EMPLOYING CONCEPTUAL METHODOLOGIES, SPARQL2HIVE PERFORMS SPARQL QUERY ON HIVE

[1]Shaik Heena, [2]Dr. M. Suresh, [3]D.Naga Mani, [4]U. Sai Bhavya, [5]Dr. T. Seshaiah
[1, 2, 3, 4] Department of Computer Science and Engineering
[5] Department of Mechanical Engineering
[1, 2,3,4,5] QIS College of Engineering and Technology, Ongole
[1]heena.sk@qiscet.edu.in, [2]suresh.m@qiscet.edu.in, [3]nagamani.d@qiscet.edu.in,
[4]saibhavya.u@qiscet.edu.in, [5]seshaiah.t@qiscet.edu.in
Corresponding Author Mail: qispublications@qiscet.edu.in

***Abstract*—** Growing development of Web documents having led to new challenges as a result of appropriately searching RDF data. Traditional RDBMS can adapt and search for dispersed data with success. The ethics of data gathering and retrieval have changed with the creation of Hdfs and its usage of the Map reduce Model with the Hive database engine. In this paper, we provide SPARQL2Hive, a Map Reduce-based, economical SPARQL querying software that allows for ad hoc SPARQL querying parsing on enormous RDF networks. As a bridge between SPARQL and Map Reduce, SPARQL2Hive makes advantage of Hive's parser rather than translating straight from one language to another. Hadoop is used to search computers and is developed into a data warehousing platform called Hive. Given the additional virtualization, our methodology is independent of the current Hdfs version, guaranteeing compatibility with any future Hadoop platform upgrades as they will be handled at the underlying Hive level. Our approach is to use the SPARQL and Hive conceptual frameworks and provide a conversion between them by means of the ATL vocabulary. In comparison to Apache Hadoop SPARQL solutions, SPARQL2Hive is used.

Keywords— RDF, ATL, SPARQL, Semantic Web, and Base.

INTRODUCTION

Over through the last few centuries, the Future Internet has become more and more significant in both research and business. The NoSQL database used to get DRF data is Elastic search, which is a W3C-standard architecture for the Future Internet. Effective data warehousing and information retrieval methodologies for DRF data must be investigated since DRF is widely used as a metadata structure in implementations. As a retrieval interface for DRF and Ontology libraries, the W3C suggests SPARQL. It is designed to be used with statistics from the World Wide Web, enabling searches across a variety of sources independent of formatting. Due to unanticipated information sources, SPARQL queries can easily grow. The development of information retrieval has created new problems with regard to correctly searching DRF data. Common information models efficiently modify and retrieve dispersed data. With the development of Hdfs and its use of the Proposed Method with the database engine Swarm, the economics of data gathering and retrieval have changed. For sure that the job information systems, information systems engineering offers a number of product and technique approaches and concepts. Business intelligence has to figure out how to manage massive volumes of information in a limited number of forms with the emergence of big information [1]. In recent years, both in the US and abroad, Model-Driven Engineering (IDM) has been a topic of dispute among computer specialists. Both in the academic world and in business IDM has made multiple significant breakthroughs in the design of complex systems by enabling a higher conceptual emphasis than standard computing [2].

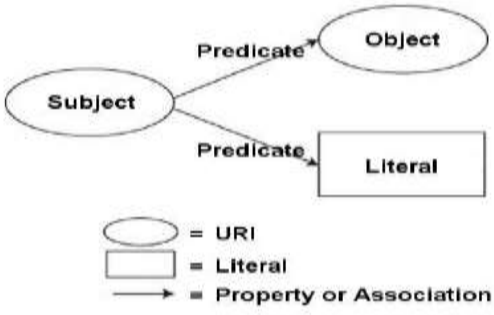
Overall arrangement of this article is as follows: A summary of the pertinent work is provided in the second section. The technologies of RDF, SPARQL, and Modeling Driven the last section talks about design tools. The explanation of the approach is covered in the fourth part. Section 5.the effectiveness of our plan concludes by presenting and analyzing the results of the tests. The results and opinions are presented in the last section.

