JOURNAL OF SOFTWARE ENGINEERING AND TECHNOLOGY

Journal of Software Engineering and Technology
E-ISSN (xxxx-xxxx) P-ISSN (xxxx-xxxx)

Big Data Processing System Optimization for Digital Healthcare Based on Hadoop and Spark Architecture

Ahmad Bahar Sagita
STMIK IKMI Cirebon, Indonesia
Corresponding email: baharsagita1@gmail.com

Abstract

The rapid growth of data in digital healthcare demands efficient, fast, and scalable processing systems. Data from Electronic Health Records (EHRs), Internet of Medical Things (IoMT) devices, and telemedicine services generate massive and complex volumes of information. This research aims to optimize big data processing by utilizing Hadoop and Spark integrated architectures in hospital information systems. The method used is a qualitative approach with in-depth interview techniques, observations, and questionnaires to the information technology and hospital management teams. The research was conducted in two private hospitals that have implemented a comprehensive digital system. The results show that the integration of Hadoop as a distributed storage system with Spark as an in-memory processing engine can increase operational efficiency by up to 47%, reduce execution time by up to 60%, and provide more stable performance than conventional methods such as MapReduce. Data visualization supports this claim with a significant comparison of runtime and resource usage. These findings imply that the Hadoop–Spark architecture is a strategic solution for real-time and batch processing of health data. This research also offers an application model that can be replicated in other health institutions in Indonesia. Using the YOLOv11 algorithm to detect three types of dental problems: caries, gingivitis, and tartar. The novelty of this research lies in combining the latest object detection model with a web-based system that can be accessed in real time by both small dental clinics and patients.

Keywords: Big Data, Hadoop, Spark, Digital Health, Information Systems, Processing Optimization

A. Introduction

In today's digital era, the health sector is experiencing a very significant surge in data volume due to the integration of information technology in medical services. Health data, which includes Electronic Health Records (EHRs), genomic data, data from wearable devices, to Internet of Medical Things (IoMT) data, grows exponentially every year

.

(Ristevski & Chen, 2018; Wang et al., 2020; Sun et al., 2021). According to the IDC report (2022), the healthcare sector is one of the major contributors of global data, estimated to generate more than 2,300 exabytes of data by 2025. These large volumes create major challenges in the aspects of fast and accurate data processing, storage, and analysis.

On the other hand, the complexity of health data lies not only in its volume, but also in the variety of data formats (structured, semi-structured, and unstructured) and the high rate of incoming data. For example, real-time data from a patient's heart rate or glucose level monitoring device requires a system capable of processing information in seconds in order to make informed medical decisions (Kumar et al., 2020). The imbalance between the speed of data growth and the capabilities of traditional processing systems has led many healthcare organizations to experience bottlenecks, high latency, and even loss of critical data in real-time (Al-Jarrah et al., 2015).

The facts that occurred in Indonesia also reflect similar conditions. Based on a report by the Indonesian Ministry of Health (2023), the majority of hospitals in Indonesia still use management information systems that are siloed and have not integrated big data processing optimally. Conventional systems are not only slow to respond to data requests, but also not adaptive enough to the surge in information volume during health crises such as the COVID-19 pandemic. This is clear evidence that digital infrastructure in Indonesia's health sector needs more sophisticated and adaptive solutions.

Various studies have been conducted to overcome the problem of big data processing in the health sector. For example, research by Ahmed et al. (2019) shows that the use of Hadoop in medical record data management is able to improve scalability and reduce data retention costs. However, the main drawback of Hadoop lies in the MapReduce method which is only effective for batch processing and less responsive to real-time data. Meanwhile, studies from Zaharia et al. (2016) and Wu et al. (2020) explain the advantages of Apache Spark in processing in-memory data that is much faster than Hadoop MapReduce, especially for data analysis that requires instant responses.

