



Evolution and Innovations of Healthcare Nosql Databases in Era of Big Data: A Systematic Review

Ahmed Ibrahim Sharqi

Faculty of Engineering -Islamic University Of Lebanon

ahmedibrahimsharqi@gmail.com

Abstract:

The rapid growth of data in digital age has necessitated advancements in database management systems, leading to the evolution of NoSQL databases. NoSQL databases have developed as a suitable substitute for relational databases, which are unable to store and analyze big data effectively. A specific NoSQL system's usefulness relies on the issue area, and there is a large variety of NoSQL systems available. This paper explores the landscape of NoSQL databases that have been used in healthcare application domains in order to ease the selection of appropriate NoSQL system for a specific purpose. To be more specific, the purpose of this study is to determine the breadth of healthcare big data analytics, as well as its features, kinds, applications, and concerns. In summary, the paper gives a concise survey of the available researches on healthcare big data exploitation utilizing NoSQL databases. Furthermore, it offers researchers a solid groundwork for future investigations within NoSQL healthcare contexts.

Keywords: Big data; NoSQL; Healthcare; Healthcare big data

تطور وابتكارات قواعد بيانات NoSQL للرعاية الصحية في عصر البيانات الضخمة: مراجعة منهجية

أحمد إبراهيم شرقي

كلية الهندسة - الجامعة الإسلامية في لبنان

ahmedibrahimsharqi@gmail.com

المخلص:

لقد أدى النمو السريع للبيانات في العصر الرقمي إلى ضرورة التقدم في أنظمة إدارة قواعد البيانات، مما أدى إلى تطور قواعد بيانات NoSQL. وقد تطورت قواعد بيانات NoSQL كبديل مناسب لقواعد البيانات العلائقية، والتي لا تستطيع تخزين وتحليل البيانات الضخمة بشكل فعال. تعتمد فائدة نظام NoSQL المحدد على مجال المشكلة، وهناك مجموعة كبيرة ومتنوعة من أنظمة NoSQL المتاحة. يستكشف هذا البحث مشهد قواعد بيانات NoSQL التي تم استخدامها في مجالات تطبيقات الرعاية الصحية من أجل تسهيل اختيار نظام NoSQL المناسب لغرض معين. وبصورة أكثر تحديداً، فإن الغرض من هذه الدراسة هو تحديد نطاق تحليلات البيانات الضخمة للرعاية الصحية، بالإضافة إلى ميزاتها وأنواعها وتطبيقاتها ومخاوفها. باختصار، يقدم البحث مسجلاً موجزاً للأبحاث المتاحة حول استغلال البيانات الضخمة للرعاية الصحية باستخدام قواعد بيانات NoSQL. وعلاوة على ذلك، فإنه يوفر للباحثين أساساً قوياً للتحقيقات المستقبلية في سياقات الرعاية الصحية NoSQL.

الكلمات المفتاحية: NoSQL : البيانات الضخمة؛ الرعاية الصحية؛ البيانات الضخمة للرعاية

1. Introduction

In recent years, large data applications have emerged, requiring the adoption of new data storage technologies in addition to regular databases. The analysis and, in particular, the storage of that amount of information is challenging. The rapid increase in the bulk of data relating to web technologies, mobile applications, and social media sites has increased the production of unstructured data from



petabytes to zettabytes, predicated on the fact that we live in the data ocean age [1]. In the last 25 years, data has grown dramatically in a variety of ways. Since 2010, the yearly volume of data collected has increased annually. Indeed, 90% of the world's data is thought to have been created in the previous two years alone. Global data generation, capture, copying, and consumption are predicted to rapidly increase, with a projected total of 64.2 zettabytes in 2020 as shown in Figure 1. It is anticipated that the amount of data created globally will surpass 181 zettabytes in 2025. The volume of data produced and replicated reached unprecedented levels in 2020. The surge was propelled by the heightened demand during the COVID-19 pandemic, as more people shifted to remote work and learning, as well as increased utilization of home entertainment options beyond initial projections. The number has grown from only 2 zettabytes in 2010 to an estimated 70x in just 14 years. It is anticipated that the 147 zettabytes produced in 2024 would rise in 2025 and reaching 181 zettabytes [2].

