

Implementation of Central Dogma Based Cryptographic Algorithm in Data Warehouse Architecture for Performance Enhancement

Rajdeep Chowdhury¹

Assistant Professor, Department of Computer Application,
JIS College of Engineering,
Block-A, Phase-III, Kalyani,
Nadia-741235, West Bengal, India

Saswata Dasgupta³

Student, Department of Computer Science & Engineering,
JIS College of Engineering,
Block-A, Phase-III, Kalyani,
Nadia-741235, West Bengal, India

Soupayan Datta²

Student, Department of Computer Science & Engineering,
JIS College of Engineering,
Block-A, Phase-III, Kalyani,
Nadia-741235, West Bengal, India

Mallika De⁴

Professor, Department of Computer Science & Engineering,
Dr. Sudhir Chandra Sur Degree Engineering College, Surer
Math, Melabagan Estate, Basak Bagan, Dum Dum,
Kolkata-700074, West Bengal, India

Abstract—Data warehouse is a set of integrated databases deliberated to expand decision-making and problem solving, espousing exceedingly condensed data. Data warehouse happens to be progressively more accepted theme for contemporary researchers with respect to contemporary inclination towards industry and executive purview. The crucial tip of the proposed work is integrated on delivering an enhanced and an exclusive innovative model based on the intention of enhancing security measures, which at times have been found wanting and also ensuring improved accessibility using Hashing modus operandi. An unsullied algorithm was engendered using the concept of protein synthesis, prevalently studied in Genetics, that is, in the field of Biotechnology, wherein three steps are observed, namely; DNA Replication, Translation and Transcription. In the proposed algorithm, the two latter steps, that is, Translation and Transcription have been taken into account and the concept have been used for competent encryption and proficient decryption of data. Central Dogma Model is the name of the explicit model that accounts for and elucidates the course of action for Protein Synthesis using the Codons which compose the RNA and the DNA and are implicated in numerous bio-chemical processes in living organisms. It could be observed that subsequently a dual stratum of encryption and decryption mechanism has been employed for optimal security. The formulation of the immaculate Hashing modus operandi ensure that there would be considerable diminution of access time, keeping in mind the apt retrieval of all indispensable data from the data vaults.

The pertinent appliance of the proposed model with enhanced security might be in its significant service in a variety of organizations where accrual of protected data is of extreme magnitude. The variety of organizations might include educational organizations, corporate houses, medical establishments, private establishments and so on and so forth.

Keywords—Data Warehouse; Central Dogma; Replication; Translation; Transcription; Codon; Data Mart; Hashing

I. INTRODUCTION

Data warehouse is an entrenched depository of an organization's electronically summative data [1, 2, 5]. Data warehouse are designed with the objective to facilitate comprehensive reporting and proficient analysis. Endowing security for the warehouse is virtually an enormous vulnerability for any organization.

The proposed work is principally unforced, as it confers the formulation of the inventive architecture with the objective of enhancing security measures and data warehouse performance enhancement in the course of action [1, 2, 5, 6].

The proposed cryptographic algorithm introduces an intuitive as well as an innovative approach by employing a unanimously accepted concept in Genetics and Molecular Biology known as 'Central Dogma.' The Central Dogma introduced by Francis Crick endows with a sequential explanation of the flow of genetic information within a biological system. The general transfer principally consists of three activities, namely; DNA Replication, Transcription and Translation.

DNA Replication – DNA Replication is the process of engendering two identical replicas from one original DNA molecule. The biological process occurs in all living organisms and is the basis for biological inheritance.

Transcription – Transcription is the initial step of gene expression, in which a particular segment of DNA is copied into RNA by the enzyme RNA polymerase.

Translation – Translation is the process by which proteins are created. In the process, apiece codon codes for a specific amino acid and proteins are synthesized or rather they are found in the form of amino acids in the body.

Improved accessibility at the data warehouse using Hashing technique would ensure performance enhancement of the data warehouse in addition to the security measures adhered.

The entire paper deals with an interdisciplinary approach of employing the biological process by which protein is synthesized in living organisms, for development of an algorithm by which a plain text is converted to a cipher text.

A real world phenomenon is being employed in computation and its appliance on the data warehouse should endow the reader with an insight into implementation of bio-inspired algorithm and its prospective impact on the manner it is being perceived and professed, thereby solving computational problems.

