JOIN QUERY ENHANCEMENT PROCESSING (JQPro) WITH BIG RDF DATA ON A DISTRIBUTED SYSTEM USING HASHING-MERGE JOIN TECHNIQUE

NAHLA MOHAMMEDELZEIN ELAWAD BABIKER

DOCTOR OF PHILOSOPHY

UNIVERSITI MALAYSIA PAHANG

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis, and, in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Doctor of Philosophy.

_____

(Supervisor's Signature)

Full Name : TS. DR. MAZLINA BINTI ABDUL MAJID

Position : ASSOCIATE PROFESSOR

Date : 05 AUGUST 2021

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.

(Student's Signature)

Full Name : NAHLA MOHAMMEDELZEIN ELAWAD BABIKER

ID Number : PCC15016

Date : 05 AUGUST 2021

JOIN QUERY ENHANCEMENT PROCESSING (JQPro) WITH BIG RDF DATA ON A DISTRIBUTED SYSTEM USING HASHING-MERGE JOIN TECHNIQUE

NAHLA MOHAMMEDELZEIN ELAWAD BABIKER

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Doctor of Philosophy

Faculty of Computing
UNIVERSITI MALAYSIA PAHANG

AUGUST 2021

ACKNOWLEDGEMENTS

My thanks go to the Almighty God, my life's giver, protector, and support. Without my friendly supervisor, **Associate Professor Ts. Dr. Mazlina Abdul Majid**, I regard this work as a great privilege and honour of working under your tutelage, it would not have been a successful task. I am also very grateful for their support in various capacities to all lecturers, consulters and course colleagues.

I would like to thank my **dad** and **Mom**, my **husband**, my **siblings** for their unimaginable sacrifices and prayers in this research. In the course of my rigorous work, I greatly thank you for your patience and love, and I spiritually support all of you in the writing of this thesis and my whole life. This research could not be carried out without precious support.

My thanks go to my friends, most especially, **Dr. Mohammed Fakherldin** and **Dr. Ibrahim Hashim,** for their encouragement during those periods of great challenge. And all those **friends** who believe in my competencies and have encouraged me tirelessly or boredomlessly.

Finally, everyone who has helped to make this work a success, I extend my respect and blessing. My prayer is that when you need it the most, you will never lack help. Thank you all and bless God.

Thank you very much for your encouragement!

ABSTRAK

Teknologi web semantik telah muncul beberapa tahun lalu merentasi pelbagai bidang kajian dan datanya bertambah dengan cepat. Secara khususnya, keupayaan penyimpanan dan penerbitan data di dalam format web telah menjadikan teknologi ini semakin maju. Jadi, data tersebut boleh dibaca oleh manusia dan dapat diproses oleh komputer. Keperluan terhadap pertanyaan *Resources Description Framework* (RDF) berganda yang kompleks menjadi penting dengan peningkatan RDF tiga kali ganda. Pertanyaan yang kompleks sebegitu kadangkala menghasilkan banyak subekspresi umum. Walaupun begitu, adalah sangat mencabar untuk mengurangkan jumlah pertanyaan RDF dan masa penghantaran bagi sebilangan besar data berkaitan RDF. Selain itu, kajian literatur terkini menunjukkan pemprosesan pertanyaan terhubung untuk RDF yang besar telah mengundang banyak masalah berkenaan masa perlaksanaan dan daya pemprosesan. Pengekodan berdasarkan hash menghasilkan masa perlaksanaan yang perlahan dan memerlukan lebih masa untuk memuatkan serta tidak mampu memuatkan kesemua graf. Hal ini kerana, RDF mengumpul dan menganalisa data yang besar di dalam kelompok, seterusnya perlu menghadapi cabaran yang wujud tentang penyimpanan data berkelompok yang efektif. Penyimpanan dan capaian data yang efektif, yang dilaksanakan ke atas sebilangan besar data tanpa skema, telah dibuktikan sangat sukar untuk penyimpanan data RDF. Sebagai contoh, adalah sukar untuk mempamerkan bahasa pertanyaan semantik dan (SPARQL) serta pola graf yang besar dan kompleks. Bagi mengatasi masalah ini, Join Query Processing Model (JQPro) telah diperkenalkan untuk data RDF yang besar. Antara objektif kajian ini adalah: (i) merumuskan rancangan algoritma penjana untuk pemprosesan pertanyaan terhubung berdasarkan kajian yang lepas, (ii) membangunkan penambahbaikan terhadap model Join Query Processing (JQPro) berdasarkan SPARQL dan Hadoop MapReduce menggunakan teknik terhubung *hashing-merge* bagi memproses data RDF yang besar dan (iii) menilai dan membandingkan prestasi masa perlaksanaan, daya pemprosesan dan penggunaan *Central Process Unit* (CPU) bagi model JQPro dengan model-model sedia ada. Masa perlaksanaan digunakan untuk mengukur masa dari permulaan kerja sehingga masa tindak balas. Selain itu, daya pemprosesan digunakan untuk mengukur unit maklumat yang boleh diproses oleh sistem di dalam setiap tempoh masa. Tambahan lagi, CPU digunakan sebagai elemen penting di dalam pemprosesan pertanyaan terhubung yang besar terutamanya semasa pemetaan bagi mengurangkan fasa-fasa. Selain itu, algoritma hash-join dan sort-merge digunakan untuk menghasilkan pemprosesan pertanyaan terhubung, dan ini adalah kerana keupayaan mereka membenarkan lebih banyak set data untuk dihubungkan. Kedua-dua proses disisih mengikut algoritma attribut terhubung dan kaitan yang digabungkan. Oleh itu, lajur terhubung mengasingkan kumpulan-kumpulan set data dengan nilai yang sama. Algoritma terhubung sort-merge mengasingkan set data pada attribut terhubung dan mencari tupel dengan menggabungkan dua set data. Seterusnya, satu kerangka pemprosesan untuk pertanyaan RDF telah diperkenalkan dan penanda aras digunakan untuk penilaian prestasi. Akhir sekali, pengesahan dilakukan dengan melakukan analisis statistik piawaian untuk mengesahkan dan membandingkan prestasi model JQPro dengan model sedia ada. Tambahan lagi, penanda aras tiruan (LUBM) dan (WatDiv) v06 digunakan sebagai pengukuran. Hasil kajian menunjukkan terdapat kaitan yang kuat antara jangka masa pelaksanaan dan daya perlaksanaan dengan kekuatan 99.9% seperti yang telah disahkan oleh pekali perkaitan Pearson. Seterusnya, hasil kajian menunjukkan bahawa penyelesaian JQPro adalah setanding dengan gStore RDF-3X, RDFox dan PARJ dan peratusan peningkatan prestasi masa pelaksanaan adalah sebanyak 87.77%. Penggunaan CPU sangat ketara meningkat dengan pemetaan yang luas dan pengurangan kod pengkomputeran. Hal ini menyimpulkan bahawa penyelesaian JQPro adalah tepat pada masanya dan inovatif, berikutan masa pelaksanaan dan penggunaan CPU yang efektif di mana pengguna dapat melaksanakan dengan sempurna pertanyaan yang lebih baik untuk pemprosesan data RDF yang besar.

