

Towards Design, Analysis and Performance Enhancement of Data Warehouse By Implementation and Simulation of P2P Technology on Proposed Pseudo Mesh Architecture

Rajdeep Chowdhury¹, Saurab Dutta², Mallika De³

Assistant Professor, Department of Computer Application, JIS College of Engineering, Block-A, Phase-III, Kalyani,
Nadia, West Bengal, India^{1,2}

Professor, Department of Engineering and Technological Studies, University of Kalyani, Kalyani, Nadia, West Bengal,
India³

ABSTRACT: The data warehouse architectural design using proposed pseudo mesh schema have been the conceptualization backbone towards design, analysis and performance enhancement of the data warehouse by implementation and simulation of P2P technology on proposed pseudo mesh architecture.

A set of dimension tables interlinked with one another, ensuring enhancement in the time complexity of the data warehouse is presented through the formulation of the proposed work. Fact tables and dimension tables have been used here, in their normalized form, to diminish redundancy. The notion of pseudo mesh schema architecture, conforming interlinking of all dimension tables and fact tables to one another, have been proposed at the very inception of the devised work. The number of links between various peers in the system could be precisely calculated using $n(n-1)/2$, wherein n exemplifies the number of peers in attendance within the structure. To comprehensively emphasize the simulation methodology adhered, diverse networks are associated to each other using routers.

The structure is evidently flexible, as any increase or decrease of one or more databases within the structure, does not affect the entire schema of the data warehouse in concern. The structure uses the perception of fact views connecting with existing dimension tables.

Peer-to-Peer or P2P technology enables unswerving allotment of computer resources amid two or more computers within a computer network. In a P2P technology, apiece computer operate as mini server, where apiece computer could share its resources with other computers proficiently.

The equivalent notion has been implemented in the data warehouse architectural design using proposed pseudo mesh schema, where apiece dimension table in the architecture act as super peer and in consequence form a peer community, with collection of peers. Apiece peer within similar peer community or from unlike peer community could have unswerving relationship amid them using view fact table(s).

The improvement in efficiency by reduction of ineffectual searches and utilization of fewer resources is ensured by formulating unswerving exchange of records amid peers using view fact table(s).

By using views, the data is desired to be stored only once and then needs to be converted into the logical view definition, thereby diminishing data redundancy, ensuring consistency and simplifying data management responsibilities, entailing no physical storage

KEYWORDS: Peer-to-Peer, Snow Flake Schema, Star Schema, Dimension Table, Fact Table, Super Peer, Peer, Pseudo Mesh

I.INTRODUCTION

Data warehouse [3, 4, 13] is a set of incorporated and integrated databases intended to support decision-making and problem solving, comprising of highly summarized data [5, 7].

Data warehouse has become an increasingly accepted topic for contemporary researchers with respect to current trends of business organizational purview [3, 4, 5, 6, 7, 8, 9, 10, 11, 12].

Data warehouses are designed distinctively to facilitate comprehensive reporting and adept analysis, such that it could be of notable assistance to the researchers [2, 4].

As inferred from literature, data warehouse [3, 4, 13] is invariably fabricated using either star schema [14] or snowflake schema [15].

Star schema embraces of one or more fact table(s) connected with dimension tables. Centre of the star schema consist of one or more fact table(s) and the fact points to distinguished dimension tables [7, 14].

Snowflake schema is an extension of star schema where apiece point of the star explodes into more points [15].

The distinction initiates from the fact that in star schema, fact tables are in normalized format and dimension tables are in un-normalized format, keeping queries uncomplicated and endowing fast response time, whereas, in snowflake schema, both fact tables and dimension tables are in normalized format, thereby plummeting the query performance, on the basis of existence of more joins [1, 14, 15].

In star schema, each dimension is represented by a solitary dimension table, whereas in snowflake schema, that particular dimension table is normalized into multiple lookup tables, each representing a level in the dimension hierarchy [1, 14, 15].

Both the architectures are well accepted by the industry, ignoring the ambiguity allied with them. In star schema [14], a solitary fact table has numerous dimension tables associated with it and in snow-flake schema [15]; it further creates smaller dimension tables from original dimension tables [12]. Communication among two or more dimension tables in both the concepts is only through the fact table.

Peer-to-Peer refers to a computer network that has no centralized server. As mentioned earlier, apiece computer within the network operate as mini server, hence any loss of information from one computer could be compensated by other computers in the network [16, 17]. Furthermore, Peer-to-Peer networked architecture would be extremely beneficial, when question of reliability would crop in [16, 17, 20].