II. LITERATURE SURVEY

The section focuses on the interrelated work available in the similar genre, encouraging the formulation of the concerned paper [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14].

There are quite a few interrelated works on data warehouse and its security measures, which have formerly been carried out, but the precise design, implementation and incorporation of the innovative algorithm using protein synthesis is something inventive as well as appealing for contemporary researchers. Whenever the perception culminates in the mind about design, implementation and incorporation of security mechanism for data warehouse as well as architectural orientation, the extensive study of few papers need meticulous mention and the concise points are affirmed in details for ease in reference [1, 2].

In the paper titled "Design and Implementation of Proposed Drawer Model Based Data Warehouse Architecture Incorporating DNA Translation Cryptographic Algorithm for Security Enhancement," the DNA Translation cryptographic algorithm is incorporated in the Proposed Drawer Model Based Data Warehouse Architecture in two distinguished tiers of security mechanism. Initially the algorithm is adhered during the transition from Operational Data Store to Data Vault and then it is adhered for the second time during the transition from Data Vault to Data Mart, thereby ensuring two-tier security enhancement for the proposed data warehouse model [1].

Through the formulation of the paper, discussions as well as Illustrations ensure how security could be implemented at distinguished tiers/levels using the innovative DNA Translation cryptographic algorithm through the proposed Drawer Model Based Data Warehouse Architecture as an Added measure over the existing Data Warehouse Architecture [1].

In the paper titled "Towards Data Security in Affordable Data Warehouse," the data warehouse technique is based on clustering and the star schema is dispersed over the nodes of the cluster. Dimension table is replicated in apiece node of the cluster and fact table is distributed using strict round robin or hash partitioning. Security is assured by using signature in

each column individually and in each row the verification is controlled by data warehouse middleware. Data warehouse middleware engender the signature for insert and update operations. In the concerned approach, encryption technique is applied on dimension table. Primary keys and foreign keys are not encrypted as they get filled with synthetic values. Encryption technique is not used in fact table due to performance issues. Fact table is fragmented into several clusters and fact data cannot be simply allied to the dimension data as they are encrypted [2, 6].

In the paper titled "An Integrated Conceptual Model for Temporal Data Warehouse Security," it has been proposed that the first integrated conceptual model for addressing temporal data warehouse security requirements needs specification. The model is the first model which combines ETL model with temporal data warehouse model in one integrated model.

ETL processes are accountable for extraction of data from heterogeneous operational data sources, their transformation (conversion, cleaning, normalization, etc.) and their loading into data warehouses. ETL proposed model has six fragmented steps, namely; 1-Source authentication, 2-Extractraction, 3-Filter process, 4- Incorrect process, 5-Surrogate process and 6-Load process [9]. These are three of numerous papers which have had an untiring impact during the formulation of the compiled paper.

III. PROPOSED WORK

For utmost simplicity in understanding, the flow chart of the entire proposed work has been designed in figure 1.

The proposed Central Dogma Based cryptographic algorithm is incorporated in the Data Warehouse Architecture for security as well as performance enhancement in distinguished tiers of security mechanism and improved accessibility using Hashing modus operandi. Initially the algorithm is adhered during the transition from Operational Data Store to Data Vault and then it is adhered for the second time during the transition from Data Vault/Storage Area to explicit segments of Data Mart, thereby ensuring two-tier security enhancement for the proposed data warehouse model. The interdisciplinary amalgamation of Biotechnology, especially genetics and Computer science, especially cryptography is an innovative approach, to say the least.

In the next fragment of the section, the detail of the proposed work has been specified and the commencement churns out with the notion about Central Dogma and the formulation of the proposed innovative Central Dogma Based cryptographic algorithm, as stated below for ease in reference:

The Central Dogma of Molecular Biology is an elucidation of the flow of genetic information within a biological system. It was initially stated by Francis Crick in 1956 and re-stated in a research paper published in 'Nature' in 1970. The Central Dogma of Molecular Biology deals with the exhaustive residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to protein or nucleic acid.

