 190,075,426/female 172,799,553)[2]
13                                    15â64 years  63.6% (male 381,446,079/female 359,802,209) (2...
14                                    65 and over  5.3% (male 29,364,920/female 32,591,030) (2009...
15                                      Sex ratio                                       Sex ratio
16                                       At birth                  1.10 male(s)/female (2013 est.)
```

Figure 6: The transformed data for the dataset "Demographics of India"

```
guest-faculty-2@guestfaculty2-OptiPlex-3060:~/tanuja/dataTransformation$ /usr/bin/python3 /home/guest-faculty-2/transform4.py
    Rank  State or union territory  ...  Population (2001 Census)[11]  Population (2011 Census)[11]
0      1             Uttar Pradesh  ...                  1.660536e+08                     199581477
1      2               Maharashtra  ...                  9.675250e+07                     112372972
2      3                     Bihar  ...                  8.287991e+07                     103804630
3      4               West Bengal  ...                  8.022130e+07                      91347736
4      5            Madhya Pradesh  ...                  6.038509e+07                      72597565
5      6                Tamil Nadu  ...                  6.211139e+07                      72138958
6      7                 Rajasthan  ...                  5.647330e+07                      68621012
7      8                 Karnataka  ...                  5.273499e+07                      61130704
8      9                   Gujarat  ...                  5.059720e+07                      60383628
9     10            Andhra Pradesh  ...                  7.572840e+07                      84580777
10    11                    Odisha  ...                  3.670790e+07                      41947358
11    12                 Telangana  ...                           NaN                      35193978
12    13                    Kerala  ...                  3.183900e+07                      33387677
13    14                 Jharkhand  ...                  2.694607e+07                      32988134
14    15                     Assam  ...                  2.663860e+07                      31169272
15    16                    Punjab  ...                  2.428913e+07                      27704236
16    17                   Haryana  ...                  2.108390e+07                      25753081
```

Figure 7: The transformed data for the dataset "List of states in India by past population"

Transformation of data remains traditionally an unavoidable phase of the web data integration mechanism. As seen from Figure 5-7, we have successfully transformed the data from web pages into a cleansed, validated, and ready-to-use form as it is very important for the respective researchers for a data source that is credible.

V. CONCLUSION

Data warehouses stores data from various sources for analysis and research. Most of the organizational decisions are based on the data stored in the warehouses. So it is imperative that the data which are being stored in the warehouses are in such a form that can be utilized appropriately. The transformation of Big data, in this context should be such that so that complexity is minimized for faster processing.

In this work, we try to transform the data generated in the web into a form which is very simple and also very common in data warehouses. The data considered for this work is basically collections of XML based on Wikipedia which is a growing source of information in today's world. In future, work can be done so as to do a comparative analysis on how this method can be made compatible the existing technologies of RDBMS.

ACKNOWLEDGMENT

The authors would like to acknowledge TEQIP III scheme of the Government of India for providing the required funds for the publication of this work.

REFERENCES

1. Syed, A., Gillela, K., & Venugopal, C. (2013). The future revolution on big data. Interna-tional Journal of Advanced Research in Computer and Communication Engineering, 2(6), 2446-2451.
2. "Gartner, IT." [Online], (2019) Available: http://www.gartner.com/it-glossary/big-data/
3. J. Dean, and S. Ghemawat, (2008) MapReduce: simplified data processing on large clusters, Communications of the ACM, vol.51, no.1, pp.107-113,
4. R. S. Chaulagain, S. Pandey, S. R. Basnet, & S. Shakya, (2017) Cloud based web scraping for big data applications, IEEE International Conference on Smart Cloud (SmartCloud, pp. 138-143, IEEE.
5. E. Rundensteiner, (1999) Special Issue on Data Transformation: ed., IEEE Techn. Bull. Data Engineering, vol.22, no.1.
6. S. Prabhu, (2007) Data mining and warehousing, New Age International.
7. B. Griesemer,(2009) Oracle Warehouse Builder 11g: Getting Started, Packt Publishing Ltd.
8. S. Soares, (2013) IBM InfoSphere: A platform for Big Data governance and process data gover-nance, Mc Press.
9. K. Haselden, and B. Baker, (2007) Microsoft SQL server 2005 integration services, Pearson Education India.
10. O. Lassila, and R. R. Swick, (1998) Resource description framework (RDF) model and syntax specification, Citeseer.
11. K. R.Malik, T. Ahmad, M. Farhan, M. Aslam, S. Jabbar, S. Khalid, and M. Kim, (2016) Big-data: transformation from heterogeneous data to semantically-enriched simplified data, Multi-media Tools and Applications, vol.75, no.20, pp.12727-12747,.
12. Halevy, A., Franklin, M., & Maier, D. (2006, June). Principles of dataspace systems. In Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 1-9). ACM.
13. Franklin, M. J. (2009, July). Dataspaces: progress and prospects. In British National Conference on Databases (pp. 1-3). Springer, Berlin, Heidelberg.
14. Hedeler, C., Belhajjame, K., Fernandes, A. A., Embury, S. M., & Paton, N. W. (2009, July). Dimensions of dataspaces. In British National Conference on Databases (pp. 55-66). Springer, Berlin, Heidelberg.
15. Mirza, H. T., Chen, L., & Chen, G. (2010). Practicability of dataspace systems. International Journal of Digital Content Technology and its Applications, 4(3), 233-243.