ABSTRACT

Semantic web technologies have emerged in the last few years across different fields of study and their data are still growing rapidly. Specifically, the increased data storage and publishing capabilities in standard open web formats have made the technology much more successful. So, the data have become readable by humans, and they can be processed on a computer. The demand for complex multiple RDF queries is becoming significant with the increasing number of RDF triples. Such complex queries occasionally produce many common subexpressions. It is therefore extremely challenging to reduce the amount of RDF queries and transmission time for a vast number of related RDF data. Moreover, Recent literature shows that join query processing of Big RDF data has introduced many problems with respect to execution time and throughput. The hash-based encoding induces low execution time, which takes a long time to load and hence does not load all graphs. This is because the Resource Description Framework (RDF) collects and analyses large data in swarms, thereby having to deal with the inherent challenge of efficient swarm storage. The effective storage and data retrieval, which could be applied to high amounts of possible schema-less data, has also proven exceedingly difficult for RDF data storage. For instance, it is particularly difficult to view semantic and SPARQL query languages, as well as huge and complex graph patterns. To address this problem, a Join Query Processing Model (JQPro) is introduced for Big RDF data. The objectives of this research are: (i) formulate plan generator algorithms for join query processing on the basis of the previous research. (ii) develop an enhancement model of Join Query Processing (JQPro) based on SPARQL and Hadoop MapReduce using hashing-merge join technique to process Big RDF Data. (iii) evaluate and compare the performance based on the execution time, throughput, and CPU utilization of the JQPro model with existing models. On the other hand, the throughput was employed to measure the units of information that a system can process in each time frame. In addition, the CPU utilization was used in the big join query processing as an important resource element particularly during the map, to reduce phases. Furthermore, the hash-join and Sort-Merge algorithms were used to generate the join query processing, and this was employed due to their capacity to allow for more data sets to be joined. Both processes were sorted by algorithms on join attributes and the sorted relations was merged. Therefore, the join column sorted the groups of datasets with the same value. The sort–merge–join algorithm sorts the datasets on the joining attribute and then searches for tuples by merging the two datasets. Then, a processing framework for RDF queries was introduced and the benchmark was used for performance evaluation. Finally, the validation was conducted by standard statistical analysis to validate and compare the performance of the JQPro model with current models. In addition, the synthetic benchmarks Lehigh University Benchmark (LUBM) and Waterloo SPARQL Diversity Test Suite (WatDiv) v06 were used for measurement. The experiment was carried out on three datasets ranging from 10 million to 1 billion RDF triples produced by the generator of WatDiv data with a scale factor of 10, 100 and 1000, respectively. A selective dataset for each experimental query was also used for the processing of RDFs with a LUBM benchmark in sizes 500, 1000 and 2000 million triples. The result revealed that there is a strong correlation between execution time and throughput with a strength of 99.9% percent as confirmed by the Pearson correlation coefficient. Furthermore, the findings show that the JQPro solution was comparable to gStore RDF-3X, RDFox and PARJ and the percentage of improved performance was 87.77% in terms of execution time. The CPU utilization was significantly increased by extensive mapping and reduced code computing. It is therefore inferred that the JQPro solution is timely and innovative, as it provides an efficient execution time and CPU utilization where users could perform better queries for Big RDF data processing in a seamless manner.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

REFERENCES

Abadi, D. J., Marcus, A., Madden, S. R., & Hollenbach, K. (2009). SW-Store: a vertically partitioned DBMS for Semantic Web data management. *The VLDB Journal, 18*(2), 385-406.

Abdelaziz, I., Harbi, R., Khayyat, Z., & Kalnis, P. (2017). A survey and experimental comparison of distributed SPARQL engines for very large RDF data. *Proceedings of the VLDB Endowment, 10*(13), 2049-2060.

Aggarwal, C. C. (2011). An introduction to social network data analytics. In *Social network data analytics* (pp. 1-15): Springer.