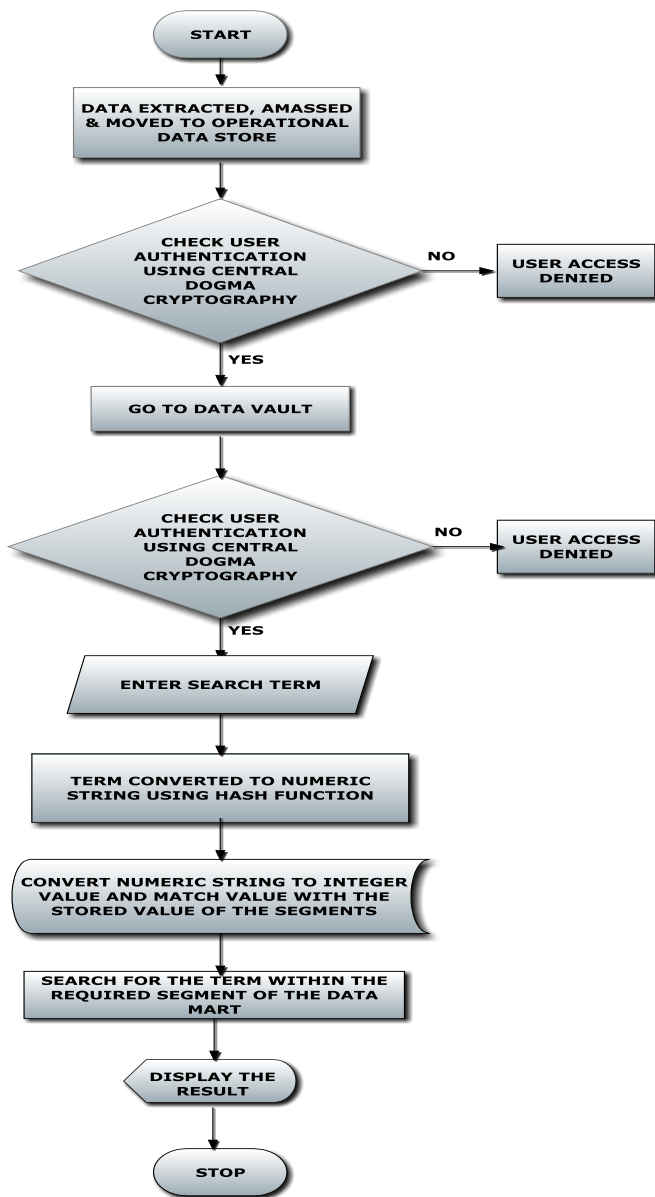


Fig. 1. Flow Chart of the Entire Proposed Work

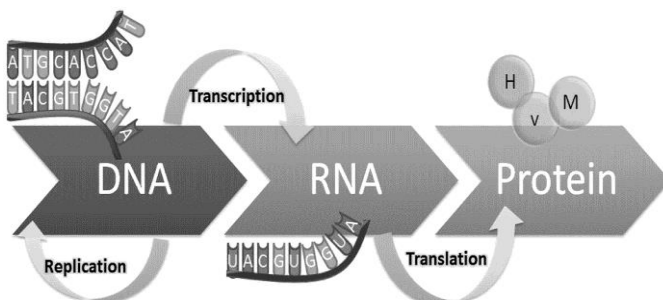


Fig. 2. Central Dogma Model

The Dogma is a framework for understanding the transfer of information sequence amid sequential information carrying biopolymers, in the most familiar or general case, in

living organisms. There are three major classes of such biopolymers, namely; **DNA** and **RNA** (Both nucleic acids) and **Protein**. There are $3 \times 3 = 9$ conceivable unswerving transfers of information that could occur amid these. The Dogma classifies these into three groups of three: Three General Transfers, which are believed to transpire normally in most cells; Three Special Transfers, which are believed to transpire, only under specific conditions, in case of some viruses or in laboratories; Three Unknown Transfers, which are believed never to transpire. The General Transfers illustrate the normal flow of biological information: DNA could be copied to DNA (DNA Replication); DNA information could be copied to mRNA (Transcription); Proteins could be synthesized using the information in mRNA as a template (Translation).

A. Proposed Central Dogma Based Cryptographic Algorithm

Proposed Central Dogma Based cryptographic algorithm employs the concept of protein synthesis which is an integral part of the ‘Central Dogma’ consisting of two steps, namely; Transcription and Translation. The procedure of Transcription engrosses transfer of information from DNA to RNA and the procedure of Translation engrosses the concluding step of protein synthesis, wherein specific group of nucleobases lead to specific protein synthesis.