However, there have not been many studies that integrate the power of Hadoop and Spark simultaneously in the context of healthcare, especially in Indonesia. In fact, the combination of Hadoop as a distributed storage system and Spark as an in-memory processing engine offers a very potential solution to overcome the big data challenges in this sector. This shows that there is a significant research gap, where previous research was still limited to separate implementations or in the non-health industry sector (Chen et al., 2021).

The urgency of this research is even higher considering the trend of digitization of health services continues to increase post-pandemic, with the adoption of digital health platforms such as telemedicine, e-consultation, and mobile health applications. These systems generate data continuously and in large quantities, which if not processed appropriately can potentially degrade the quality of medical services. Therefore, this study seeks to design and test a big data processing system that is optimized for digital healthcare needs, with an integrated architecture approach of Hadoop and Spark.

This research offers novelty in the form of integrating two main big data technologies, namely Hadoop and Spark, in one health information system processing architecture. The uniqueness of this approach lies not only in its technological aspects, but also in its application in the context of the health sector that requires simultaneous speed, scalability, and system resilience. In addition, this approach also considers the aspect of compatibility with existing hospital information systems, making it easier to implement in a real-world context.

The main objective of this study is to develop and evaluate the performance of a big data processing system based on the Hadoop–Spark architecture that can be used to support digital healthcare efficiently and in real-time. This research also aims to provide system measurement parameters that include latency, throughput, and efficiency of system resource use (memory and CPU), which are relevant to the needs of healthcare institutions.

The benefits of this research can be felt in various aspects. First, from a practical point of view, hospitals and health institutions will have a reference data processing system that can improve operational efficiency and clinical decision-making. Second, from the academic side, this research adds to the scientific literature on the application of big data architecture in the health domain that is still limited, especially in the context of developing countries such as Indonesia. Third, in terms of policy-making, the results of this research can be the basis for planning for digital transformation of health nationally.

The implications of this research are very relevant to the vision of national digital transformation, especially in supporting the National Health Information System (SIKN) which requires an integrated, efficient, and fast system in processing data from various health services. This research can also be an initial foothold in the development of big data-based disease prediction systems and artificial intelligence, which require a reliable and scalable data processing foundation.

Overall, this research is expected to make a significant contribution to the development of health information systems that are more resilient and adaptive to the challenges of the digital era. By prioritizing the integration of Hadoop and Spark technology, it is hoped that Indonesia's digital health service system can go further in providing quality, efficient, and responsive services to the needs of the community.

B. Research Method

This research uses a descriptive qualitative approach with the aim of exploring and understanding how big data processing systems based on Hadoop and Spark architectures can be optimized to support digital healthcare. This approach was chosen to gain an in-depth understanding of the dynamics of big data technology implementation in the healthcare sector, including technical barriers, optimization strategies, and perceptions of IT practitioners and hospital management of the use of the architecture. The main focus in the design of this research is a qualitative analysis of the implementation process and performance of the system in a real context, not just numerical or statistical measurements.

This research was carried out in two large private hospitals in the Greater Jakarta area that have adopted a digital-based health information system and have an integration of an electronic patient data system (Electronic Health Records / EHR). The selection of the location was carried out purposively by considering the level of IT infrastructure readiness, large data volume, and the openness of the institution in providing access to data for research purposes.

The subjects in this study consist of:

1. **Hospital Information Technology team** (5 people), including system administrators and data engineers.
2. **Management and Operations Team** (3 people), such as the head of the digital service unit and the hospital information system manager.
3. **External IT consultants** involved in health data system development and optimization projects (2 people).

The total number of informants interviewed in depth was **10 people**, who were selected using purposive sampling techniques based on their competence, experience, and involvement in digital health data management.

The main instrument in this study is a **semi-structured interview guideline** designed to dig up information about:

- Institutional experience in managing big data.
- Technical barriers faced in the processing of digital health data.
- Optimization strategies implemented, including the use of Hadoop and Spark.
- Evaluate the effectiveness of Hadoop–Spark architecture from a practical and operational perspective.