Ali, W., Saleem, M., Yao, B., Hogan, A., & Ngomo, A.-C. N. (2020). Storage, indexing, query processing, and benchmarking in centralized and distributed rdf engines: A survey. *arXiv preprint arXiv:2009.10331*.

Amann, B., Curé, O., & Naacke, H. (2018). Distributed SPARQL Query Processing: a Case Study with Apache Spark. *NoSQL Data Models: Trends and Challenges, 1*, 21-55.

Antoniou, G., & Van Harmelen, F. (2004). *A semantic web primer*: MIT press.

Aranda-Andújar, A., Bugiotti, F., Camacho-Rodríguez, J., Colazzo, D., Goasdoué, F., Kaoudi, Z., & Manolescu, I. (2012). *AMADA: web data repositories in the amazon cloud.* Paper presented at the Proceedings of the 21st ACM international conference on Information and knowledge management.

Arfat, Y., Usman, S., Mehmood, R., & Katib, I. (2020). Big Data Tools, Technologies, and Applications: A Survey. In *Smart Infrastructure and Applications* (pp. 453-490): Springer.

Ashraf, J., Hussain, O. K., & Hussain, F. K. (2015). Making sense from big RDF data: OUSAF for measuring ontology usage. *Software: Practice and Experience, 45*(8), 1051-1071.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web* (pp. 722-735): Springer.

Bakhshi, M., Nematbakhsh, M., Mohsenzadeh, M., & Rahmani, A. M. (2020). Data-driven construction of SPARQL queries by approximate question graph alignment in question answering over knowledge graphs. *Expert Systems with Applications, 146*, 113205.

Bauer, F., & Kaltenböck, M. (2011). Linked open data: The essentials. *Edition mono/monochrom, Vienna, 710*.

Beckett, D. (2014). Rdf 1.1 n-triples. *URL: https://www. w3. org/TR/n-triples*.

Berman, J. J. (2013). *Principles of big data: preparing, sharing, and analyzing complex information*: Newnes.

Bermudez-Edo, M., Elsaleh, T., Barnaghi, P., & Taylor, K. (2017). IoT-Lite: a lightweight semantic model for the internet of things and its use with dynamic semantics. *Personal and Ubiquitous Computing, 21*(3), 475-487.

Berners-Lee, T. (2000). Semantic Web. Retrieved from http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html

Berners-Lee, T. (2000). The semantic web stack from a 2000 presentation. In.

Berners-Lee, T. (2006). Linked data-design issues. *http://www. w3. org/DesignIssues/LinkedData. html*.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific american, 284*(5), 34-43.

Bilidas, D., & Koubarakis, M. (2019). *Scalable Parallelization of RDF Joins on Multicore Architectures.* Paper presented at the EDBT.

Bishop, B., Kiryakov, A., Ognyanoff, D., Peikov, I., Tashev, Z., & Velkov, R. (2011). OWLIM: A family of scalable semantic repositories. *Semantic Web, 2*(1), 33-42.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far.. sl: sn.

Blin, G., Curé, O., & Faye, D. C. (2012). A survey of RDF storage approaches. *Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées, 15*.

Bloom, B. H. (1970). Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM, 13*(7), 422-426.

Bohlouli, M., Schulz, F., Angelis, L., Pahor, D., Brandic, I., Atlan, D., & Tate, R. (2013). Towards an integrated platform for big data analysis. In *Integration of practice-oriented knowledge technology: Trends and prospectives* (pp. 47-56): Springer.

Brisaboa, N. R., Cerdeira-Pena, A., de Bernardo, G., Fariña, A., & Navarro, G. (2020). Space/time-efficient RDF stores based on circular suffix sorting. *arXiv preprint arXiv:2009.10045*.

Broekstra, J., Kampman, A., & Van Harmelen, F. (2002). *Sesame: A generic architecture for storing and querying rdf and rdf schema.* Paper presented at the International semantic web conference.

Cabellos, L., Campos, I., Fernández-del-Castillo, E., Owsiak, M., Palak, B., & Płóciennik, M. (2011). Scientific workflow orchestration interoperating HTC and HPC resources. *Computer Physics Communications, 182*(4), 890-897.

Cai, H., & Vasilakos, A. V. (2017). Web of things data storage. In *Managing the Web of Things* (pp. 325-354): Elsevier.

Chansler, R. J. (2012). Data availability and durability with the hadoop distributed file system. *The USENIX Magzine, 37*(1).

Chantrapornchai, C., & Choksuchat, C. (2018). TripleID-Q: RDF query processing framework using GPU. *IEEE Transactions on Parallel and Distributed Systems, 29*(9), 2121-2135.

Chawla, T., Singh, G., & Pilli, E. S. (2018). *JOTR: Join-Optimistic Triple Reordering Approach for SPARQL Query Optimization on Big RDF Data.* Paper presented at the 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT).

Che, D., Safran, M., & Peng, Z. (2013). *From big data to big data mining: challenges, issues, and opportunities.* Paper presented at the International conference on database systems for advanced applications.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly, 36*(4).

Cheng, L., & Kotoulas, S. (2015). Scale-out processing of large RDF datasets. *IEEE Transactions on Big Data, 1*(4), 138-150.

Chiarcos, C., McCrae, J., Cimiano, P., & Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources* (pp. 7-25): Springer.

Choi, P., Jung, J., & Lee, K.-H. (2013). *RDFChain: Chain Centric Storage for Scalable Join Processing of RDF Graphs using MapReduce and HBase.* Paper presented at the International Semantic Web Conference (Posters & Demos).

Consortium, T. G. O. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics, 25*(1), 25-29.

Consortium, W. C. W. W. W. (2008). SPARQL query language for RDF. *Online: http://www. w3. org/TR/rdf-sparql-query*.