ACGT stands for the four nucleic acid bases that make up DNA. ‘A’ stands for Adenine and pairs with ‘T’, which stands for Thymine. ‘C’ stands for Cytosine and pairs with ‘G’, which stands for Guanine. These four nucleic acids make up an organism’s genetic code or DNA.

ACGU stands for the four nucleic acids that make up RNA. RNA pairs up in the similar manner as DNA, except that Thymine is replaced with Uracil.

B. Extended ASCII Code and Encryption Key

The Extended ASCII code has been employed for ease in working, taking into account all feasible characters possible during authentication process. Extended ASCII is an 8-bit or larger character encoding modulus operandi that includes 7-bit ASCII characters as well as others. All the characters embodied within the Extended ASCII code could be represented by their equivalent 8-bit binary form.

Consequently for smooth functionality, the grouping mechanism for primary key has been employed, nevertheless to overlook the fact that this implicitly places the key into the category of public and symmetric encryption key.

Subsequently, as exemplified, the 8 bits have been clustered into 2 bits apiece, wherein apiece cluster of 2 bits embodies a specific nucleobase found in DNA (Ribonucleic Acid). After grouping in the mentioned manner, it is observed that there are 4 combinations of groups feasible, namely; 00, 01, 10, 11.

The feasible groups have been used to embody the specific nucleobase (Found in DNA) in the subsequent manner:
00 – A
01 – C
10 – G
11 – T

IV. ILLUSTRATION AND EXAMPLE

A. Steps to obtain Cipher Text

Step 1 – Transcription Phase →

1. The binary representation of the character concerned from the Extended ASCII Table is referred and are clustered/grouped into 2-bits apiece. Subsequently, four groups are obtained.
2. The groups are reinstated by the nucleobases as mentioned above.
3. A text is obtained which would contain the four nucleobases, which might look like → UUTC.
4. The code/text thus obtained in its DNA form, gets converted to its RNA form by the process of Transcription, following the rules mentioned below: T (Thymine) is replaced with U (Uracil), following the rule of complementary base pairing. For complementary base pairing, the following replacement is followed:

A → U & Vice Versa

G → C & Vice Versa

Step 2 – Translation Phase →

1. The Genetic Code Table is referred and it is observed that cluster/group of three nucleobases leads to the synthesis of a specific protein.
2. Employing the concept of Translation, the cipher text obtained by Transcription mechanism is further processed.

To illustrate the concept exclusively, it is assumed that the obtained cipher text from Transcription, for a random word of three letters is as mentioned below:

ACUGUCGACUAA

In order to translate the cipher text, the following rules are adhered:

1) Subsequent to the implementation of Transcription mechanism, for apiece character, a cipher text is obtained which consists of a combination of four nucleobases.

2) If the first nucleobase is left out, then the rest three nucleobases refer to a specific codon that code for a specific amino acid, as observed in the Genetic Code Table mentioned already.

3) It is observed that certain codon refers to the similar amino acid or rather a specific amino acid could be synthesized from multiple codons.

4) In order to avoid ambiguity, numbering mechanism should be employed, that is, '1' before the name of the amino acid (As visible in Genetic Code Table) should refer to the first codon from top, within a specific block, accountable for the synthesis of that amino acid.

5) As there are three STOP codons, UAA would be referred to as **1sto**, UAG would be referred to as **2sto** and UGA would be referred to as **3sto**.



Fig. 3. Nucleobase Pairing

		Second Position								
		U		C		A		G		
First Position		code	Amino Acid	code	Amino Acid	code	Amino Acid	code	Amino Acid	Third Position
		U	UUU	phe	UCU	ser	UAU	tyr	UGU	
	UUC		UCC		UAC		UGC		C	
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A	
	UUG		UCG		UAG	STOP	UGG	trp	G	
C	CUU		CCU		CAU	his	CGU		U	
	CUC	leu	CCC	pro	CAC		CGC	arg	C	
	CUA		CCA		CAA		CGA		A	
	CUG		CCG		CAG	gln	CGG		G	
A	AUU		ACU		AAU	asn	AGU	ser	U	
	AUC	ile	ACC	thr	AAC		AGC		C	
	AUA		ACA		AAA	lys	AGA	arg	A	
	AUG	met	ACG		AAG		AGG		G	
G	GUU		GCU		GAU	asp	GGU		U	
	GUC	val	GCC	ala	GAC		GGC	gly	C	
	GUA		GCA		GAA	glu	GGA		A	
	GUG		GCG		GAG		GGG		G	

