World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

Scalable big data architectures for healthcare analytics using Spark and SQL- based pipelines

Jagadeeswar Alampally *

IQVIA Inc., USA.

World Journal of Advanced Research and Reviews, 2021, 09(03), 429-434

Publication history: Received on 21 February 2021; revised on 24 March 2021; accepted on 28 March 2021

Article DOI: https://doi.org/10.30574/wjarr.2021.9.3.0122

Abstract

The emergence of healthcare data poses challenges in data processing, storage, and analysis. This paper discusses scalable big data solutions in healthcare analytics and the application of Apache Spark and SQL-based pipelines in this context. The proposed architecture provides the means to perform real-time analytics on big data in healthcare through the use of Spark's distributed computing features and data transformation with the help of SQL. This paper discusses the design and implementation of a scalable data pipeline to suit healthcare applications and its potential use to support real-time decision-making, predictive analytics, and health monitoring systems. Performance assessment proves the scalability, performance, and capability of the architecture to process both structured and unstructured data, which opens the way to the enhanced healthcare output and efficiency in operations.

Keywords: Big Data; Healthcare Analytics; Apache Spark; SQL Pipelines; Scalable Architectures; Real-Time Data Processing; Data Management; Healthcare Systems.

1. Introduction

Big data analytics are becoming central to the healthcare industry for improving decision-making, providing optimal patient care, and enhancing operational efficiency. Nevertheless, the amount and diversity of healthcare data pose serious processing, data storage, and real-time challenges. Conventional data management systems are mostly unable to scale properly as the need for faster analytics increases. Apache Spark, with SQL-based pipelines, has become a strong solution in recent years with respect to solving these issues. The real-time processing of large volumes of data provided by Spark and the effectiveness of SQL in participating in queries of structured data are potential architectures for scalable healthcare analytics [1], [2]. This study addresses scalable big data architectures with the assistance of Spark and SQL with the application to healthcare analytics and the benefits that come with the use of these tools in relation to dealing with large and complex data to obtain real-time insights.

Integration of distributed processing frameworks and structured query pipelines tailored to healthcare analytics environments. It explores how layered architectures built on Spark clusters and SQL-driven data orchestration can accommodate heterogeneous datasets originating from health records, imaging systems, wearable devices, and administrative repositories. Emphasis is placed on mechanisms that support elasticity, fault tolerance, and throughput optimization while maintaining data integrity and governance requirements that are inherent to clinical contexts.

Furthermore, attention is directed toward the design of ingestion and transformation workflows that enable continuous data acquisition and normalization without creating latency. By leveraging columnar storage formats, partitioning strategies, and query optimization techniques, modern pipelines can deliver high-performance analytics for both batch

* Corresponding author: Jagadeeswar Alampally

and streaming workloads. Such capabilities are increasingly critical for predictive modeling, clinical decision support, and population-level health monitoring, where delayed insights may diminish their practical impact.

In addition, this study considers architectural considerations associated with interoperability and compliance, including schema harmonization, metadata management, and privacy-aware computation. The alignment of scalable computing layers with standardized healthcare data models contributes to reproducibility and cross-institutional data exchange. The investigation not only evaluates technological scalability but also situates these frameworks within broader operational and regulatory ecosystems that shape real-world deployment.

From this perspective, this study positions Spark-enabled SQL pipeline architectures as a foundational pathway toward responsive and data-driven healthcare infrastructures, establishing a basis for further discussion on implementation strategies, performance evaluation, and optimization methodologies in subsequent sections.

2. Literature Review

2.1. Big Data in Healthcare

Big data is a key element of contemporary healthcare, allowing the analysis of large volumes of patient data to enhance treatment outcomes, lower costs, and reduce operational complexity. Apache Hadoop and Apache Spark are important technologies for the storage, processing, and analysis of large datasets [2]. Big data analytics can be used for predictive modelling, monitoring patients, and optimizing clinical decisions.

2.2. Processing Frameworks for Big Data.

Healthcare analytics uses various frameworks for big data processing, one of the most conspicuous of which is Apache Spark. The Spark distributed computing feature provides a higher ability to process data faster, particularly compared with Hadoop MapReduce [4]. With the help of SQL-based pipelines and Spark support, healthcare data can be efficiently queried and transformed to guarantee the use of real-time analytics and accurate data [5].

Table 1 Comparison of Big Data Processing Frameworks for Healthcare Analytics.