Consortium, W. W. W. (2014a). Best practices for publishing linked data.

Consortium, W. W. W. (2014b). RDF Schema 1.1 W3C Recommendation 25 February 2014.

Cossu, M., Färber, M., & Lausen, G. (2018). PRoST: distributed execution of SPARQL queries using mixed partitioning strategies. *arXiv preprint arXiv:1802.05898*.

Cumbley, R., & Church, P. (2013). Is "big data" creepy? *Computer Law & Security Review, 29*(5), 601-609.

De Virgilio, R., Del Nostro, P., Gianforme, G., & Paolozzi, S. (2011). A scalable and extensible framework for query answering over RDF. *World Wide Web, 14*(5-6), 599-622.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107-113.

Dendek, P. J., Czeczko, A., Fedoryszak, M., Kawa, A., Wendykier, P., & Bolikowski, L. (2013). How to perform research in Hadoop environment not losing mental equilibrium-case study. *arXiv preprint arXiv:1303.5234*.

Ekanayake, J., Li, H., Zhang, B., Gunarathne, T., Bae, S.-H., Qiu, J., & Fox, G. (2010). *Twister: a runtime for iterative mapreduce.* Paper presented at the Proceedings of the 19th ACM international symposium on high performance distributed computing.

Escriva, R., Wong, B., & Sirer, E. G. (2012). *HyperDex: A distributed, searchable key-value store.* Paper presented at the Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication.

Esteves, D., Rula, A., Reddy, A. J., & Lehmann, J. (2018). Toward veracity assessment in RDF knowledge bases: An exploratory analysis. *Journal of Data and Information Quality (JDIQ), 9*(3), 1-26.

Feng, J., Meng, C., Song, J., Zhang, X., Feng, Z., & Zou, L. (2017). *SPARQL query parallel processing: a survey.* Paper presented at the 2017 IEEE International Congress on Big Data (BigData Congress).

Fernández, A., del Río, S., López, V., Bawakid, A., del Jesus, M. J., Benítez, J. M., & Herrera, F. (2014). Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4*(5), 380-409.

Fox, E. A., Heath, L. S., Chen, Q. F., & Daoud, A. M. (1992). Practical minimal perfect hash functions for large databases. *Communications of the ACM, 35*(1), 105-121.

Freitas, A., Curry, E., Oliveira, J. G., & O'Riain, S. (2011). Querying heterogeneous datasets on the linked data web: challenges, approaches, and trends. *IEEE Internet Computing, 16*(1), 24-33.

Gai, L., Wang, X., & Wang, T. (2018). *ROSIE: Runtime Optimization of SPARQL Queries over RDF Using Incremental Evaluation.* Paper presented at the International Conference on Knowledge Science, Engineering and Management.

Galárraga, L., Hose, K., & Schenkel, R. (2014). *Partout: a distributed engine for efficient RDF processing.* Paper presented at the Proceedings of the 23rd International Conference on World Wide Web.

Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003). The Google file system.

Gillick, D., Faria, A., & DeNero, J. (2006). Mapreduce: Distributed computing for machine learning. *Berkley, Dec, 18*.

Goranko, V., Kyrilov, A., & Shkatov, D. (2010). Tableau tool for testing satisfiability in LTL: Implementation and experimental analysis. *Electronic Notes in Theoretical Computer Science, 262*, 113-125.

Gorton, I., Greenfield, P., Szalay, A., & Williams, R. (2008). Data-intensive computing in the 21st century. *Computer, 41*(4), 30-32.

Gracia, J., Villegas, M., Gomez-Perez, A., & Bel, N. (2018). The apertium bilingual dictionaries on the web of data. *Semantic Web, 9*(2), 231-240.

Graux, D. (2016). *On the efficient distributed evaluation of SPARQL queries.* Université Grenoble Alpes, Retrieved from https://tel.archives-ouvertes.fr/tel-01618366 (2016GREAM058)

Groppe, S. (2011). *Data management and query processing in semantic web databases*: Springer Science & Business Media.

Grover, P., & Johari, R. (2015). *BCD: BigData, cloud computing and distributed computing.* Paper presented at the 2015 Global Conference on Communication Technologies (GCCT).

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition, 5*(2), 199-221.

Guo, X., Gao, H., & Zou, Z. (2019). *Leon: A Distributed RDF Engine for Multi-query Processing.* Paper presented at the International Conference on Database Systems for Advanced Applications.

Guo, Y., Pan, Z., & Heflin, J. (2005). LUBM: A benchmark for OWL knowledge base systems. *Journal of Web Semantics, 3*(2-3), 158-182.

Gurajada, S., Seufert, S., Miliaraki, I., & Theobald, M. (2014). *TriAD: a distributed shared-nothing RDF engine based on asynchronous message passing.* Paper presented at the Proceedings of the 2014 ACM SIGMOD international conference on Management of data.

Hammoud, M., Rabbou, D. A., Nouri, R., Beheshti, S.-M.-R., & Sakr, S. (2015). DREAM: distributed RDF engine with adaptive query planner and minimal communication. *Proceedings of the VLDB Endowment, 8*(6), 654-665.

Han, J., Haihong, E., Le, G., & Du, J. (2011). *Survey on NoSQL database.* Paper presented at the 2011 6th international conference on pervasive computing and applications.

Harter, T., Borthakur, D., Dong, S., Aiyer, A., Tang, L., Arpaci-Dusseau, A. C., & Arpaci-Dusseau, R. H. (2014). *Analysis of {HDFS} Under HBase: A Facebook Messages Case Study.* Paper presented at the Proceedings of the 12th {USENIX} Conference on File and Storage Technologies ({FAST} 14).