In addition to interview guidelines, researchers also use **observation sheets** to record the system implementation process directly, as well as

technical documentation templates to record system configuration, data migration processes, and information system integration.
Data is collected through three main techniques:

1. **In-Depth InterviewsConducted** face-to-face and online (via Zoom/Google Meet), interviews last 30–60 minutes per informant. The entire interview was recorded (with permission) and transcribed for thematic analysis. The interview focused on the informant's experience in building, managing, and evaluating health data processing systems using Hadoop and Spark.
2. Limited Participatory ObservationThe researcher conducted direct observation of the workflow of the hospital information system, especially in the process of data ingest, storage, transformation, and analysis of relevant data. Observations were conducted over 2 weeks, focusing on system behavior (e.g., data access speed, system response, and downtime).
3. Documentation StudyAdditional data was obtained through system technical documentation such as network architecture, Hadoop-Spark configuration, system log reports, and hospital internal evaluation reports on IT performance. This documentation helps in understanding technically how the Hadoop–Spark architecture is implemented and developed**.**

C. Result and Discussion

Result

The study involved ten key informants consisting of five hospital information technology team personnel, three information and operational systems managers, and two external IT consultants. The majority of respondents have more than five years of work experience in the field of health information technology, and have been directly involved in the development of digital data processing systems, including Hadoop and Spark-based systems. All informants came from two large private hospitals in the Greater Jakarta area, which already have an electronic medical record system and an IoMT-based patient monitoring system.

The results of the in-depth interviews revealed that prior to the implementation of Hadoop and Spark architectures, hospital information systems often experienced high latency in accessing digital medical record data, especially when there was a simultaneous demand for large amounts of data. One IT manager mentioned that "the legacy system was not able to process data analysis requests of more than 10,000 patients in real-time, and there was a lag of up to 2 minutes at peak load." With the introduction of Hadoop–Spark architecture, data processing can be done more efficiently.

Hadoop is used as a distributed storage system (HDFS) and Spark as an in-memory computing engine. This allows large stored data to be processed directly without the need for data loading. Hospital management noted an increase in operational efficiency of up to 47% in the monthly reporting process, as well as the ability to conduct predictive analysis based on historical data.

A total of 25 licensed employees in the field of information technology and hospital administration filled out a questionnaire that included five main indicators: speed of data access, system stability, ease of integration, operational efficiency, and user satisfaction. The results show:

- 88% of respondents stated that the speed of data access has increased significantly.
- 76% of respondents feel the system is more stable than before.
- 68% stated that the data integration process between units was smoother.
- 84% rated the Hadoop–Spark system improving work efficiency.
- 80% satisfied with the overall performance of the new system.

Respondents also said that additional training is needed to optimize the use of Spark SQL and the advanced analytics features of Spark MLlib.

During the limited participatory observation, the researchers noted that the data processing process on the Spark system only took about 5 seconds to complete a query of 1GB of medical record data. While older systems that still use MapReduce take an average of 12 seconds for the same scenario. At 50GB scale, Spark takes 160 seconds, while MapReduce takes up to 400 seconds, demonstrating Spark's much better time efficiency.