Since, information is not amassed in a central server and has back-up in an assortment of computers within the Peer-to-Peer network; recovery is not of much concern. There is no concept of central storage and in case any peer in the network would break down, it would in no way affect the entire system and could be effortlessly managed and recovered with the assistance of other peers in the network [18, 19].

The proposed system is based on the concept of Peer-to-Peer (P2P) architectural implementation, functional over the proposed pseudo mesh architectural data warehouse [1], where apiece node or sub dimensional table(s) emerging from dimension tables or super peers operate as individual peer, within the entire pseudo mesh schema data warehouse [1]. The super peer(s) essentially maintain the reference of the records with respect to the particular port number. Apiece record has individualistic port number and a group of port numbers specify a unique peer within the system. Apiece record is categorized beneath some explicit peer(s) based on certain characteristics/distinctiveness. The system diminishes the work load from the central fact table, thereby diminishing the load from the dimension tables [1].

Moreover, no additional space would be necessary for query processing, since the architecture itself implements the concept of view, which being a logical perception. A view relationship would be essentially established every time a query would be processed using two or more records from dissimilar peers.

The process of searching is essentially accomplished by use of the indexing technique. The indexing technique used would be small in size and would be able to operate with other indexes for filtering out the records before accessing the actual information [3].

Concerning performance enhancement, the proposed schema would effectively speed up the query performance by ensuring connection amid all super peers, thereby diminishing the exploration time for any record from the super peer table. The simulated result engendered using CISCO 2811 router connecting various nodes in the network and ARP,

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 6, February 2014

National Conference on Emerging Technology and Applied Sciences-2014 (NCETAS 2014)

On 15th to 16th February, Organized by

Modern Institute of Engineering and Technology, Bandel, Hooghly 712123, West Bengal, India.

ICMP protocols used for simulation ensures comprehensive analysis and proficient utilization of P2P architectural design for performance enhancement of data warehouse [21, 22].

II. PROPOSED WORK

The proposed schema is based on the pseudo mesh architecture, where each node, rather dimension table(s) is connected to other dimensional table(s) within the data warehouse purview, with the concept of views generated from the original fact table. Based on requirement, the number of dimension tables could be increased in number. With increase in dimensional tables, there would be no amend in data within an existing design, but only connectivity between dimensional tables would increase.

The design devised is said to be flexible and in apiece fragment, there has to be a fact table included compulsorily, which would contain K_n keys corresponding to n dimensional tables within the data warehouse.

The fact tables are consequently connected to all other dimensional tables.

The proposed work essentially implements the concept of Peer-to-Peer (P2P) file sharing technique on the existing pseudo mesh architectural data warehouse. A pseudo mesh architectural data warehouse comprises of a central fact table which remains connected to various dimension tables.

The primary logic of the architecture is that the dimension tables which remain inter-connected to each other diminish superfluous searches all the way through the data warehouse.

In the proposed system, apiece dimension table operate as a super peer which is itself connected to an assortment of sub peers, forming a peer community. Apiece record within a peer is updated to its corresponding super peer with the assistance of a unique peer port number, representing apiece peer. The super peer essentially maintains a table within that contains the record name and an associated unique port number that would direct the peer containing the record. The concept of port number is pertinent for all peers, wherein apiece record within the data warehouse could be acknowledged and identified with the unique port number.

A pseudo mesh architectural data warehouse could have N dimension tables, where apiece dimension table could be linked up with n peers, thereby forming a peer community amid themselves.

A record would be assigned a unique port number at the time the record would be inserted enduringly into the data warehouse, as soon as the ETCL process would be accomplished.

The unique port number could be further utilized for effortless searches of record(s) within the data warehouse.

Any peer within a peer community requesting for a record would initially search its own peer community by means of super peer table that maintains all the names of the record along with the associated port number.

If the record that the requesting peer would be asking for is found in the super peer table, then it would instantaneously return the port number of the peer that would contain the record of the requesting peer.

Based on the port number received from the super peer, the requesting peer would communicate with the peer containing the record using a view fact table. Since the concept of view is logical, therefore it does not occupy any additional space within the data warehouse.

An alternative circumstance could arise for the preceding case.

If the requested record is not found in the given peer community or the port number for the requested record does not exist in the given super peer, then the search would be continued in the neighboring super peer(s) until the requested peer obtain the port number of the peer with the located record.