Hasan, A., Hammoud, M., Nouri, R., & Sakr, S. (2016). *DREAM in action: A distributed and adaptive RDF system on the cloud.* Paper presented at the

Proceedings of the 25th International Conference Companion on World Wide Web.

Hernández-Illera, A., Martínez-Prieto, M. A., & Fernández, J. D. (2020). RDF-TR: Exploiting structural redundancies to boost RDF compression. *Information Sciences, 508*, 234-259.

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence, 194*, 28-61.

Hogenboom, A., Frasincar, F., & Kaymak, U. (2013). Ant colony optimization for RDF chain queries for decision support. *Expert Systems with Applications, 40*(5), 1555-1563.

Horrocks, I. (2002). DAML+OIL: A Description Logic for the Semantic Web. *IEEE Data Eng. Bull., 25*(1), 4-9.

Huang, J., Abadi, D. J., & Ren, K. (2011). Scalable SPARQL querying of large RDF graphs. *Proceedings of the VLDB Endowment, 4*(11), 1123-1134.

Husain, M., Doshi, P., Khan, L., & McGlothlin, J. (2009). Efficient query processing for large rdf graphs using hadoop and mapreduce. In *Technical report*: University of Texas Dallas.

Husain, M., McGlothlin, J., Masud, M. M., Khan, L., & Thuraisingham, B. M. (2011). Heuristics-based query processing for large RDF graphs using cloud computing. *IEEE Transactions on Knowledge and Data Engineering, 23*(9), 1312-1327.

Isard, M., Budiu, M., Yu, Y., Birrell, A., & Fetterly, D. (2007). *Dryad: distributed data-parallel programs from sequential building blocks.* Paper presented at the ACM SIGOPS operating systems review.

Islam, M., Huang, A. K., Battisha, M., Chiang, M., Srinivasan, S., Peters, C., . . . Abdelnur, A. (2012). *Oozie: towards a scalable workflow management system for hadoop.* Paper presented at the Proceedings of the 1st ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies.

Islam, M., & Reza, S. (2019). The Rise of Big Data and Cloud Computing. *Internet of Things and Cloud Computing, 7*(2), 45.

Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: a technological perspective and review. *Journal of Big Data, 3*(1), 25.

Jarrar, M., & Dikaiakos, M. D. (2011). A query formulation language for the data web. *IEEE Transactions on Knowledge and Data Engineering, 24*(5), 783-798.

Jung, I.-Y., Kim, K.-H., Han, B.-J., & Jeong, C.-S. (2014). Hadoop-based distributed sensor node management system. *International Journal of Distributed Sensor Networks, 10*(3), 601868.

Junghanns, M., Petermann, A., Neumann, M., & Rahm, E. (2017). Management and analysis of big graph data: current systems and open challenges. In *Handbook of Big Data Technologies* (pp. 457-505): Springer.

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). *Big data: Issues and challenges moving forward.* Paper presented at the 2013 46th Hawaii International Conference on System Sciences.

Kang, S., Shim, J., & Lee, S.-g. (2013). Tridex: A lightweight triple index for relational database-based semantic web data management. *Expert Systems with Applications, 40*(9), 3421-3431.

Kaoudi, Z., & Manolescu, I. (2014). *Cloud-based RDF data management.* Paper presented at the Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data.

Karnstedt, M., Sattler, K.-U., & Hauswirth, M. (2012). Scalable distributed indexing and query processing over Linked Data. *Web Semantics: Science, Services and Agents on the World Wide Web, 10*, 3-32.

Karvinen, P., Díaz-Rodríguez, N., Grönroos, S., & Lilius, J. (2019). RDF stores for enhanced living environments: an overview. In *Enhanced Living Environments* (pp. 19-52): Springer.

Khadilkar, V., Kantarcioglu, M., Thuraisingham, B., & Castagna, P. (2012). *Jena-HBase: A distributed, scalable and efficient RDF triple store.* Paper presented at the Proceedings of the 11th International Semantic Web Conference Posters & Demonstrations Track, ISWC-PD.

Khelil, A., Mesmoudi, A., Galicia, J., Bellatreche, L., Hacid, M.-S., & Coquery, E. (2020). Combining graph exploration and fragmentation for scalable rdf query processing. *Information Systems Frontiers*, 1-19.

Kim, K., Moon, B., & Kim, H.-J. (2014). RG-index: An RDF graph index for efficient SPARQL query processing. *Expert Systems with Applications, 41*(10), 4596-4607.

Kiryakov, A., Bishoa, B., Ognyanoff, D., Peikov, I., Tashev, Z., & Velkov, R. (2010). *The features of BigOWLIM that enabled the BBC's World Cup website.* Paper presented at the Workshop on Semantic Data Management (SemData).

Klyne, G., & Carroll, J. (2006). Resource Description Framework (RDF): Concepts and Abstract Syntax (feb 2006). *World Wide Web Consortium, Recommendation REC-rdf-concepts-20040210, 1*(2.6).

Konopnicki, D., & Shmueli, O. (1995). *W3qs: A query system for the world-wide web.* Paper presented at the VLDB.

Kulcu, S., Dogdu, E., & Ozbayoglu, A. M. (2016). *A survey on semantic web and big data technologies for social network analysis.* Paper presented at the 2016 IEEE International Conference on Big Data (Big Data).

Laney, D. (2012). The importance of'big data': A definition. *Gartner. Retrieved, 21*, 2014-2018.