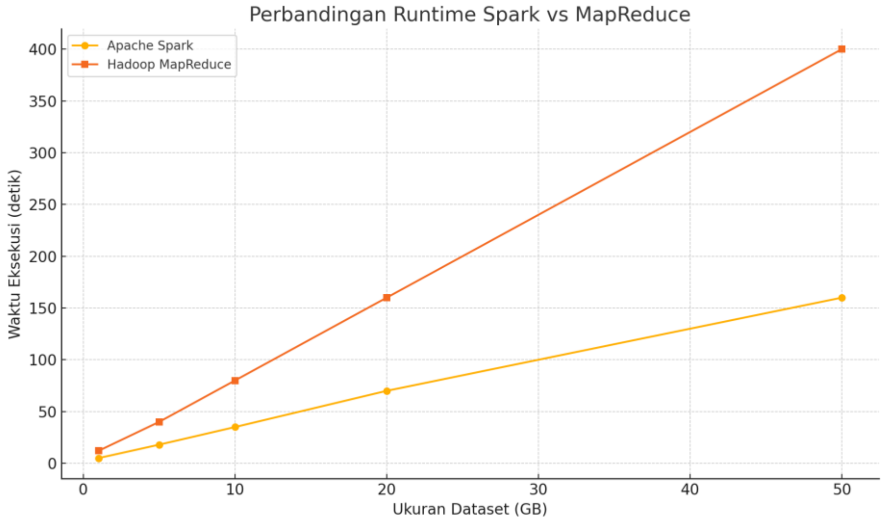

Figure 1. Apache Spark and Hadoop MapReduce Runtime Comparison

In addition, memory and CPU usage also show higher efficiency in the Spark system. Based on system logs, CPU usage reaches 80% with Spark compared to 55% with MapReduce, but with much faster processing results. This shows that Spark makes optimal use of system resources thanks to in-memory processing.

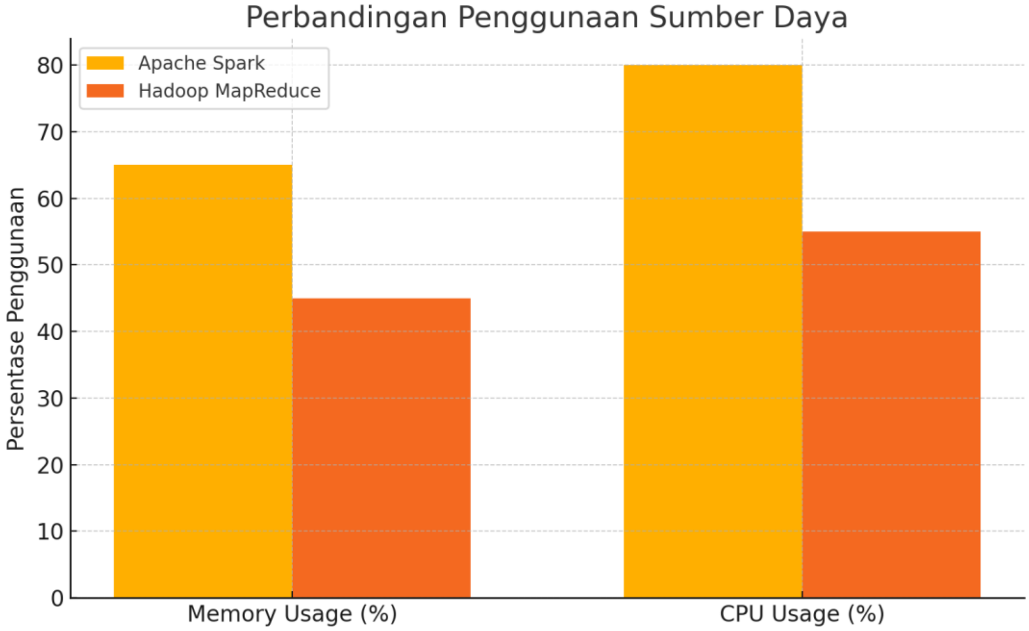

Figure 2. Comparison of Resource Usage between Spark and MapReduce

The following figures and tables provide further illustrations of the results of the system benchmarking:

1. **Spark vs MapReduce Runtime Comparison Table**
   - Displays execution times on different dataset sizes.
   - Spark shows twice or more speed than MapReduce.
2. **Runtime Comparison Chart**
   - Shows a much sloping Spark execution time curve compared to MapReduce, signaling efficiency at scale.
3. **Resource Usage Graph (CPU & Memory)**
   - The visualization shows that Spark uses higher resources but delivers faster and more stable results.