RELATED WORKS

Implementing big data platforms like Hadoop, HDFS, Mongo DB, and others has been recommended in several earlier studies as a way to solve scaling issues with successful field implementation [7]. Hive A method for creating sizable data centers is called the Hive platform [8]. It is based on a second storage system called memstore, which connects to other NoSQL databases (mostly HBase) and an elastic search system called HiveQL, which is a query language comparable to SQL. For analytics on Big Information. Beehive, Pig, and Hadoop were built and compared by [9]. The selection is based on the type of vocabulary, graphical user, procedures, and data volume of each product. We conducted research and published it in [10] on the use of Mongo dB monitoring and managing enormous data collections utilizing the following 4 JSON concepts: Centered columnar, important directed, memo, and bar chart are common formats used in online transaction processing. The research [11] provided an overview of DRF triple stores based on Data stores as well as a brief synopsis of the information retrieval tools used for each experimental dataset. In between them is PigSPARQL [12], a technique for handling data sets. Pig Latin [13] linguistic curriculum is plotted and translated into SQL queries via PigSPARQL. Another Domain Specific method based on Big Information appears to be Jena-HBase [14], a highly scalable RDF store predicated on the NoSQL dbms HBase [15]. Data are stored in HDFS and processed using Dremel in a graph database, which is a paragraph directory. This system's architecture is divided into two sections: storing and questioning. The RDF documents are kept in HDFS databases, and SPARQL searches are made using the Maya Foundation [17].

A World Wide Web consortium standard called Background domain Reference Format (DRF) was developed to integrate computer information to Web content. A group of sequences that together make up logical information are what are referred to as DRF instances. Three elements comprise a DRF trio: Topics, conditions, and arguments that are frequently employed (S, P, O):

The subject is a website's identity, or URI (Universal Resource Identification). A property may be a piece of digital text, a portion of a homepage, a collection of online pages, or even a non-internet object that may be given a URI. The feature that defines the item is the prerequisite. So every field describes the assets it may represent, the possible interpretations for those values, and the relationships it has with other variables. The asset's price is expressed as an item, which might be either a textual integer or a URI (or resource). Figure 1 shows a triple RDF in action. The phrase "social compact created by William Daniel Thoreau" might be expressed as an RDF



triple.

Triple RDF in Figure 1.

The SPARQL Interface and RDF Linguistic Search are suggested by the W3C. (SPARQL). You may define searches on DRF databases using this language, and you can also decide how the results will be presented. So long as they are stated in DRF, it is a generic programming vocabulary for any types of data. The SPARQL system supports both the building of a new graph from a set of predetermined components via its Build stipulation in addition to the querying of an RDF network using its Search verb. This choice makes it possible to think about SPARQL as a transformational rules language for RDF-based modelling. By placing the paradigm at the center of

the construction process, Test Driven Technology [19] enables it to shift from a contemplative position to a unifying function in respect to other activities in the computer industry method. Therefore, the Break core should be seen as a method of fusing several technical fields together in order to advance toward advanced software manufacturing. Then, prototype development gives means to model, metamodel, choose a topic, convert, and design and apply ideas to aid in the creation of this new software. In addition to testing and evaluating the frames' adherence to metamodels, these methodologies are supplemented with coding tools and development techniques [20].

Proposed Approach

We now discuss the three primary components of our strategy: an origin meant to represent, the SPARQL met model, a destination met model, the Hive effectiveness, and a third basic attribute. As shown in Picture 2, the third element is the transition between these two modelling techniques, which we do out using the language ATL [18]. Each SPARQL query submitted by a client complies with our SPARQL conceptual schema, thus this request must undergo a converter using the ATL languages, which converts the SPARQL clauses into Hive clauses. This Hive programming will then also abide by our Hive conceptual.