Feature	Apache Spark	Apache Hadoop	Apache Flink
Processing Speed	High (Real-time processing)	Slower (Batch processing)	High (Real-time processing)
Ease of Use	Easy (Supports SQL, Python, Scala)	Complex (MapReduce paradigm)	Moderate (Stream & batch modes)
Fault Tolerance	Built-in fault tolerance	Strong fault tolerance	Strong fault tolerance
Data Processing Type	Batch and Real-time	Batch only	Batch and Real-time
Scalability	High scalability with in-memory processing	High scalability but slower performance	High scalability, optimized for stream processing
Integration with SQL	Native support for SQL queries	Limited SQL support	Limited SQL support

3. Scalable healthcare data architecture

3.1. Major Building Blocks of Scalable Architecture

Effective ingestion, storage, processing, and real-time analytics of healthcare data are necessary to develop a scalable healthcare data architecture. Some of the major elements are distributed systems, raw storage in data lakes, and real-time streaming platforms such as Apache Kafka and Spark [2]. Apache Spark can help identify risk factors in healthcare.

3.2. Apache Spark in Healthcare.

Apache Spark has the computing capacity required to handle big healthcare data. It has the advantage of in-memory processing, which makes analytics quick and SQL-based pipelines, which provide smooth data transformation and querying to support real-time applications [5].

Table 2 Key Benefits of Apache Spark in Healthcare.

Feature	Description	Benefit in Healthcare
In-Memory Processing	Spark processes data in memory instead of writing to disk	Faster data analytics, reducing latency for real-time insights
Distributed Computing	Uses cluster-based architecture to distribute computation	Scalable for handling large, complex healthcare datasets
Real-Time Analytics	Supports Spark Streaming for real-time data processing	Enables continuous patient monitoring and timely interventions
SQL-Based Pipelines	Allows seamless integration with SQL for data transformation and querying	Simplifies data cleaning and querying, optimizing decision-making

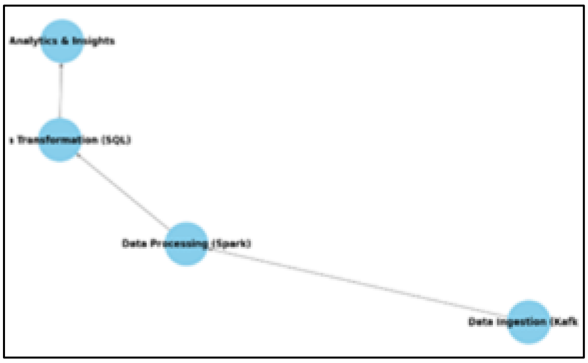


Figure 1 Pipeline Architecture Diagram

3.3. SQL Pipelines for Data Transformation.

SQL pipelines are essential to healthcare data architectures because they provide an opportunity to transform and manage structured data. These pipelines make intricate data tasks easier, making querying and integration with other healthcare systems more effective [4].

4. Suggested pipeline of analytics based on big data

4.1. Architecture Design

The suggested pipeline integrates Apache Spark for processing distributed data and SQL pipelines for data transformation efficiency, making it scalable for large healthcare datasets. Apache Kafka is applied in real-time ingestion of data, whereas Spark operates real-time on the data, which is that of high speed. Data cleaning and transformation are applied to SQL pipelines, which simplify the data cleaning process and prepare the data for further analysis, predictive modelling, and reporting [2], [5].

4.2. Data Flow and Processing

Apache Kafka is used to ingest raw healthcare data in sources such as electronic health records (EHR) and sensors in the pipeline. The data were then processed in real time using Apache Spark, and data transformation and cleaning were performed using SQL pipelines. In this way, healthcare data can be organized and prepared for use in analytics and clinical setting decisions in a timely and proper manner [4].

5. Performance analysis

5.1. Scalability

The proposed pipeline is highly scalable because Spark can process big data in distributed mode. Kafka ingestion of real-time data and Spark processing result in the architecture being able to process increasing amounts of healthcare data, even in real time, with no performance degradation [2].

5.2. Processing Efficiency

The high processing efficiency of the pipeline can be explained by the fact that Spark supports in-memory processing, which allows for fast transformation and analytics of the data. The pipeline also allows the processing of healthcare data more quickly, which eliminates the necessity of using disks in operation; thus, it is appropriate for time-sensitive systems, such as patient monitoring [5].