Once it is a triumphant endeavor to accomplish the port number from any super peer, it would communicate to the peer unswervingly by establishing an association with that particular peer of the neighboring peer community, using a view fact table.

The maximum number of view tables possible within the proposed architecture would be $n(n-1)/2$, wherein n exemplifies the number of peers in attendance within the structure.

The very conception could be well understood from Figure-1, which comprises of a fact table and two dimension tables of super peers. From apiece super peer, two peers emerge. The maximum number of view table(s) that could formulate relations amid the peers would be equivalent to $n(n-1)/2$, which is equivalent to six herein, since the total number of

peers in the given system is four.

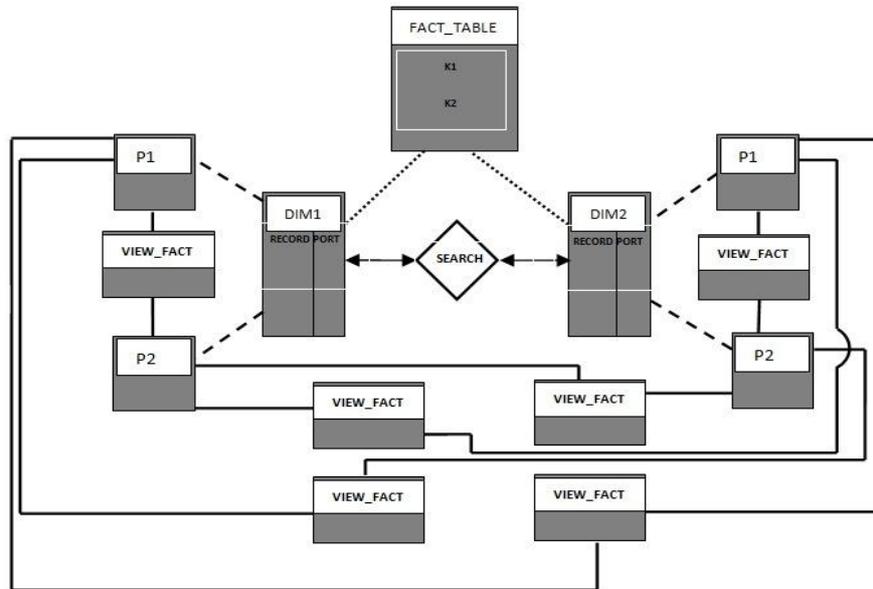


Fig-1: Proposed pseudo mesh schema for two dimension tables with two peers under a piece super peer, showing an assortment of likely relations between peers

The entire functioning of the system could be well exemplified with the assistance of the diagram embodied in Figure-2. The figure embodies the pseudo mesh architectural data warehouse design having a centralized fact table and three super peers, wherein a piece super peer makes up a peer community with n peers respectively. Subsequently, the total number of peers present in the system would be the summation of all the peers beneath all the super peers. The maximum number of possible view fact tables in the system would be calculated accordingly with $n(n-1)/2$, wherein n exemplifies the number of peers in attendance within the structure.

In Figure-2, presume a query that needs to be processed using the combined records from peers P1 and P2 respectively. For instance, P1 is searching for a record X that is present in any peer under super peer or DIM1. P1 entails sending a request for the record it requires. Based on the request from peer P1, dimension table DIM1 or super peer 1 checks its table and if the record is found then it instantaneously returns the port number to the requesting peer. On retrieving the port number, peer P1 communicates with peer P3 as per the return of port number from super peer 1, thereby establishing a relationship between them, using view fact table and conclusively, the query gets processed. Similarly, peer 2 or P2 of DIM2 would be requesting for a record Y to its DIM2 or super peer 2. DIM2 searches its table and if the record is not found present in the table, it continues its search to the neighboring super peers, that is, DIM2 and DIM3.

The record P2 of DIM2 requested would be found in peer P3 of DIM3. The port number with respect to the requested record gets returned to P2 of DIM 2. P2 instantaneously communicates with P3 of DIM3, as per the port number received.