Le-Phuoc, D., Parreira, J. X., Reynolds, V., & Hauswirth, M. (2010). *Rdf on the go: An rdf storage and query processor for mobile devices.* Paper presented at the 9th International Semantic Web Conference (ISWC2010).

Lee, E. A., & Messerschmitt, D. G. (1987). Static scheduling of synchronous data flow programs for digital signal processing. *IEEE Transactions on computers, 100*(1), 24-35.

Lehmann, D., Fekete, D., & Vossen, G. (2016). *Technology selection for big data and analytical applications*. Retrieved from

Leng, Y., Zhikui, C., Zhong, F., Li, X., Hu, Y., & Yang, C. (2017). BRGP: a balanced RDF graph partitioning algorithm for cloud storage. *Concurrency and Computation: Practice and Experience, 29*(14), e3896.

Li, R., Hu, H., Li, H., Wu, Y., & Yang, J. (2016). MapReduce parallel programming model: a state-of-the-art survey. *International Journal of Parallel Programming, 44*(4), 832-866.

Li, Y., Chen, W., Wang, Y., & Zhang, Z.-L. (2013). *Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships.* Paper presented at the Proceedings of the sixth ACM international conference on Web search and data mining.

Lu, Y., Zhang, M., Witherspoon, S., Yesha, Y., Yesha, Y., & Rishe, N. (2013). *sksOpen: efficient indexing, querying, and visualization of geo-spatial big data.* Paper presented at the 2013 12th International Conference on Machine Learning and Applications.

Ma, Z., Capretz, M. A., & Yan, L. (2016). Storing massive Resource Description Framework (RDF) data: a survey. *The Knowledge Engineering Review, 31*(4), 391.

Ma, Z., & Yan, L. (2019). Towards Massive RDF Storage in NoSQL Databases: A Survey. In *Emerging Technologies and Applications in Data Processing and Management* (pp. 263-284): IGI Global.

Mami, M. N., Graux, D., Scerri, S., Jabeen, H., Auer, S., & Lehmann, J. (2019). *Squerall: Virtual ontology-based access to heterogeneous and large data sources.* Paper presented at the International Semantic Web Conference.

Manyika, J. (2011). Big data: The next frontier for innovation, competition, and productivity. *http://www. mckinsey. com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation*.

McBride, B. (2001). *Jena: Implementing the rdf model and syntax specification.* Paper presented at the Proceedings of the Second International Conference on Semantic Web-Volume 40.

McGuinness, D. L., & Van Harmelen, F. (2004). OWL web ontology language overview. *W3C recommendation, 10*(10), 2004.

Mell, P., & Grance, T. (2011). The NIST definition of cloud computing.

Mountantonakis, M., & Tzitzikas, Y. (2020). Content-based Union and Complement Metrics for Dataset Search over RDF Knowledge Graphs. *Journal of Data and Information Quality (JDIQ), 12*(2), 1-31.

Muhleisen, H., & Dentler, K. (2012). Large-scale storage and reasoning for semantic data using swarms. *IEEE Computational Intelligence Magazine, 7*(2), 32-44.

Nenov, Y., Piro, R., Motik, B., Horrocks, I., Wu, Z., & Banerjee, J. (2015). *RDFox: A highly-scalable RDF store.* Paper presented at the International Semantic Web Conference.

Neumann, T., & Weikum, G. (2008). RDF-3X: a RISC-style engine for RDF. *Proceedings of the VLDB Endowment, 1*(1), 647-659.

Nitta, K., & Savnik, I. (2014). *Survey of RDF storage managers.* Paper presented at the Proceedings of the 6th international conference on advances in databases, knowledge, and data applications (DBKDA'14), Chamonix, France.

O'Sullivan, B. (2009). *Mercurial: The Definitive Guide: The Definitive Guide*: " O'Reilly Media, Inc.".

Odom, P. S., & Massey, M. J. (2003). Tiered hashing for data access. In: Google Patents.

Oh, H., Chun, S., Eom, S., & Lee, K.-H. (2015). *Job-optimized map-side join processing using mapreduce and hbase with abstract RDF data.* Paper presented at the 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT).

Olston, C., Chiou, G., Chitnis, L., Liu, F., Han, Y., Larsson, M., . . . Seth, S. (2011). *Nova: continuous pig/hadoop workflows.* Paper presented at the Proceedings of the 2011 ACM SIGMOD International Conference on Management of data.

Oren, E., Heitmann, B., & Decker, S. (2008). ActiveRDF: Embedding Semantic Web data into object-oriented languages. *Journal of Web Semantics, 6*(3), 191-202.

Oussous, A., Benjelloun, F.-Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences, 30*(4), 431-448.

Padhy, R. P. (2013). Big data processing with Hadoop-MapReduce in cloud systems. *International Journal of Cloud Computing and Services Science, 2*(1), 16.

Papailiou, N., Konstantinou, I., Tsoumakos, D., & Koziris, N. (2012). *H2RDF: adaptive query processing on RDF data in the cloud.* Paper presented at the Proceedings of the 21st International Conference on World Wide Web.

Papailiou, N., Tsoumakos, D., Konstantinou, I., Karras, P., & Koziris, N. (2014). *H2RDF+ an efficient data management system for big RDF graphs.* Paper presented at the Proceedings of the 2014 ACM SIGMOD international conference on Management of data.

Patil, N., Kiran, P., Kiran, N., & KM, N. P. (2018). A survey on graph database management techniques for huge unstructured data. *International Journal of Electrical and Computer Engineering, 8*(2), 1140.