5.3. Performance Evaluation Metrics

To gauge the performance quality of the pipeline, the data throughput, processing speed, and latency were considered. These measures can be used to effectively evaluate how well the system can process massive data in real time, so that medical practitioners can depend on the system to provide information on an on-demand basis.

Table 3 Performance Evaluation Metrics

Metric	Description	Measurement Type
Throughput	Rate of data processed per second	Records per second
Processing Speed	Time taken to process data stages	Seconds per stage
Latency	Delay between data ingestion and analytics	Milliseconds
Scalability	Ability to scale with data volume	Data volume vs. processing time

5.4. Comparison with Other Systems

The proposed Spark-based pipeline is significantly faster and more scalable than the classical system. Although a system such as Hadoop proves useful in the context of batch processing, the real-time features of Spark provide a decisive edge over healthcare analytics, where real-time data processing is an essential requirement for decision-making [4]

6. Use cases and applications

6.1. Real-Time Health Monitoring

The specified big data architecture is particularly suitable for real-time health monitoring systems, in which constant data ingestion and processing of devices, wearables, and sensors of medical significance are essential. Real-time data processing allows healthcare providers to obtain instant data regarding patients' conditions, thus facilitating immediate interventions [1].

6.2. Disease Detection Predictive Analytics

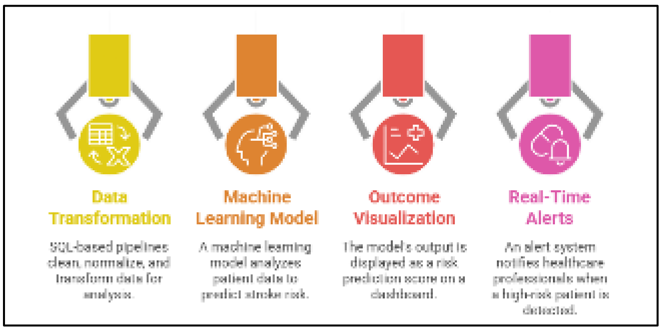
The proposed pipeline can assist healthcare systems in making predictions using real-time and historical data to improve patient outcomes. For example, using the system, the possibility of diseases, stroke, or heart failure can be predicted based on patient data over time, enhancing early diagnosis and preventive care [8].

6.3. The sixth issue is the operational efficiency of hospitals

Administrative tasks, including patient scheduling, resource allocation, and stock management in hospitals, can be streamlined by incorporating the proposed architecture into hospital management systems. This minimizes the costs of operation and enhances the general experience when treating patients [6].

6.4. Stroke Disease Prediction App.

The proposed architecture may be applied specifically to stroke disease prediction to monitor patient history, vital signs, and other data. Predictive analytics will be applied to evaluate the risk of stroke among patients to undertake early interventions via the system. Machine learning models can be used to increase the accuracy of predictions as the pipeline is improved.



SQL-based pipelines clean, normalize, and transform data for analysis.

A machine learning model analyzes patient data to predict stroke risk.

The model's output is displayed as a risk prediction score on a dashboard.

An alert system notifies healthcare professionals when a high-risk patient is detected.

Figure 2 Stroke Risk Prediction System

7. Conclusion

Apache Spark and SQL-based pipelines can be integrated to create a powerful solution to scalable big data analytics in the healthcare sector. It is an efficient architecture capable of processing large amounts of structured and unstructured data in real time and providing insights into the data. Its applications in health monitoring, predictive analytics, and operational efficiency illustrate its potential to significantly contribute to healthcare decision making. These systems must be further optimized for application in even more complex healthcare settings, and more complex machine learning models must be integrated to increase prediction accuracy in the future [1]; [4].

This combined architecture enables the efficient processing of large volumes of structured and unstructured data, supports near real-time analysis, and facilitates actionable insight generation. Its demonstrated applicability in areas such as patient health monitoring, predictive diagnostics, population-level trend detection, and operational workflow optimization highlights its relevance in improving evidence-driven clinical and administrative decision-making. By leveraging distributed computing and declarative query processing, the architecture also enhances throughput, fault tolerance, and resource utilization compared with conventional single-node analytics approaches.

Beyond technical performance, the strategic value of such integration lies in its capacity to bridge fragmented healthcare data silos and promote interoperability across electronic health records, wearable device streams, imaging repositories and laboratory systems. When implemented effectively, these pipelines can support continuous learning healthcare environments, where insights derived from large-scale analytics inform policy formulation, resource allocation, and clinical guideline refinement. Consequently, the integration of scalable data processing technologies represents not only infrastructural improvement but also an enabler of digital transformation within modern healthcare ecosystems.