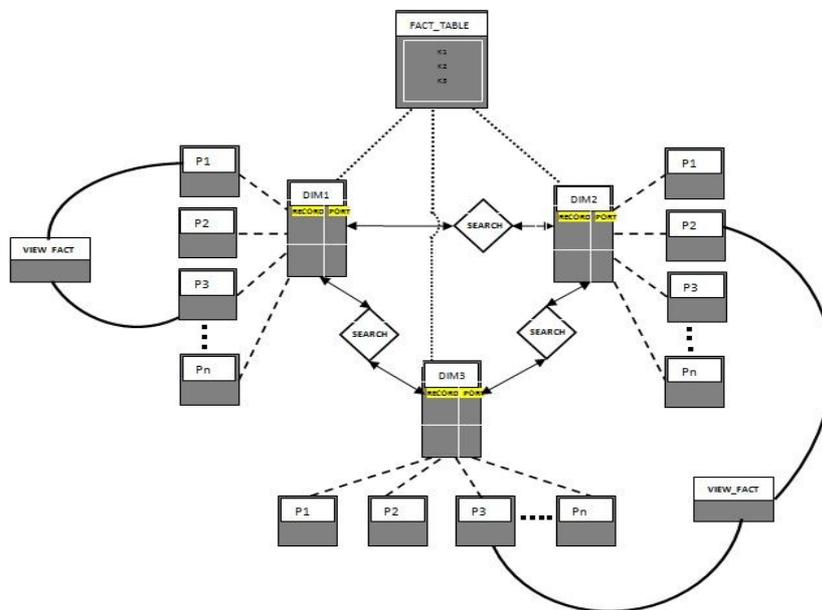


Fig-2: Proposed pseudo mesh schema with three super peers or dimension table

III. MATHEMATICAL REPRESENTATION

Let, K_1, K_2, \dots, K_n be the set of keys for any fact table.

K_i embodies the set of keys for any fact table which would be used to link up with an assortment of dimension tables, wherein value of i varies between $0 < i < n$

As per the proposed architecture, the numeral feasible dimension tables or super peers that could exist in the system are n .

Each super peer could be de-normalized and could have N number of peers connected to it.

The number of peers in the system could be mathematically represented as shown below:

The number of peers depends on the number of dimension tables taking part in the system.

Solitary dimension table can have n number of peers connected to it.

Hence,

Dimension table D1 has N_1 peers

Dimension table D2 has N_2 peers

.....

.....

Dimension table Dn has N_n peers

Therefore, total number of peers in the system could be represented as:

$$TP = N_1 + N_2 + \dots + N_{n-1} + N_n$$

or

$$\sum_{i=1}^n N_i$$

Apiece peer under alike or unlike super peer could establish connection between view fact table(s).

Therefore, maximum number of view fact table(s) possible depends on the value of TP.

VF represents the number of view fact tables.

This could be represented as:

$$VF = TP(TP-1)/2$$

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume3, Special Issue 6, February 2014

National Conference on Emerging Technology and Applied Sciences-2014 (NCETAS 2014)

On 15th to 16th February, Organized by

Modern Institute of Engineering and Technology, Bandel, Hooghly 712123, West Bengal, India.

In addition to the unanimously accepted existing schema architectures, the proposed schema finds its unfathomable foundation too, as it furnishes the user(s) with distinct manner of designing and formulating a data warehouse and its allied distinctiveness.

Moreover, P2P technology has diverse edge over centralized architecture, and the use of the concept along with the proposed pseudo mesh schema furnishes an option.

P2P architecture is effortless to be ascertained and could be configured with utmost ease.

All resources could be effortlessly shared with other peers within the network.

P2P architecture ascertains reliability as central dependency gets eradicated and could be supportive for developing distributed data warehouse(s).

The concerned architecture also diminishes the strain allied with the central administrator functionality, as apiece super peer operates as the administrator for apiece peer community.

In reference to the table of comparative study, the performance of the proposed schema could be recognized with utmost ease. Lesser are the number of queries amid the primary dimension table(s), however, complex queries might appear in the internal processing of apiece dimension table, due to additional number of foreign keys, which could be further exterminated through apt query optimization techniques.

Apiece dimension table would be expected to be highly normalized, in lieu with the architectural design devised. Apiece dimension table could be conked out into n number of sub dimension tables. The architectural design comprises only of solitary fact table and n number of dimension tables attached to the fact table.

Nevertheless, the speed could be optimized as there are unswerving relationships amongst the various dimension tables, which in the process diminishes needless comparisons amongst the various dimension tables in a data warehouse. The concept not only diminishes the query processing time but also diminishes storage complexity.

Establishment of relationships amid various dimension tables entail some space and for substantial data warehouses, the space gradually intensifies. The relation amid various table(s) is established through view, which is a logical concept and requires no memory space.