16. Davenport, R. J. (2008). ETL vs ELT a subjective view. Insource IT Consultancy Ltd.
17. Quinto, B. (2018). Big Data Warehousing. In Next-Generation Big Data (pp. 375-406). Apress, Berkeley, CA.
18. Alooma. 2018. "ETL Tools." January 4. https://www.etltools.net/
19. Haithcoat, T. (1999). Relational Database Management Systems, Database Design, and GIS. Missouri Spatial Data Information Service presentations.
20. Pearlson, K. E., Saunders, C. S., & Galletta, D. F. (2016). Managing and using information systems, binder ready version: a strategic approach. John Wiley & Sons.
21. Batra, R. (2018). A History of SQL and Relational Databases. In SQL Primer (pp. 183-187). Apress, Berkeley, CA.
22. Grand, A. (1989). U.S. Patent No. 4,823,310. Washington, DC: U.S. Patent and Trademark Office.
23. Patil, P. S., Rao, S., & Patil, S. B. (2011, February). Data integration problem of structural and semantic heterogeneity: data warehousing framework models for the optimization of the ETL processes. In Proceedings of the International Conference & Workshop on Emerging Trends in Technology (pp. 500-504). ACM.

24. Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002, November). Conceptual modeling for ETL processes. In Proceedings of the 5th ACM international workshop on Data Warehous-ing and OLAP (pp. 14-21). ACM.
25. Singh, M., & Jain, S. K. (2015). Transformation rules for decomposing heterogeneous data into triples. Journal of King Saud University-Computer and Information Sciences, 27(2), 181-192.
26. Tomingas, K., Kliimask, M., & Tammet, T. (2014). Mappings, Rules and Patterns in Tem-plate Based ETL Construction. In The 11th International Baltic DB & IS2014 Conference.
27. Homayouni, H., Ghosh, S., & Ray, I. (2018, June). An Approach for Testing the Extract-Transform-Load Process in Data Warehouse Systems. In Proceedings of the 22nd International Database Engineering & Applications Symposium (pp. 236-245). ACM.
28. Köppen, V., Brüggemann, B., & Berendt, B. (2011). Designing data integration: the ETL pattern approach. UPGRADE: the European Journal for the Informatics Professional, (3), 49-55.
29. Dhanda, P., & Sharma, N. (2016). Extract Transform Load Data with ETL Tools. Interna-tional Journal of Advanced Research in Computer Science, 7(3).
30. Khan, Z., & Vorley, T. (2017). Big data text analytics: an enabler of knowledge management. Journal of Knowledge Management, 21(1), 18-34.
31. Pouyanfar, S., Yang, Y., Chen, S. C., Shyu, M. L., & Iyengar, S. S. (2018). Multimedia big data analytics: A survey. ACM Computing Surveys (CSUR), 51(1), 10.
32. Wang, Y., Kung, L., Wang, W. Y. C., & Cegielski, C. G. (2018). An integrated big data analytics-enabled transformation model: Application to health care. Information & Management, 55(1), 64-79.
33. Saeed, M. R., Chelmis, C., & Prasanna, V. K. (2017). Automatic integration and querying of semantic rich heterogeneous data: Laying the foundations for semantic web of things. In Managing the Web of Things (pp. 251-273). Morgan Kaufmann.

34. Robie J, "What is the Document Object Model?", Texcel Research, July 19, 1998 [Online]. Available: https://www.w3.org/TR/WD-DOM/introduction.html[Accessed 29 June 2019].
35. "Religion in India", Oct. 19, 2018 [Online]. Available: https://en.wikipedia.org/wiki/Religion_in_India [Accessed 29 June 2019].
36. "Demographics Of India", Oct. 19, 2018 [Online]. Available: https://en.wikipedia.org/wiki/List_of_states_and_union_territories_of_India_by_population [Accessed 29 June 2019].
37. "List of states in India by past population", Oct. 19, 2018 [Online]. Available: https://en.wikipedia.org/wiki/List_of_states_in_India_by_past_population [Accessed 29 June 2019].

AUTHORS PROFILE

Tanuja Das, received the B.Tech degree in Information Technology from North Eastern Hill University, Shillong, India in 2012. She received the M.Tech degree from Tezpur University, Tezpur, India in 2014. Currently, she is working as an Assistant Professor (Under TEQIP III) in the Department of Information Technology, GUIST, Gauhati University, Guwahati, India and pursuing Ph.D from the Department of Information Technology, North Eastern Hill University, Shillong, India under the supervision of Dr. Goutam Saha, Professor, Department of IT, North Eastern Hill University, Shillong, India.

Ramesh Saha, received the B.Tech degree in Information Technology from Kalyani Govt. Engineering College, West Bengal , India in 2013. He received the M.Tech degree from N.I.T.T.T.R., Kolkata, Kolkata, West Bengal, India in 2015. Currently, He is working as an Assistant Professor (Under TEQIP III) in the Department of Information Technology, GUIST, Gauhati University, Guwahati, India and pursuing Ph.D. from the Department of Computer Science, Maulana Abul Kalam Azad University of Technology, West Bengal, India.

Goutam Saha, received the B. E. degree in Electrical Engineering and the M. E. degree in Electronics and Telecommunication Engineering from the Bengal Engineering College, Shibpur under the University of Calcutta, Kolkata, India in 1984 and 1989, respectively. He received the Ph. D. degree from the Indian Institute of Technology, Kharagpur, India in 1999. He also has Post-Doctoral Research experience at the Indian Institute of Technology, Kharagpur, India and the Ben Gurion University, Israel. Presently, he is working as the Professor at the Department of Information Technology, North Eastern Hill University, Shillong, India.