However, implementation challenges remain significant. Concerns surrounding patient data confidentiality, regulatory compliance, and ethical governance must be systematically addressed to prevent the misuse or unintended exposure of sensitive information. Technical barriers, such as heterogeneous data formats, integration latency, and infrastructure costs, may also hinder deployment in resource-constrained environments. Furthermore, reliance on large-scale predictive models introduces risks associated with algorithmic bias, interpretability limitations, and overfitting, which could compromise clinical trust and adoption of these models. These constraints underscore the need for rigorous validation frameworks, transparent model evaluations, and standardized interoperability protocols to ensure responsible system utilization.

Therefore, future research should focus on enhancing distributed optimization strategies, incorporating advanced machine learning and deep learning frameworks for improved predictive accuracy, and embedding explainable artificial intelligence mechanisms to increase transparency in clinical decision support. Additional exploration of privacy-preserving techniques, including federated learning, secure multiparty computation, and differential privacy, may further strengthen trust and compliance. Investigating energy-efficient processing strategies and adaptive workload management can also improve sustainability and accessibility, particularly in emerging healthcare infrastructures.

In summary, the convergence of Apache Spark and SQL-driven pipelines presents a robust pathway for scalable, intelligent healthcare analytics. While technical and ethical challenges must be addressed, continued innovation and interdisciplinary collaboration are likely to position such architectures as foundational components in the evolution of data-centric, resilient, and patient-focused healthcare systems.

References

[1] Baljak, V., Ljubovic, A., Michel, J., Montgomery, M., & Salaway, R. (2018). A scalable realtime analytics pipeline and storage architecture for physiological monitoring big data. Smart Health, 9, 275-286.

[2] Nazari, E., Shahriari, M. H., & Tabesh, H. (2019). BigData analysis in healthcare: apache hadoop, apache spark and apache flink. Frontiers in Health Informatics, 8(1), 14.

[3] Ankam, V. (2016). Big data analytics. Packt Publishing Ltd.

[4] Muller, J., & Fischer, L. (2020). Scalable Data Architectures for Real-Time Big Data Analytics: A Comparative Study of Hadoop, Spark, and Kafka. International Journal of AI, BigData, Computational and Management Studies, 1(4), 8-18.

[5] Gupta, Y. K., & Kumari, S. (2020, December). A study of big data analytics using apache spark with Python and Scala. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 471-478). IEEE.

[6] El Aboudi, N., & Benhlima, L. (2018). Big data management for healthcare systems: architecture, requirements, and implementation. Advances in bioinformatics, 2018(1), 4059018.

[7] Chrimes, D., Moa, B., Zamani, H., & Kuo, M. H. (2016, August). Interactive healthcare big data analytics platform under simulated performance. In 2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech) (pp. 811-818). IEEE.

[8] Senbato, A. S. S. E. F. A. (2019). Designing Healthcare Data Analytics Framework Based on Big Data Approach: In Case of Stroke Disease Prediction. Addis Ababa Science and Technology University.

[9] Mittal, M., Balas, V. E., Goyal, L. M., & Kumar, R. (Eds.). (2019). Big data processing using spark in cloud (Vol. 12). Berlin, Germany: Springer.

[10] Fahimimoghaddam, G. (2021). A customizable on-demand big data health analytics platform using cloud and container technologies. Ecole Polytechnique, Montreal (Canada).

[11] Orozco-GómezSerrano, A. (2020). Adaptive Big Data Pipeline.

[12] Shaikh, E., Mohiuddin, I., Alufaisan, Y., & Nahvi, I. (2019, November). Apache spark: A big data processing engine. In 2019 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM) (pp. 1-6). IEEE.

[13] Chrimes, D., Kuo, M. H., Moa, B., & Hu, W. (2017). Towards a real-time big data analytics platform for health applications. International Journal of Big Data Intelligence, 4(2), 61-80.

[14] Ghane, K. (2020, March). Big data pipeline with ML-based and crowd sourced dynamically created and maintained columnar data warehouse for structured and unstructured big data. In 2020 3rd International Conference on Information and Computer Technologies (ICICT) (pp. 60-67). IEEE.

[15] Furtado, P. (2016). Scalability and realtime on big data, MapReduce, NoSQL and Spark. In European Business Intelligence Summer School (pp. 79-104). Cham: Springer International Publishing.