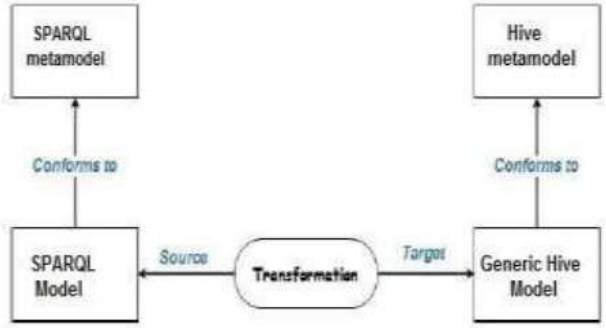
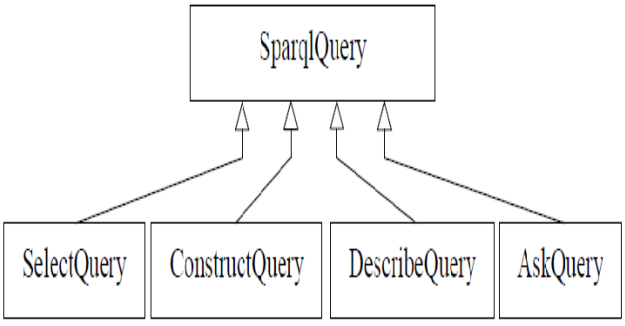


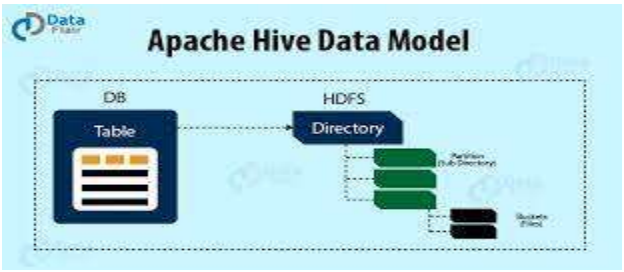
Figure 2. Proposed Approach

The true nature of the queries must be understood in order to build a SPARQL object model. This architecture is first reminiscent of SQL queries. Additionally, Choose, Build, and Request are three separate categories of SPARQL queries. The Choose inquiry will be our first choice because it is the most popular and significant. This question is a question because it may separate Data sets in accordance with the specifications listed in the within which requirement. Every SELECT statement contains the key words Pick, Source, and the conditional WHERE.



SPARQL Metamodel, Figure 3.

The "database server" in Hdfs called "Hive" offers a language for extracting "interpersonal" architecture from ambiguous or insufficient information (flat files, JSON, web logs, Hbase, Cassandra ...). In a Hadoop Sql search, the phrases With, Holding, Part On, and Sort According are all used.



using Hive Metamodel, Fig 4.

Throughout supplement to changing existing theories, ATL offers a new form of inquiry that enables users to define inquiries on the models (conversions of type arrangement towards language, or "enquiry"). One of the characteristics that makes the ATL term unique is its dual nature (unambiguous and propulsive) [22]. A component of the transformation's origin conceptual schema may be explicitly linked to a component of the transformation's destination conceptual schema in the prescriptive section. Wow also included QVT [23] as a component of the Protein carbonyl methodology (Query, Look, and Transition). A statement that takes a modelling source and selects particular components from it is referred to as an enquiry. A viewpoint is a paradigm that combines elements from various representations. An intake version is modified or a new one is constructed.

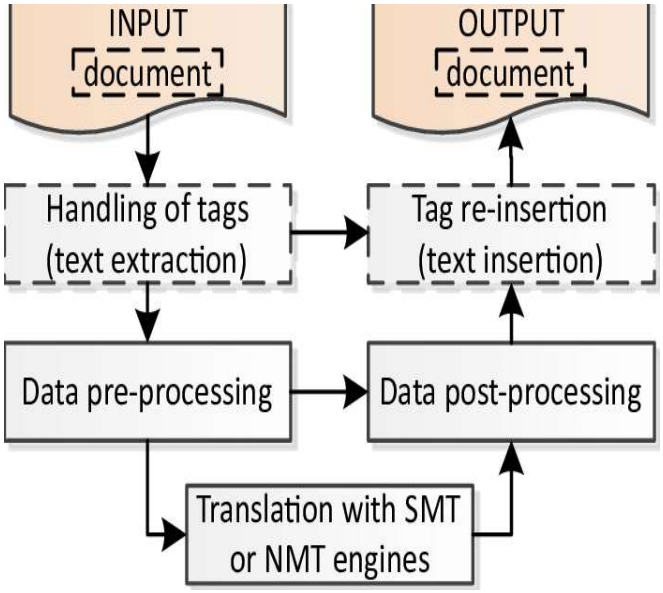


Figure 5: The Translation Machine Process

Advertisement programs include instructions that specify how the parts of the destination figure should be put together based on the parts of the parent figure. These criteria are always established while adhering to the