Peng, P., Ge, Q., Zou, L., Özsu, M. T., Xu, Z., & Zhao, D. (2019). Optimizing Multi-Query Evaluation in Federated RDF Systems. *IEEE Transactions on Knowledge and Data Engineering*.

Peng, P., Zou, L., Özsu, M. T., Chen, L., & Zhao, D. (2016). Processing SPARQL queries over distributed RDF graphs. *The VLDB Journal, 25*(2), 243-268.

Pham, C. M., Dogaru, V., Wagle, R., Venkatramani, C., Kalbarczyk, Z., & Iyer, R. (2014). *An evaluation of zookeeper for high availability in system S.* Paper presented at the Proceedings of the 5th ACM/SPEC international conference on Performance engineering.

Potter, A., Motik, B., Nenov, Y., & Horrocks, I. (2016). *Distributed RDF query answering with dynamic data exchange.* Paper presented at the International Semantic Web Conference.

Prud'hommeaux, E. (2008). SPARQL query language for RDF, W3C recommendation. *http://www. w3. org/TR/rdf-sparql-query/*.

Pujari, A. K. (2001). *Data mining techniques*: Universities press.

Punnoose, R., Crainiceanu, A., & Rapp, D. (2012). *Rya: a scalable RDF triple store for the clouds.* Paper presented at the Proceedings of the 1st International Workshop on Cloud Intelligence.

Punnoose, R., Crainiceanu, A., & Rapp, D. (2015). SPARQL in the cloud using Rya. *Information Systems, 48*, 181-195.

Purohit, S., Paulson, P., & Rodriguez, L. (2016). *User-Centric Approach for Benchmark RDF Data Generator in Big Data Performance Analysis.* Paper presented at the 2016 IEEE Tenth International Conference on Semantic Computing (ICSC).

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health information science and systems, 2*(1), 3.

Rajith, A., Nishimura, S., & Yokota, H. (2016). *JARS: Join-Aware Distributed RDF Storage.* Paper presented at the Proceedings of the 20th International Database Engineering & Applications Symposium.

Ranichandra, C., & Tripathy, B. (2019). Architecture for distributed query processing using the RDF data in cloud environment. *Evolutionary Intelligence*, 1-9.

Redaschi, N., & Consortium, U. (2009). Uniprot in RDF: Tackling data integration and distributed annotation with the semantic web. *Nature precedings*, 1-1.

Ren, X. (2018). *Distributed RDF stream processing and reasoning.*

Rietveld, L., Verborgh, R., Beek, W., Vander Sande, M., & Schlobach, S. (2015). *Linked data-as-a-service: the semantic web redeployed.* Paper presented at the European Semantic Web Conference.

Rohloff, K., & Schantz, R. E. (2010). *High-performance, massively scalable distributed systems using the MapReduce software framework: the SHARD triple-store.* Paper presented at the Programming support innovations for emerging distributed applications.

Rong, C. (2011). *Using Mahout for clustering Wikipedia's latest articles: A comparison between k-means and fuzzy c-means in the cloud.* Paper presented at the 2011 IEEE Third International Conference on Cloud Computing Technology and Science.

Roy, R., Paul, A., Bhimjyani, P., Dey, N., Ganguly, D., Das, A. K., & Saha, S. (2020). A Short Review on Applications of Big Data Analytics. In *Emerging Technology in Modelling and Graphics* (pp. 265-278): Springer.

Russom, P. (2011). Big data analytics. *TDWI best practices report, fourth quarter, 19*(4), 1-34.

Sakr, S., Wylot, M., Mutharaju, R., Le Phuoc, D., & Fundulaki, I. (2018). Distributed RDF Query Processing. In *Linked Data* (pp. 51-83): Springer.

Schätzle, A., Przyjaciel-Zablocki, M., Berberich, T., & Lausen, G. (2015). S2X: graph-parallel querying of RDF with GraphX. In *Biomedical Data Management and Graph Online Querying* (pp. 155-168): Springer.

Schätzle, A., Przyjaciel-Zablocki, M., Hornung, T., & Lausen, G. (2013). *PigSPARQL: A SPARQL Query Processing Baseline for Big Data.* Paper presented at the International Semantic Web Conference (Posters & Demos).

Shang, W., Jiang, Z. M., Hemmati, H., Adams, B., Hassan, A. E., & Martin, P. (2013). *Assisting developers of big data analytics applications when deploying on hadoop clouds.* Paper presented at the 2013 35th International Conference on Software Engineering (ICSE).

Shankar, D., Lu, X., Wasi-ur-Rahman, M., Islam, N., & Panda, D. K. D. (2014). *A micro-benchmark suite for evaluating Hadoop MapReduce on high-performance networks.* Paper presented at the Workshop on Big Data Benchmarks, Performance Optimization, and Emerging Hardware.

Song, M., Oh, H., Seo, S., & Lee, K.-H. J. J. o. D. M. (2019). Map-Side Join Processing of SPARQL Queries Based on Abstract RDF Data Filtering. *30*(1), 22-40.

Stillerman, M. A., & Joyce, R. A. (2014). Scalable distributed processing of RDF data. In: Google Patents.

Storey, V. C., & Song, I.-Y. (2017). Big data technologies and management: What conceptual modeling can do. *Data & Knowledge Engineering, 108*, 50-67.

Su, X., Zhang, H., Riekki, J., Keränen, A., Nurminen, J. K., & Du, L. (2014). *Connecting IoT Sensors to Knowledge-based Systems by Transforming SenML to RDF.* Paper presented at the ANT/SEIT.

Tatu, M., Werner, S., Balakrishna, M., Erekhinskaya, T., & Moldovan, D. (2016). *Semantic question answering on big data.* Paper presented at the Proceedings of the International Workshop on Semantic Big Data.