Fig. 4. Genetic Code Table

6) Hence, a random character, viz, **ACUG**, after implementation of Translation mechanism should be:

A4leu

7) After implementation of Transcription mechanism, the entire cipher text representing a random word of three letters, that is, **ACUGUCGACUAA**, enduring the Translation mechanism should be:

A4leuU3proC1sto

Demonstration/Illustration of General Data Warehouse Architecture

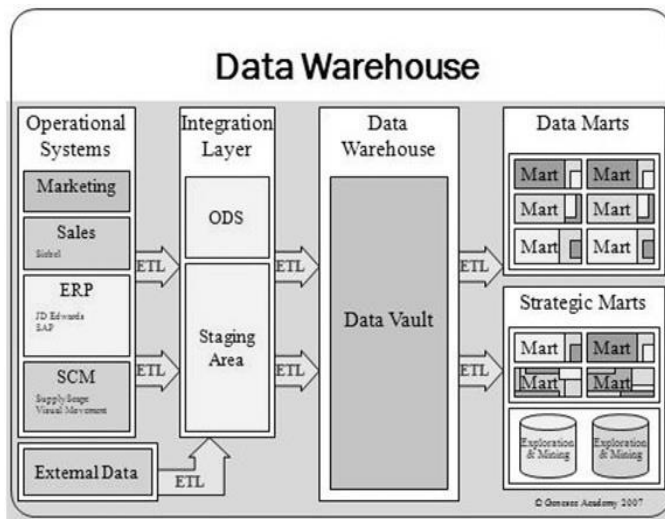


Fig. 5. General Data Warehouse Working Architecture

B. General Working of Data Warehouse

1) The elucidation of the working exemplifies that data being extracted from various **Operational Systems**. The extracted data are of distinguished format.

2) After extraction, the data are amassed at **Staging Area**, which is an intermediary storage locale used for data processing.

3) From staging area, the data is moved to **Operational Data Store (ODS)**, which assimilates all data and place them into a common format.

4) From ODS, data is moved to **Data Vault**, which is deliberated to endow with long term historical storage of data.

5) From data vault, the data is moved to different **Data Marts**, which is the access layer of data warehouse environment.

C. Data Vault Working

The working of the data vault could be fragmented into three components/processes performed on data, namely;

- a) OLAP Analysis
- b) Adept Reporting
- c) Data Mining

Proposed Hashing Technique Based Data Warehouse Architecture

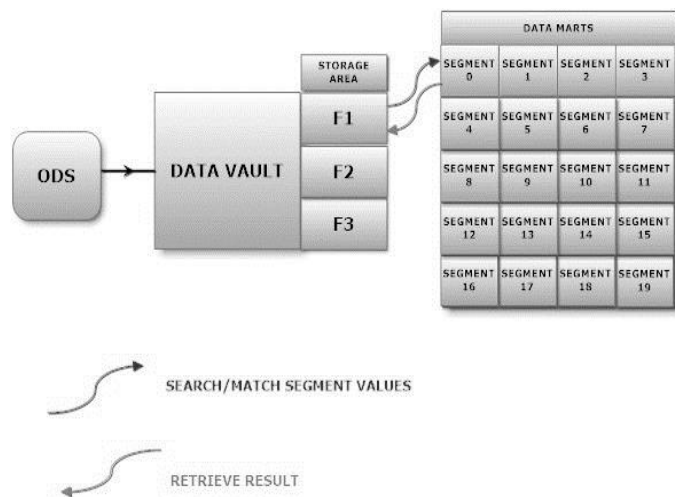


Fig. 6. Proposed Hashing Modus Operandi Based Data Warehouse Architecture

D. General Working of Proposed Hashing Modus Operandi Based Data Warehouse employing Central Dogma Based Cryptographic Algorithm for security enhancement

1) Initially, data is extracted from various **Operational Systems**, then amassed in **Staging Area** and then moved to **Operational Data Store**.