paradigm. The schematic illustration illustrates how this tweak will affect things (Fig.6)

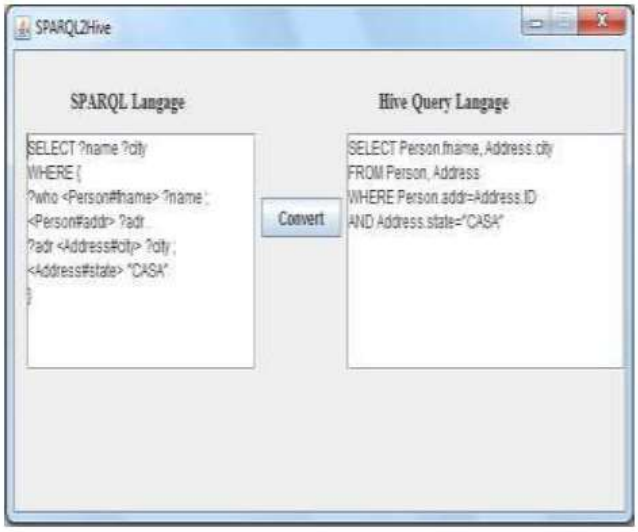
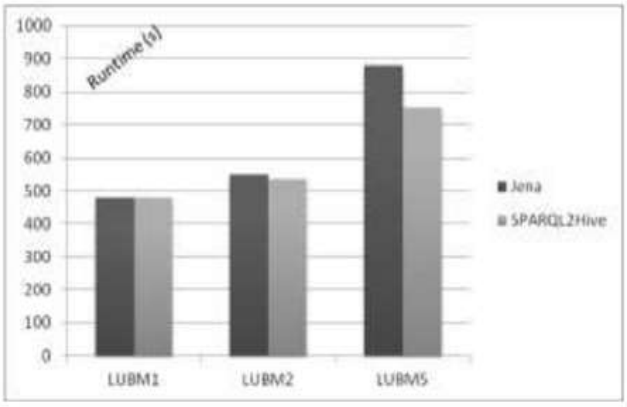


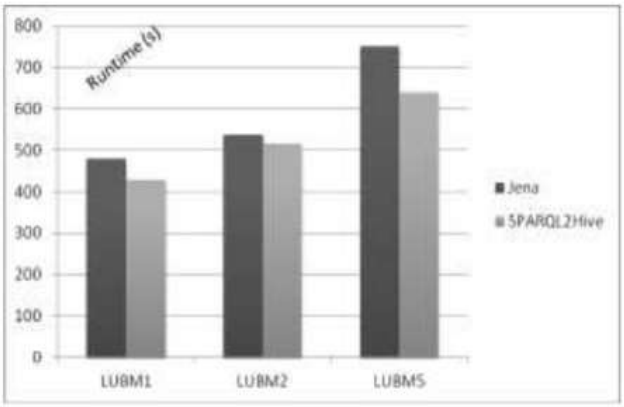
Fig. 6 shows an example of converting a SPARQL query into a Hive application.