Tez., A. (2015). Retrieved from http://tez.apache.org/

Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., . . . Murthy, R. (2009). Hive: a warehousing solution over a map-reduce framework. *Proceedings of the VLDB Endowment, 2*(2), 1626-1629.

Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Zhang, N., . . . Murthy, R. (2010). *Hive-a petabyte scale data warehouse using hadoop.* Paper presented at the 2010 IEEE 26th international conference on data engineering (ICDE 2010).

Urbani, J., Dutta, S., Gurajada, S., & Weikum, G. (2016). KOGNAC: efficient encoding of large knowledge graphs. *arXiv preprint arXiv:1604.04795*.

Wang, J., Zhang, Y., Gao, Y., & Xing, C. (2013). *PLSM: a highly efficient LSM-tree index supporting real-time big data analysis.* Paper presented at the 2013 IEEE 37th Annual Computer Software and Applications Conference.

Wayner, P. (2012). 7 Top Tools for Taming Big Data. Retrieved from http://www.networkworld.com/reviews/2012/041812-7-top-tools-for-taming-258398.html

Weihua, M., Hong, Z., Qianmu, L., & Bin, X. (2014). Analysis of information management and scheduling technology in Hadoop. *Journal of Digital Information Management, 12*(2), 133.

Wu, B., Zhou, Y., Yuan, P., Liu, L., & Jin, H. (2015). *Scalable SPARQL querying using path partitioning.* Paper presented at the 2015 IEEE 31st International Conference on Data Engineering.

Wylot, M., & Cudré-Mauroux, P. (2015). Diplocloud: Efficient and scalable management of rdf data in the cloud. *IEEE Transactions on Knowledge and Data Engineering, 28*(3), 659-674.

Wylot, M., Hauswirth, M., Cudré-Mauroux, P., & Sakr, S. (2018). RDF data storage and query processing schemes: A survey. *ACM Computing Surveys (CSUR), 51*(4), 84.

Xu, H., Chen, X., & Fan, G. (2019). *Ecosystem Description of Hadoop Platform Based on HDFS, MapReduce and Data Warehouse Tool Hive.* Paper presented at the The International Conference on Cyber Security Intelligence and Analytics.

Yang, C., Zhang, X., Zhong, C., Liu, C., Pei, J., Ramamohanarao, K., & Chen, J. (2014). A spatiotemporal compression based approach for efficient big data processing on cloud. *Journal of Computer and System Sciences, 80*(8), 1563-1583.

Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. *International journal of information management, 36*(6), 1231-1247.

Yoon, J., Jeong, W. S., & Ro, W. W. (2020). *Check-in: in-storage checkpointing for key-value store system leveraging flash-based SSDs.* Paper presented at the 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA).

Yu, L. (2011). Linked open data. In *A Developer's Guide to the Semantic Web* (pp. 409-466): Springer.

Yu, Q., & Bouguettaya, A. (2011). Efficient service skyline computation for composite service selection. *IEEE Transactions on Knowledge and Data Engineering, 25*(4), 776-789.

Yu, Y., Isard, M., Fetterly, D., & Budiu, M. (2008). *Erlingsson, l., Gunda, PK, and Currey, J. Dryadlinq: A system for general-purpose distributed data-parallel computing using a high-level language.* Paper presented at the OSDI'08: Eighth Symposium on Operating System Design and Implementation.

Yuan, P., Xie, C., Jin, H., Liu, L., Yang, G., & Shi, X. (2014). Dynamic and fast processing of queries on large-scale RDF data. *Knowledge and information systems, 41*(2), 311-334.

Zaslavsky, A., Perera, C., & Georgakopoulos, D. (2013). Sensing as a service and big data. *arXiv preprint arXiv:1301.0159*.

Zeng, K., Yang, J., Wang, H., Shao, B., & Wang, Z. (2013). *A distributed graph engine for web scale RDF data.* Paper presented at the Proceedings of the VLDB Endowment.

Zhang, X., Chen, L., Tong, Y., & Wang, M. (2013). *EAGRE: Towards scalable I/O efficient SPARQL query evaluation on the cloud.* Paper presented at the 2013 IEEE 29th International Conference on Data Engineering (ICDE).

Zhang, X., Chen, L., & Wang, M. (2012). *Towards efficient join processing over large RDF graph using mapreduce.* Paper presented at the International Conference on Scientific and Statistical Database Management.

Zhang, X., Song, D., Priya, S., Daniels, Z., Reynolds, K., & Heflin, J. (2014). Exploring linked data with contextual tag clouds. *Web Semantics: Science, Services and Agents on the World Wide Web, 24*, 33-39.

Zhang, X., Zhang, M., Peng, P., Song, J., Feng, Z., & Zou, L. (2019). *A Scalable Sparse Matrix-Based Join for SPARQL Query Processing.* Paper presented at the International Conference on Database Systems for Advanced Applications.

Zhong, Y., Fang, J., & Zhao, X. (2013). *VegaIndexer: A distributed composite index scheme for big spatio-temporal sensor data on cloud.* Paper presented at the 2013 IEEE International Geoscience and Remote Sensing Symposium-IGARSS.

Zikopoulos, P., Deroos, D., Parasuraman, K., Deutsch, T., Giles, J., & Corrigan, D. (2012). *Harness the power of big data The IBM big data platform*: McGraw Hill Professional.

Zou, L., Özsu, M. T., Chen, L., Shen, X., Huang, R., & Zhao, D. (2014). gStore: a graph-based SPARQL query engine. *The VLDB Journal, 23*(4), 565-590.