Conclusively, it could be inferred that maintaining a data warehouse using pseudo mesh schema could be relatively difficult, as with increase in dimension table(s), the number of view table(s) also increases. However, increase in view tables does not engender the predicament of space complexity rather it further enhances unswerving relation amid dimension tables.

V.SIMULATION RESULT VIA SCREENSHOT(S)

The section comprehensively emphasizes on the simulation result analysis, wherein CISCO packet tracer software has been used for simulation. All nodes act as mini server and an optional DNS server is also reserved. As packet is sent from source to destination and acknowledgement from destination is obtained, the total simulation time is recorded.

Figure-3, internetworks are connected via 2811 router and within solitary network, all nodes are connected via 2950 (24 port switch). All nodes within the network are connected using fast Ethernet.

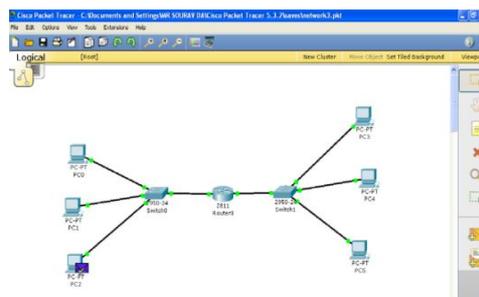


Fig-3: Internetwork

As packet is sent from apiece node and acknowledgement is obtained therein, the total simulation time is recorded and the experimental result is 177.717 seconds, as ensured by the sequential fragmented steps of Figure-4.