RESULTS AND DISCUSSION

The seven LUBM searches are carried out on these three databases of different sizes in order to better grasp our SPARQL2Hive technology. In the first section, specifics about the databases, the edition of Hive, and the settings and backgrounds of our interactions are all described. After then, the results will be analyzed. Additionally, we will evaluate how the sampling data size affects the performance of the data modelling, and we will explain the results of this evaluation in a way that can be visually represented and shows the effectiveness of our SPARQL2Hive technology. Hdfs 3.xy and Hive 3.1.0 are used to execute SPARQL2Hive on a 2.3 GHz CPU-powered machine with up to 4 Terabytes of solid-state storage and 16 Gigabytes of RAM. All three characteristics, designated as LUBM1, LUBM2, and LUBM5, were used in this analysis and had the following sextuplet's numbers: These two databases are 8.5 gigabytes, 22 gigabytes, 78 gigabytes, and 56 gigabytes in size, with 139 billion references, 277 billion triples, and 488 billion triples, respectively. The findings for the constraints of these 3 matches are summarized in table below. By using the three databases LUBMI, LUBM2, and LUBM5, we conduct LUBM Benchmarking searches to assess our SPARQL to hive technology to Maya. When that comes to LUBM Benchmark performance, SPARQL2hive surpasses Maya by a wide margin. Figure 7 displays a comparison of the results for each LUBM search. SPARQL2Hive is a modular, robust, and fault-tolerant platform, according to the aforementioned conclusions. These results show that SPARQL2Hive is useful for processing huge volumes of DRF data [25]. The information is inserted quickly using SPARQL2Hive. Due to the simplicity with which a SQL query may be easily converted into a HiveQL application. In contrast, the Jasper architecture's operation is a little trickier since the demand must go through a number of lengthy procedures, particularly when uploading preparation and analysis data for restoration, and Jasper uses a lot of resources, such Memory. Datasets loading time

Dataset	***LUBM1***	***LUBM2***	***LUBM5***
Loading time in milliseconds	1.18	2.95	2.7





Runtime using LUBM1 (Fig. 7)

Fig 8: Runtime using LUBM2

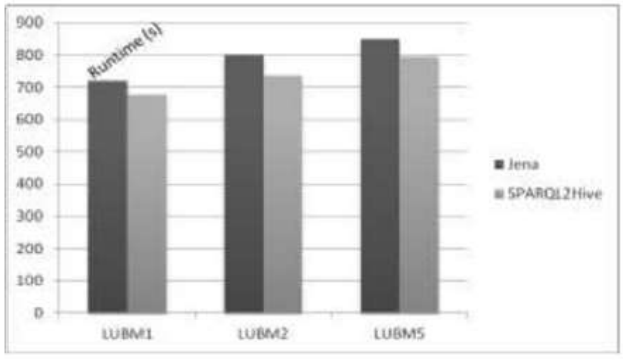


Fig 9:Runtime using LUBM5

CONCLUSION

academic institutions now face a new challenge in maintaining such a large volume of drf data due to the growing volume of drf data. the focus of the study is on large-scale digital retention technologies like hbase [14], and administration is handled via querying technologies like hive. in this paper, we introduce sparql2hive, a novel structure for turning sparql searches into hiveql programmes that is based on a conceptual approach.

References

[1] H. Thakkar, R. Angles, M. Rodriguez, S. Mallette and J. Lehmann, "Let's build Bridges, not Walls: SPARQL Querying of TinkerPop Graph Databases with Sparql-Gremlin," 2020 IEEE 14th International Conference on Semantic Computing (ICSC), 2020, pp. 408-415, doi: 10.1109/ICSC.2020.00080.

[2] M. Atzori, "Computing Recursive SPARQL Queries," 2014 IEEE International Conference on Semantic Computing, 2014, pp. 258-259, doi: 10.1109/ICSC.2014.54.

[3] K. M. Kyu and A. N. Oo, "Enhancement of Query Execution Time in SPARQL Query Processing," 2020 International Conference on Advanced Information Technologies (ICAIT), 2020, pp. 153-158, doi: 10.1109/ICAIT51105.2020.9261805.

[4] S. Malik, A. Goel and S. Maniktala, "A comparative study of various variants of SPARQL in semantic web," 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM), 2010, pp. 471-474, doi: 10.1109/CISIM.2010.5643493.

[5] M. Banane, A. Erraissi and A. Belangour, "SPARQL2Hive: An approach to processing SPARQL queries on Hive based on meta-models," 2019 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO), 2019, pp. 1-5, doi: 10.1109/ICMSAO.2019.8880393.