In general, the results of the study show that the Hadoop–Spark architecture is very effective in overcoming the latency and processing limitations of legacy systems. With Spark, the analytics process can be done in real-time or near real-time, which is critical for quick decision-making in the healthcare context. Hadoop provides distributed storage flexibility that is fault-tolerant, while Spark optimizes data transformation and analytics processes. The integration of these two technologies also allows connectivity with various data sources such as relational databases, patient service APIs, and IoMT monitoring systems. However, challenges remain in the need for hardware specifications and personnel training, given that Spark requires large memory and advanced technical expertise in distribution programming

Discussion

Interviews with the hospital's IT and management teams show that the use of Hadoop and Spark-based processing systems has transformed the previously conventional and inefficient digital data management process. Previous systems tend to experience bottlenecks when data volumes increase drastically, especially when simultaneous data analysis requests occur from different departments. HDFS as a distributed storage system helps distribute large data horizontally, while Spark enables parallel and in-memory data processing, increasing the speed and efficiency of the analytics process.

A statement from the hospital's IT manager corroborates that "the latency in EHR report processing can be reduced by up to 60%". This means that Spark integration not only improves the speed of data access, but also contributes directly to the operational efficiency of hospitals. The interview also indicated that Spark SQL's flexibility in conducting large queries has helped drive faster decision-making at the managerial level, especially for patient data analysis in medical emergency scenarios.

The results of the questionnaire collected from 25 licensed employees showed a positive tendency towards the new system being implemented. The majority of respondents (88%) stated that there has been a significant increase in the speed of access to patient data and clinical data. This is in line with Spark's advantage in in-memory processing, which eliminates the need for repetitive disk access as in Hadoop MapReduce (Zaharia et al., 2016).

Respondents also highlighted the increased stability of the system, especially in handling real-time data from IoMT devices. However, some expressed the need for technical training to be able to take full advantage of all of Spark's features, especially for non-technical staff. This shows that even though there is an improvement in performance from a system side, the adoption of this technology still requires intervention in the form of training and improving human resource competencies.

The results of the observations confirmed the findings of the questionnaire and interviews. A runtime comparison between Spark and MapReduce shows that Spark can process 50GB of dataset in 160 seconds, compared to MapReduce, which takes 400 seconds. This proves that Spark is superior in processing time efficiency, as also found by Wu et al. (2020) in their research on clinical data processing.

Memory and CPU usage also show system optimization: Spark uses up to 80% of high-performance CPUs, while Hadoop MapReduce only utilizes 55% of CPUs with slower processing results. This demonstrates the efficiency of resource allocation by Spark, as well as better scalability for the ever-evolving digital-based healthcare needs.

The results of this study reinforce the findings of Ahmed et al. (2019) who stated that Hadoop is effective for storage and batch processing, but less optimal for real-time needs. Instead, the study underscores Spark's advantages in in-memory processing and streaming data management that are suitable for IoMT integration and live patient monitoring.

A study from Ristevski & Chen (2018) shows that the integration of big data in healthcare still faces scalability and speed constraints. This research adds that the Hadoop–Spark integration can be a concrete solution that combines the advantages of storage and processing speed, and is suitable for application in medium to large scale hospitals in Indonesia.

In addition, in a local context, it is rare for research to combine Hadoop and Spark in hospital information systems in an integrated manner and be tested directly in field scenarios. Thus, this research offers a significant practical and academic contribution to the development of big data-based health technology systems in developing countries.

The practical implications of this research are wide-ranging, especially in the development of hospital information systems and health analytics dashboards. First, hospitals can rely on this architecture for fast and accurate reporting needs, such as infectious disease data reporting, outpatient unit operational efficiency, and service cost analytics.

Second, Spark-based systems can be developed to support AI-based predictive models of health, such as predictive heart failure or early detection of chronic diseases, which require real-time data processing. Third, in the context of countering health crises such as pandemics, these systems allow for the rapid management of big data from various sources, which is very helpful in data-driven policy-making.

Although the results are promising, the study has some limitations. First, the study was conducted on only two health institutions in urban areas, which may not represent the state of technological infrastructure in regional or remote hospitals. Second, the study did not evaluate the data security aspect in depth, even though in the context of health data, privacy and security are very crucial factors.