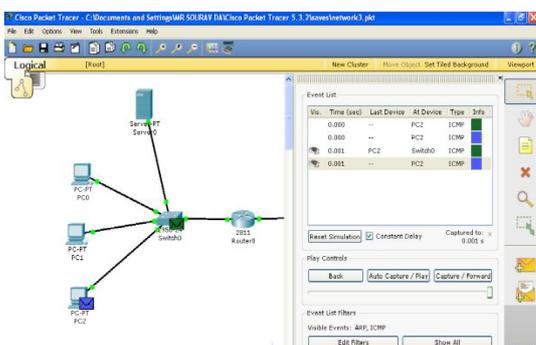


Fig-4(a): Step-1

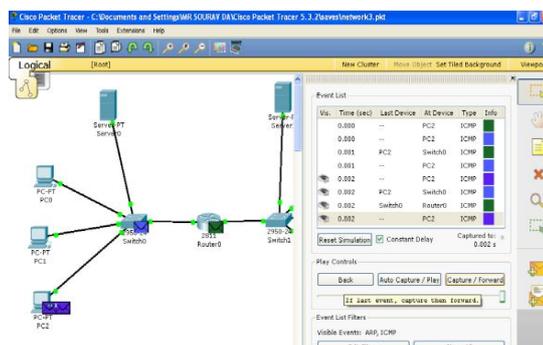


Fig-4(b): Step-2

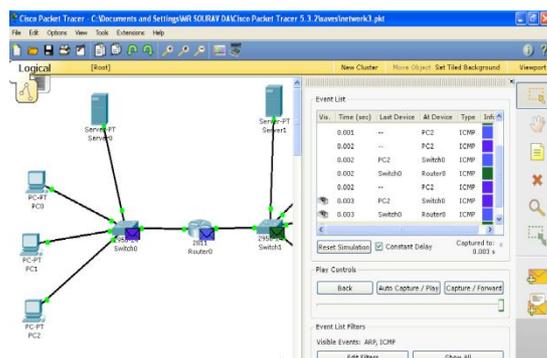


Fig-4(c): Step-3

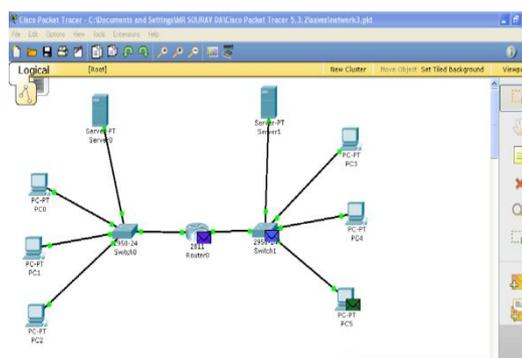


Fig-4(d): Step-4

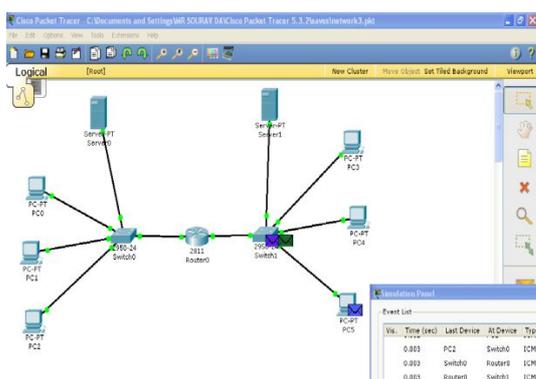


Fig-4(e): Step-5

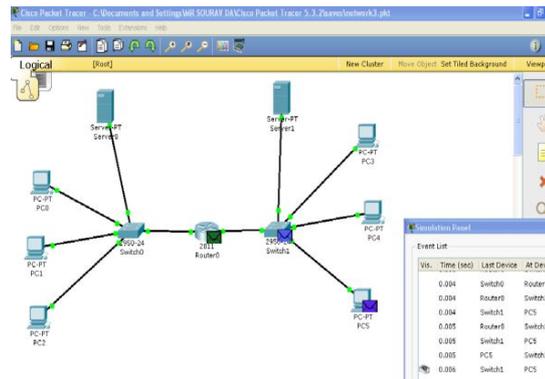


Fig-4(f): Step-6

utilization of P2P technology could only enhance during development of distributed data warehouse(s). Distributed data warehouse(s) could be well utilized by any organization which would be geographically dispersed. As far as performance enhancement is concerned, the proposed schema could speed up the query performance by connecting all super peers to each other, thereby diminishing the search time for any record from the super peer table. The search process is essentially accomplished by using the technique of indexing. The indexing technique to be used would be small in size and would be able to operate with other indexes for filtering out the records before accessing the actual information.

Apart from the use of the indexing technique towards performance enhancement, the system also implements the concept of view fact table. Views present data to users in a way they would be expecting to observe it, even if they define the identical element in a different way. By using views, the data is desired to be stored only once and then needs to be converted into the logical view definition, thereby diminishing data redundancy, ensuring consistency and simplifying data management responsibilities.

The simulated result engendered using router linking various nodes in the network and protocols used for simulation, is definitively established in the Simulation Result via Screenshot(s) section with justifiable screenshots establishing internetwork as well as entire simulation procedure.

REFERENCES

- [1] 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 '12, 47th Annual National Convention of Computer Society of India, Science City Auditorium, Kolkata, Tata McGraw Hill Education Private Limited, (2012), pp. 138–141.
- [2] 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.
- [3] KALIDO Dynamic Information Warehouse – A Technical Overview, KALIDO White Paper, (2004).
- [4] Chaudhuri, S., Dayal, U., "An Overview of Data Warehousing and OLAP Technology", ACM SIGMOD Record, 26 (1), (1997).
- [5] Bebel, B., Eder, J., Koncilia, C., Morzy, T., Wrembel, R., "Creation and Management of Versions in Multiversion Data Warehouse", Proceedings of 2004 ACM Symposium on Applied Computing, SAC '04, Nicosia, Cyprus, ACM, (2004), pp. 717–723.
- [6] Blaschka, M., Sapia, C., Höfling, G., "On Schema Evolution in Multidimensional Databases", Proceedings of First International Conference, Data Warehousing and Knowledge Discovery, DaWaK '99, Florence, Italy, LNCS, Springer, (1999), pp. 153–164.
- [7] Patel, A., Patel, J., M., "Data Modeling Techniques for Data Warehouse", International Journal of Multidisciplinary Research, 2 (2), (2012), pp. 240–246.
- [8] Cosmadakis, S., S., Kanellakis, P., C., "Functional and Inclusion Dependencies: A Graph Theoretic Approach", Proceedings of Third ACM SIGACT–SIGMOD Symposium on Principles of Database Systems, PODS '84, ACM, (1984), pp. 29–37.
- [9] De Castro, C., Grandi, F., Scalas, M., R., "On Schema Versioning in Temporal Databases", Proceedings of the International Workshop on Temporal Databases, Zurich, Switzerland, Springer, (1995), pp. 272–291.
- [10] Eder, J., Koncilia, C., Morzy, T., "The COMET Metamodel for Temporal Data Warehouses", Proceedings of Fourteenth International Conference on Advanced Information Systems Engineering, CAISE '02, London, United Kingdom, Springer, (2002), pp. 83–99.
- [11] Eder, J., Koncilia, C., "Changes of Dimension Data in Temporal Data Warehouses", Proceedings of Third International Conference, Data Warehousing and Knowledge Discovery, DaWaK '01, Munich, Germany, LNCS, Springer, (2001), pp. 284–293.
- [12] 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.
- [13] 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.
- [14] http://en.wikipedia.org/wiki/Star_schema.
- [15] http://en.wikipedia.org/wiki/Snowflake_schema.
- [16] Xiong, L., Liu, L., "Peer Trust: Supporting Reputation-Based Trust for Peer-to-Peer Electronic Communities", IEEE Transactions on Knowledge and Data Engineering, 16 (7), (2004), pp. 843–857.
- [17] Yang, X., de Veciana, G., "Service Capacity of Peer to Peer Networks", Proceedings of IEEE INFOCOM '04, (2004).
- [18] Pouwelse, J., A., Garbacki, P., Epema, D., H., J., Sips, H., J., "The Bittorrent P2P File-sharing System: Measurements and Analysis", Proceedings of Peer-to-Peer Systems IV, Springer, (2005), pp. 205–216.
- [19] Venkataraman, V., Francis, P., Calandrino, J., "ChunkySpread: Multi-tree Unstructured Peer-to-Peer Multicast", Proceedings of Fifth International Workshop on Peer-to-Peer Systems, Cornell University, Santa Barbara, CA, USA, (2006).
- [20] Stutzbach, D., Rejaie, R., "Understanding Churn in Peer-to-Peer Networks", Proceedings of Sixth ACM SIGCOMM Conference on Internet Measurement, IMC '06, ACM, (2006), pp. 189–202.
- [21] <http://www.cisco.com/en/US/products/ps5881>.
- [22] <https://learningnetwork.cisco.com/thread/36117>.