[6] W. Li et al., "SHOE: A SPARQL Query Engine Using MapReduce," 2013 International Conference on Parallel and Distributed Systems, 2013, pp. 446-447, doi: 10.1109/ICPADS.2013.78.

[7] X. Zhang and J. Van den Bussche, "On the Power of SPARQL in Expressing Navigational Queries," in The Computer Journal, vol. 58, no. 11, pp. 2841-2851, Nov. 2015, doi: 10.1093/comjnl/bxu128.

[8] F. Bamashmoos, I. Holyer, T. Tryfonas and P. Woznowski, "Towards Secure SPARQL Queries in Semantic Web Applications Using PHP," 2017 IEEE 11th International Conference on Semantic Computing (ICSC), 2017, pp. 276-277, doi: 10.1109/ICSC.2017.29.

[9] L. Lv, H. Jiang and L. Ju, "Research and Implementation of the SPARQL-TO-SQL Query Translation Based on Restrict RDF View," 2010 International Conference on Web Information Systems and Mining, 2010, pp. 309-313, doi: 10.1109/WISM.2010.44.

[10] H. Wang, Z. M. Ma and J. Cheng, "fp-Sparql: An RDF fuzzy retrieval mechanism supporting user preference," 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, 2012, pp. 443-447, doi: 10.1109/FSKD.2012.6234114.

[11] X. Chen, T. Wu, Q. Xie and J. He, "Data Flow-Oriented Multi-Path Semantic Web Service Composition Using Extended SPARQL," 2017 IEEE International Conference on Web Services (ICWS), 2017, pp. 882-885, doi: 10.1109/ICWS.2017.112.

[12] J. M. Almendros-Jiménez, A. Becerra-Terón and G. Moreno, "A fuzzy extension of SPARQL based on fuzzy sets and aggregators," 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2017, pp. 1-6, doi: 10.1109/FUZZ-IEEE.2017.8015411.

[13] R. Gupta and S. K. Malik, "SPARQL Semantics and Execution Analysis in Semantic Web Using Various Tools," 2011 International Conference on Communication Systems and Network Technologies, 2011, pp. 278-282, doi: 10.1109/CSNT.2011.67.

[14] H. Jabeen, E. Haziiev, G. Sejdiu and J. Lehmann, "DISE: A Distributed in-Memory SPARQL Processing Engine over Tensor Data," 2020 IEEE 14th International Conference on Semantic Computing (ICSC), 2020, pp. 400-407, doi: 10.1109/ICSC.2020.00079.

[15] D. A. C. Amat, C. Buil-Aranda and C. Valle-Vidal, "A Neural Networks Approach to SPARQL Query Performance Prediction," 2021 XLVII Latin American Computing Conference (CLEI), 2021, pp. 1-9, doi: 10.1109/CLEI53233.2021.9639899.

[16] D. S. Ferru and M. Atzori, "Write-Once Run-Anywhere Custom SPARQL Functions," 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), 2016, pp. 176-178, doi: 10.1109/ICSC.2016.64.

[17] A. Kashlev and A. Chebotko, "SPARQL-to-SQL Query Translation: Bottom-Up or Top-Down?," 2011 IEEE International Conference on Services Computing, 2011, pp. 757-758, doi: 10.1109/SCC.2011.79.

[18] X. Zhai, L. Huang and Z. Xiao, "Geo-spatial query based on extended SPARQL," 2010 18th International Conference on Geoinformatics, 2010, pp. 1-4, doi: 10.1109/GEOINFORMATICS.2010.5567605.

[19] M. Krommyda and V. Kantere, "Understanding SPARQL Endpoints through Targeted Exploration and Visualization," 2019 First International Conference on Graph Computing (GC), 2019, pp. 21-28, doi: 10.1109/GC46384.2019.00012.

[20] J. M. Almendros-Jiménez, A. Becerra-Terón, G. Moreno and J. A. Riaza, "Tuning Fuzzy SPARQL Queries in a Fuzzy Logic Programming Environment," 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2019, pp. 1-7, doi: 10.1109/FUZZ-IEEE.2019.8858958.