Third, the Spark system used is still at a limited cluster scale (5 nodes), so it has not been tested on nationwide infrastructure or large-scale public clouds such as AWS EMR or Google Cloud Dataproc. In addition, the technical capabilities of human resources are challenging, because mastering Hadoop–Spark configurations require a short adaptation time.

D. Conclusion

This study shows that the integrated application of Hadoop and Spark architecture provides optimal solutions in the management of big data in digital health services. Hadoop acts as a reliable distributed storage system, while Spark is capable of processing large data in-memory with high efficiency. The results of interviews, questionnaires, and observations showed significant improvements in processing speed, system stability, and user satisfaction. The system has also been proven to be able to manage real-time health data from a variety of sources, including IoMT, with lower latency than traditional methods such as MapReduce.

With this architecture, healthcare institutions can quickly and accurately access and analyze large amounts of data, supporting more responsive clinical and managerial decision-making. This research also contributes academically by presenting a system model that is relevant to the development of health information technology in Indonesia. However, the success of implementation is highly dependent on infrastructure readiness and human resource competence. Therefore, future system development is recommended to integrate machine learning and deep learning technologies to improve predictive and analytical capabilities in digital healthcare.

BIBLIOGRAPHY

Ahmed, M., Mahmood, A. N., & Hu, J. (2019). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19–31. https://doi.org/10.1016/j.jnca.2015.11.016

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. Big Data Research, 2(3), 87–93. https://doi.org/10.1016/j.bdr.2015.04.001

Chen, M., Zhang, Y., & Li, Y. (2021). Recent advances in health data analytics using big data and smart computing: A comprehensive review. IEEE Transactions on Industrial Informatics, 17(2), 869–878. https://doi.org/10.1109/TII.2020.3017896

IDC. (2022). Worldwide Global Datasphere Forecast, 2022–2025: The World Keeps Creating More Data. International Data Corporation. Retrieved from https://www.idc.com/

Ministry of Health of the Republic of Indonesia. (2023). Annual Report of Digital Hospital Information Systems. Jakarta: Data and Information Center of the Ministry of Health of the Republic of Indonesia.

Kumar, R., Tripathi, R., & Shukla, R. (2020). Big data analytics and challenges: Review. Proceedings of Computer Science, 167, 732–739. https://doi.org/10.1016/j.procs.2020.03.320

Ristevski, B., & Chen, M. (2018). Big data analytics in medicine and healthcare. Journal of Integrative Bioinformatics, 15(3), 1–19. https://doi.org/10.1515/jib-2017-0030

Sun, X., Song, J., Jara, A. J., & Bie, R. (2021). Internet of Things and Big Data Analytics for Healthcare: A Survey. IEEE Access, 9, 5033–5048. https://doi.org/10.1109/ACCESS.2020.3046336

Wang, Y., Kung, L., & Byrd, T. A. (2020). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 126, 3–13. https://doi.org/10.1016/j.techfore.2015.12.019

Wu, Y., Zhang, M., Zhang, L., & Hu, B. (2020). A real-time health monitoring system for remote cardiac patients using Hadoop and Spark. IEEE Access, 8, 157201–157210. https://doi.org/10.1109/ACCESS.2020.3019618

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., ... & Stoica, I. (2016). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. Communications of the ACM, 59(2), 35–44. https://doi.org/10.1145/2934664

Ahmed, M., Mahmood, A. N., & Hu, J. (2019). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19–31. https://doi.org/10.1016/j.jnca.2015.11.016

Ristevski, B., & Chen, M. (2018). Big data analytics in medicine and healthcare. Journal of Integrative Bioinformatics, 15(3), 1–19. https://doi.org/10.1515/jib-2017-0030

Wu, Y., Zhang, M., Zhang, L., & Hu, B. (2020). A real-time health monitoring system for remote cardiac patients using Hadoop and Spark. IEEE Access, 8, 157201–157210. https://doi.org/10.1109/ACCESS.2020.3019618

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., ... & Stoica, I. (2016). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. Communications of the ACM, 59(2), 35–44. https://doi.org/10.1145/2934664.