2) From **ODS**, data is moved to **Data Vault**, deliberated to endow with long term historical storage.

3) From **Data Vault**, data are shifted to newly designed **Storage Area**, where data is separated and amassed in distinguished folders. Apiece folder is specified to accumulate data for explicit **Data Mart**.

4) All the **Data Marts** are initially empty, wherein data are transmitted through folders to the dedicated **Data Marts**.

5) At a particular time, only apiece folder could be transmitted from **Storage Area** to **Data Mart**.

6) After triumphant authentication employing **Central Dogma Based Cryptographic Algorithm**, the user is allowed to access the **Data Mart**, wherein the **Data Mart** is fragmented into **Segments** numbered from 0 to 19.

7) A database pertaining to a bank should contain **Employee Unique ID, Name, Age, Sex, etc.**, which might be treated as **Segments**.

8) Gaining the access to the **Data Mart**, the search term concerning the requisite data is entered by the user, wherein the search term is converted into a numeric string employing the **Hashing function**:

$$F(n) = \text{Sum of the (digits or alphabets)} \bmod 20 \quad (1)$$

Set of rules are followed to engender an integer value which would denote the segment number.

The rules followed are:

(i) If digits would be present, then they would be added as usual.

Example: If Unique ID is 120110, then Sum is 5.

(ii) If alphabets would be present, then the respective digit value would be considered.

Example: a = 1, b = 2, c = 3, etc.

(iii) If an amalgamation of digits and alphabets would be present, then they would be converted in respective manner described already, followed by addition to obtain the Sum.

9) It is reasonably undemanding to search for an integer rather than an assortment of characters, ensuring diminution of accessibility time for retrieving the requisite data from the specific **Segment**.

10) Subsequent to finding a match with respect to the integer value obtained after **Hashing**, the search progresses by seeking the requisite and precise data within that specific integer **Segment**.

V. CONCLUSION

Whenever the term safety comes in intellect and initiative, security is synonymous, but from time to time implementing security mechanism(s) like cryptographic techniques, biometric methodologies, genetic algorithm, quick response code mechanisms, etc. has not only been sturdy but cost constrained as well. On the other hand, design, implementation and incorporation of any security methodology at internal structure of the data warehouse are affable.

The accessibility is explicitly improved by employing Hashing Modus Operandi, thereby ensuring diminution of access time and validating performance enhancement for the data warehouse. The design, implementation and incorporation of the Central Dogma Based cryptographic algorithm is the core of the conferred security amid predicament at bay like malicious intrusions.

Through the formulation of the paper, deliberations as well as illustrations ensure how security could be implemented and incorporated at distinguished tiers/levels using the innovative Central Dogma Based cryptographic algorithm through the proposed Hashing Modus Operandi Based Data Warehouse Architecture, aided as an additional quantifier over the existing Data Warehouse Architecture.