[21] K. M. Nguyen, T. -H. Nguyen and X. H. Huynh, "Automated translation between RESTful/JSON and SPARQL messages for accessing semantic data," 2016 International Conference on Electronics, Information, and Communications (ICEIC), 2016, pp. 1-4, doi: 10.1109/ELINFOCOM.2016.7562981.

[22] J. Euzenat, A. Polleres and F. Scharffe, "Processing Ontology Alignments with SPARQL," 2008 International Conference on Complex, Intelligent and Software Intensive Systems, 2008, pp. 913-917, doi: 10.1109/CISIS.2008.126.

[23] S. Chun, S. Seo, W. Ro and K. -H. Lee, "Proactive Plan-Based Continuous Query Processing over Diverse SPARQL Endpoints," 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015, pp. 161-164, doi: 10.1109/WI-IAT.2015.168.

[24] T. Chawla, G. Singh and E. S. Pilli, "A shortest path approach to SPARQL chain query optimisation," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp. 1778-1778, doi: 10.1109/ICACCI.2017.8126102.

[25] N. Kumar and S. Kumar, "Querying RDF and OWL data source using SPARQL," 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 2013, pp. 1-6, doi: 10.1109/ICCCNT.2013.6726698.

[26] P Ramprakash, M Sakthivadivel, N Krishnaraj, J Ramprasath. "Host-based Intrusion Detection System using Sequence of System Calls" International Journal of Engineering and Management Research, Vandana Publications, Volume 4, Issue 2, 241-247, 2014

[27] N Krishnaraj, S Smys."A multihoming ACO-MDV routing for maximum power efficiency in an IoT environment" Wireless Personal Communications 109 (1), 243-256, 2019.

[28] N Krishnaraj, R Bhuvanesh Kumar, D Rajeshwar, T Sanjay Kumar, Implementation of energy aware modified distance vector routing protocol for energy efficiency in wireless sensor networks, 2020 International Conference on Inventive Computation Technologies (ICICT),201-204

[29] Ibrahim, S. Jafar Ali, and M. Thangamani. "Enhanced singular value decomposition for prediction of drugs and diseases with hepatocellular carcinoma based on multi-source bat algorithm based random walk." Measurement 141 (2019): 176-183. https://doi.org/10.1016/j.measurement.2019.02.056

[30] Ibrahim, Jafar Ali S., S. Rajasekar, Varsha, M. Karunakaran, K. Kasirajan, Kalyan NS Chakravarthy, V. Kumar, and K. J. Kaur. "Recent advances in performance and effect of Zr doping with ZnO thin film sensor in ammonia vapour sensing." GLOBAL NEST JOURNAL 23, no. 4 (2021): 526-531. https://doi.org/10.30955/gnj.004020 , https://journal.gnest.org/publication/gnest_04020

[31] N.S. Kalyan Chakravarthy, B. Karthikeyan, K. Alhaf Malik, D.Bujji Babbu,. K. Nithya S.Jafar Ali Ibrahim , Survey of Cooperative Routing Algorithms in Wireless Sensor Networks, Journal of Annals of the Romanian Society for Cell Biology ,5316-5320, 2021

[32] Rajmohan, G, Chinnappan, CV, John William, AD, Chandrakrishan Balakrishnan, S, Anand Muthu, B, Manogaran, G. Revamping land coverage analysis using aerial satellite image mapping. Trans Emerging Tel Tech. 2021; 32:e3927. https://doi.org/10.1002/ett.3927

[33] Vignesh, C.C., Sivaparthipan, C.B., Daniel, J.A. et al. Adjacent Node based Energetic Association Factor Routing Protocol in Wireless Sensor Networks. Wireless Pers Commun 119, 3255–3270 (2021). https://doi.org/10.1007/s11277-021-08397-0.

[34] 9. C Chandru Vignesh, S Karthik, Predicting the position of adjacent nodes with QoS in mobile ad hoc networks, Journal of Multimedia Tools and Applications, Springer US,Vol 79, 8445-8457,2020