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume3, Special Issue 6, February 2014

National Conference on Emerging Technology and Applied Sciences-2014 (NCETAS 2014)

On 15th to 16th February, Organized by

Modern Institute of Engineering and Technology, Bandel, Hooghly 712123, West Bengal, India.

BIOGRAPHY



Mr. Rajdeep Chowdhury is presently engaged with pursuing his Ph.D. degree and is registered at Department of Engineering and Technological Studies, University of Kalyani under the supervision of Dr. Mallika De. Mr. Chowdhury was awarded Masters of Computer Application (MCA) from JIS College of Engineering under West Bengal University of Technology. Mr. Chowdhury's fields of interest include Network Security and Cryptography, Database Management System, Data Warehouse and Data Mining. Mr. Chowdhury has authored/co-authored several research articles/papers in refereed National/International Journals and National/International Conferences. Mr. Chowdhury has worked as Reviewer for few National/International Conferences as well as National/International Journals and has been part of Editorial Boards as well. Mr. Chowdhury has been recipient of numerous Awards and Honour by Corporate bigwigs and Academia. Few notable of which being; Winner at INSPIRE-Infosys Faculty Contest Series/INSPIRE-Infosys Faculty Excellence Awards in 2012 and 2013, Certificate of Recognition for Outstanding Contribution by Infosys for the year 2011-2012, Top Practitioner at Wipro Mission 10X Bookrack, Best Paper Awards in many National/International Conferences and Contests, etc.



Mr. Saurab Dutta is presently engaged in Research and Development. Mr. Dutta was awarded M.Tech from Bengal Engineering and Science University, Shibpur and Masters of Computer Application (MCA) from Haldia Institute of Technology under West Bengal University of Technology. Mr. Dutta's fields of interest include Network Security, Network-on-chip and Routing Algorithm.



Dr. Mallika De was awarded M.Tech degree in Computer Science from Indian Statistical Institute. Dr. De was awarded Ph.D. degree in Engineering from Jadavpur University. Dr. De is currently a senior faculty in the Department of Engineering and Technological Studies at University of Kalyani. Dr. De's research interest includes Parallel Algorithms & Architectures, Fault-tolerant Computing, Image Processing, Soft Computing, Data Warehousing and Data Mining. Dr. De has authored/co-authored several research papers in refereed National/International Journals and National/International Conferences. Dr. De has worked as Reviewer for few International Conferences such as Advanced Computing and Communication, High Performance Computing and Asian Test Symposium and has been part of Editorial Boards as well.