REFERENCES

- [1] Chowdhury, R., Dey, K., S., Datta, S., Shaw, S., "Design and Implementation of Proposed Drawer Model Based Data Warehouse Architecture Incorporating DNA Translation Cryptographic Algorithm for Security Enhancement", Proceedings of International Conference on Contemporary Computing and Informatics, (IC3I 2014), Organized by Sri Jayachamarajendra College of Engineering, Mysore, Proceedings in USB: CFP14AWQ-USB, ISBN-978-1-4799-6628-8, INSPEC Accession Number-14881472, Published and Archived in IEEE Digital Xplore, ISBN-978-1-4799-6629-5, pp. 55-60.
- [2] Chowdhury, R., Chatterjee, P., Mitra, P., Roy, O., "Design and Implementation of Security Mechanism for Data Warehouse Performance Enhancement using Two Tier User Authentication Techniques", International Journal of Innovative Research in Science, Engineering and Technology, An ISO 3297:2007 Certified Organization, 3 (6), 2014, Special Issue, pp. 441-449, ISSN (Online)-2319 8753, ISSN (Print)-2347 6710, Proceedings of National Conference on Emerging Technology and Applied Sciences-2014 (NCETAS 2014).
- [3] Chowdhury, R., Pal, B., Ghosh, A., De, M., "A Data Warehouse Architectural Design using Proposed Pseudo Mesh Schema", Proceedings of First International Conference on Intelligent Infrastructure (CSI ICII 2012), 47th Annual National Convention of Computer Society of India, Science City Auditorium, Kolkata, Tata McGraw Hill Education Private Limited, 2012, pp. 138-141.
- [4] Chowdhury, R., Pal, B., "Proposed Hybrid Data Warehouse Architecture based on Data Model", International Journal of Computer Science and Communication, 1 (2), 2010, pp. 211-213.
- [5] Saurabh, A., K., Nagpal, B., "A Survey on Current Security Strategies in Data Warehouses", International Journal of Engineering Science and Technology, 3 (4), 2011, pp. 3484-3488.
- [6] Vieira, M., Vieira, J., Madeira, H., "Towards Data Security in Affordable Data Warehouse", 7th European Dependable Computing Conference, (2008).
- [7] Patel, A., Patel, J., M., "Data Modeling Techniques for Data Warehouse", International Journal of Multidisciplinary Research, 2 (2), 2012, pp. 240-246.
- [8] Chaudhuri, S., Dayal, U., "An Overview of Data Warehousing and OLAP Technology", ACM SIGMOD Record, 26 (1), (1997).
- [9] Farhan, M., S., Marie, M., E., El-Fangary, L., M., Helmy, Y., K., "An Integrated Conceptual Model for Temporal Data Warehouse Security", Computer and Information Science, 4 (4), (2011), pp. 46-57.
- [10] Golfarelli, M., Maio, D., Rizzi, S., "The Dimensional Fact Model: A Conceptual Model for Data Warehouses", International Journal of Cooperative Information Systems, 7 (2-3), (1998), pp. 215-247.
- [11] Golfarelli, M., Rizzi, S., "A Methodological Framework for Data Warehouse Design", Proceedings of ACM First International Workshop on Data Warehousing and OLAP, DOLAP, Washington, (1998), pp. 3-9.
- [12] Chowdhury, R., Bose, R., Sengupta, N., De, M., "Logarithmic Formula Generated Seed based Cryptographic Technique using Proposed Alphanumeric Number System and Rubik Rotation Algorithm", Proceedings of IEEE 2012 International Conference on Communications, Devices and Intelligent Systems (CODIS 2012), Organized by Jadavpur University, Kolkata, Proceedings in CD: IEEE Catalog Number-CFP1207U-CDR, ISBN-978-1-4673-4698-6, Proceedings in Print: IEEE Catalog Number-CFP1207U-PRT, ISBN-978-1-4673-4697-9, INSPEC Accession Number-13285714, DOI-10.1109/CODIS.2012.6422265, Published and Archived in IEEE Digital Xplore, ISBN-978-1-4673-4700-6, pp. 564-567.
- [13] Chowdhury, R., Ghosh, S., De, M., "String Graphification based Asymmetric Key Cryptographic Algorithm using Proposed Concepts of GDC and S-Loop Matrix", Proceedings of IEEE/OSA/IAPR International Conference on Informatics, Electronics & Vision 2012 (ICIEV 2012), Organized by University of Dhaka, Dhaka, Bangladesh, Proceedings in CD: IEEE Catalog Number-CFP1244S-CDR, ISBN-978-1-4673-1152-6, Proceedings in Print: IEEE Catalog Number-CFP1244S-PRT, ISBN-978-1-4673-1151-9, Conference Proceedings: ISSN-2226 2105, INSPEC Accession Number-13058551, DOI-10.1109/ICIEV.2012.6317483, Published and Archived in IEEE Digital Xplore, ISBN-978-1-4673-1153-3, pp. 1152-1157.
- [14] Chowdhury, R., Gupta, S., Saha, A., "Stochastic Seed based Cryptographic Technique [SSCT] using Dual Formula Key [DFK]", Proceedings of International Conference on Communication and Industrial Applications (ICCIA 2011), Organized by Narula Institute of Technology at Science City, Kolkata, Proceedings in CD: IEEE Catalog Number-CFP1135R-CDR, ISBN-978-1-4577-1916-5, Proceedings in Print: IEEE Catalog Number-CFP1135R-PRT, ISBN-978-1-4577-1915-8, INSPEC Accession Number-12540264, DOI-10.1109/ICCIndA.2011.6146660, Published and Archived in IEEE Digital Xplore, ISBN-978-1-4577-1915-8, pp. 1-5.