the results of this study can be a source of reference for future research.

IV. Conclusion

This study displays the process of cleaning data from dirty data to obtain the reliability aspect of data quality. Data cleansing is done using Spark SQL and Oracle. It is known that the two tools have different architecture and processing techniques. Therefore, based on the results obtained, it confirmed that the performance of Apache Spark in terms of query processing time is superior to Oracle Database. However, both are not similar tools. Thus, it can be said that Oracle Database is more suitable for use as a data management tool in large sizes. Meanwhile, Apache Spark is more suitable for use as a data analysis tool. Therefore, to do data analytics, data should be first loaded in Apache Spark from Oracle Database and then moved into Spark Warehouse using JDBC connector.

Given the limitations of the material and methodology used in this study, then for future research, it is recommended that a comparison can be made in several more aspects, such as memory usage and CPU usage. Other than that, future research can consider adding query optimization procedures into the testing queries. Thus, the comparison is not only done with similar queries but also with more diverse queries.

References

[1] I. Taleb and M. A. Serhani, "Big Data Pre-Processing: Closing the Data Quality Enforcement Loop," in *IEEE International Congress on Big Data (BigData Congress)*, Honolulu, 2017.

[2] H. A. Sulistyo, T. F. Kusumasari and E. N. Alam, "Implementation of Data Cleansing Null Method for Data Quality Management Dashboard using Pentaho Data Integration," in *International Conference on Information and Communications Technology*, Yogyakarta, 2020.

[3] F. Boufarez, A. B. Salem, M. Rehab and S. Correia, "Similar Data Elimination: MFB Algorithm," *2013 International Conference on Control, Decision and Information Technologies (CoDIT),* p. 289, 2013.

[4] I. Taleb, M. A. Serhani and R. Dssouli, "Big Data Quality Assessment Model for Unstructured Data," in *International Conference on Innovations in Information Technology (IIT)*, Al Ain, 2018.

[5] A. Juneja and N. N. Das, "Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con)*, Faridabad, 2019.

[6] T. Hongxun, W. Honggang and Z. Kun, "Data Quality Assessment for On-line Monitoring and Measuring System of Power Quality Based on Big Data and Data Provenance Theory," in *3rd IEEE International Conference on Cloud Computing and Big Data Analysis*, Chengdu, 2018.

[7] S. Loetpipatwanich and P. Vichitthamaros, "Sakdas: A Python Package for Data Profiling and Data Quality Auditing," in *1st International Conference on Big Data Analytics and Practices*, Bangkok, 2020.

[8] S. R. Amethyst, T. F. Kusumasari and M. A. Hasibuan, "Data Pattern Single Column Analysis for Data Profiling using an Open Source Platform," in *IOP Conference Series Materials Science and Engineering*, 2018.

[9] S. Juddoo and C. George, "A Qualitative Assessment of Machine Learning Support for Detecting Data Completeness and Accuracy Issues to Improve Data Analytics in Big Data for the Healthcare Industry," in *3rd International Conference on Emerging Trends in Electrical, Electronic, and Communications Engineering*, Balaclava, 2020.

[10] F. Haneem, N. Kama, R. Ali and S. Basri, "Resolving data duplication, inaccuracy and inconsistency issues using Master Data Management," 2017.

[11] E. Rahm and H. H. Do, "Data Cleaning: Problems and Current Approaches," *IEEE Bulletin of the Technical Committee on Data Engineering,* vol. 23, pp. 3-13, 2000.

[12] T. F. Kusumasari and Fitria, "Data Profiling for Data Quality Improvement with Openrefine," in *International Conference on Information Technology Systems and Innovation* , Bali, 2016.

[13] V. Kumar and C. Khosla, "Data Cleaning-A Thorough Analysis and Survey on Unstructured Data," in *8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, 2018.

[14] I. F. Ilyas and X. Chu, Data Cleaning, Association for Computing Machinery, 2019.

[15] F. Ridzuan and W. M. N. W. Zainon, "A Review on Data Cleansing Methods for Big Data," in *The Fifth Information Systems International Conference 2019*, 2019.

[16] J. Yin, J. Zhang, D. Li, T. Wang and K. Jing, "Big data cleaning model of smart grid based on Tensor Tucker decomposition," in *International Conference on Big Data & Artificial Intelligence & Software Engineering* , Bangkok, 2020.

[17] A. Wakde, P. Shende, S. Waydande, S. Uttarwar and G. Deshmukh, "Comparative Analysis of Hadoop Tools and Spark Technology," in *Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, 2018.

[18] X. Li and W. Zhou, "Performance Comparison of Hive, Impala and Spark SQL," in *International Conference on Intelligent Human-Machine Systems and Cybernetics*, Hangzhou, 2015.

[19] G. Gousios, "Big Data Software Analytics with Apache Spark," in *International Conference on Software Engineering: Companion Proceedings*, 2018.

[20] H. Müller and J.-C. Freytag, "Problems, methods, and challenges in comprehensive data cleansing," Humboldt University Berlin, 2003.

[21] C. Li, Y. Hou and Z. Yu, "Research on data cleaning technology based on instance level," *Journal of Physics,* pp. 1-4, 2019.

[22] M. Smallcombe, "Top Data Cleansing Tools for 2022," integrate.io, 12 January 2021. [Online]. Available: https://www.integrate.io/blog/top-10-data-cleansing-tools/. [Accessed 6 January 2022].

[23] F. Chen and L. Jiang, "A parallel algorithm for data cleansing in incomplete information systems using MapReduce," in *International Conference on Computational Intelligence and Security*, Kunming, 2014.

[24] S. Padhy and G. M. M. Kumaran, "A Quantitative Performance Analysis between Mongodb and Oracle NoSQL," in *International Conference on Computing for Sustainable Global Development*, New Delhi, 2019.

[25] A. V. Hazarika, G. J. S. R. Ram and E. Jain, "Performance Comparision of Hadoop and Spark Engine," in *International conference on I-SMAC*, Palladam, 2017.

[26] R. Poljak, P. Poscic and D. Jaksic, "Comparative Analysis of the Selected Relational Database Management Systems," in *International Convention on Information and Communication Technology, Electronics and Microelectronics*, Opatija, 2017.

[27] J. Powers, "Apache Spark Performance Compared to a Traditional Relational Database using Open Source Big Data Health Software," *Project Paper for CSE8803 Big Data Analytics for Health Care,* pp. 1-5, 24 April 2016.