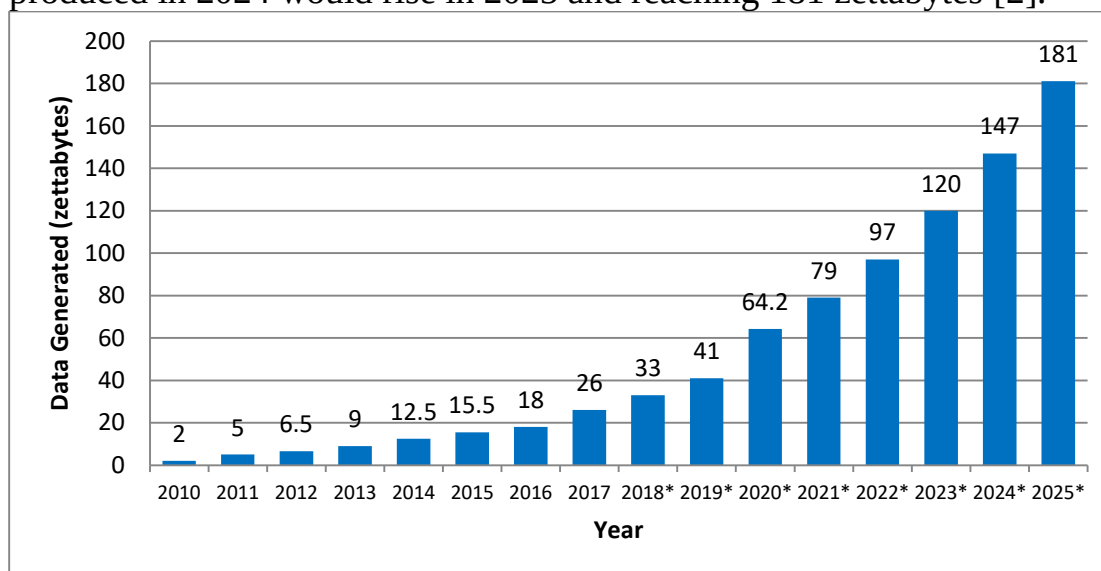


Figure 1: Volume of data generated, recorded, duplicated, and used globally between 2010 and 2020, with projections extending until 2025 [2]

The healthcare area in Big Data is mainly the collecting of patient electronic health records (EHRs) and other pertinent medical data. The goal is to deliver healthcare services by using historical data. Finding the right insight is made feasible by applying analytics and data science to huge data. Rather of striving to assemble all available data, data analytics focuses on the relevant data, giving the doctor with the necessary insights. However, there is still a disparity between what data is needed quickly and what is stored in big data. On the other hand, the vast volume of data may not always deliver the exact information that is required. Similarly, large data alone may not provide a clinician with accurate and meaningful insights [3].

For the purpose of storing data, there exist two main categories: relational databases and non-relational databases. These databases vary in their storage mechanisms, construction processes, and the types of data they are suited for.



Relational databases (RDBs), which are rooted in the relational model, were originally developed over 50 years ago to facilitate commercial data processing tasks. They serve as an excellent choice for storing a wide range of information, from financial records to personal data. The relational model, conceptualized by E.F. Codd at IBM in 1970, has been widely adopted for its simplicity and versatility in managing structured data [4]. Relational databases operate using Structured Query Language (SQL), which is utilized for executing interactive searches and manipulating data within these databases. Data within relational databases is organized into relations, containing of columns and rows. Each column holds specific data attributes, while each row represents a unique instance. Columns are also known as fields or attributes, and rows may be alternatively referred to as tuples or records. Every table within the database contains a unique identifier called a primary key, which is used to retrieve individual entries. Additionally, foreign keys, which are primary keys from other tables, establish connections between different tables. Due to their fixed schema, relational databases lack flexibility [5].

NoSQL databases, pioneered by Strozzi Carlo in 1998 [6], refer to non-relational databases designed as an alternative to traditional relational database systems that do not rely on SQL queries. They emerged as competitors to relational databases, boasting horizontal scalability and flexible structures, whether dynamic or non-dynamic. Given their need for horizontal scalability, NoSQL databases have become the preferred choice for handling large volumes of data in big data and cloud computing environments. Their scalability and capacity to manage vast amounts of data effectively position them as viable alternatives to relational databases. NoSQL databases, renowned for their traits like strong availability, optimized overall performance, and seamless scalability, have end up critical components of the infrastructure helping main organizations and on-line systems which includes Google, Twitter, Facebook, and Amazon [7].

Various methodologies were employed in the subject of healthcare huge records. Machine mastering techniques had been applied to anticipate epidemics and screen diseases, aiding within the early detection of ailments [8]. Data mining techniques had been deployed to forecast coronary heart disorder in its nascent degrees, exemplified in fitness analytics and the identity of epidemics [9,10]. Additionally, Natural Language Processing (NLP) techniques were leveraged to enhance the high-quality of care and precision of scientific selections, along with reducing costs and identifying high-danger elements [11, 12].

2. Systematic Review Method

The sources that provided the basis for our investigation are included in this part, together with the research questions that were created to support them and the query that were entered into the search strings.



2.1 The Source of Data

In order to locate research articles that answer our questions, a sufficient selection of electronic databases was selected prior to beginning the search process. Before the search even started, a sufficient selection of electronic databases was selected in order to identify research papers closely relevant to our work. The primary electronic resources utilized in this investigation were IEEE Xplore, Google Scholar, PubMed, and Springer Link. The keywords for the search phrases used in the online databases that we considered and the number of result papers are shown in Table 1.

Table 1: The search strings keywords

Search Keywords	Databases	Number of papers	Number of articles after filtering (2018-2024)
(NoSQL OR Non-relational databases) AND (Healthcare OR Medical) AND (Big data OR Big data analysis)	Google Scholar	21000	14400
	PubMed	20	11
	Springer	2988	2548
	IEEE Xplore	64	33

Since browsing sites often provide ability for direct filtering, in this work only articles published between 2018 and 2024 take in consideration, we were left with 23087 items overall after direct exclusion. Not all of these studies were included in this paper. The following four criteria were used to pick papers:

- (1) Articles on NoSQL approaches and big data in healthcare were considered.
- (2) Exclusion of articles published before 2018 (see Table 1).
- (3) Exclude books, chapters, masters and doctoral theses.
- (4) The articles written in the English language only take in consideration.

We retrieved 43 articles based on the previously mentioned criteria. We read each article in its entirety and conducted an analysis. The results of earlier investigations are analyzed and combined. The technique for sorting and selection is summed up and explained in Figure 2.

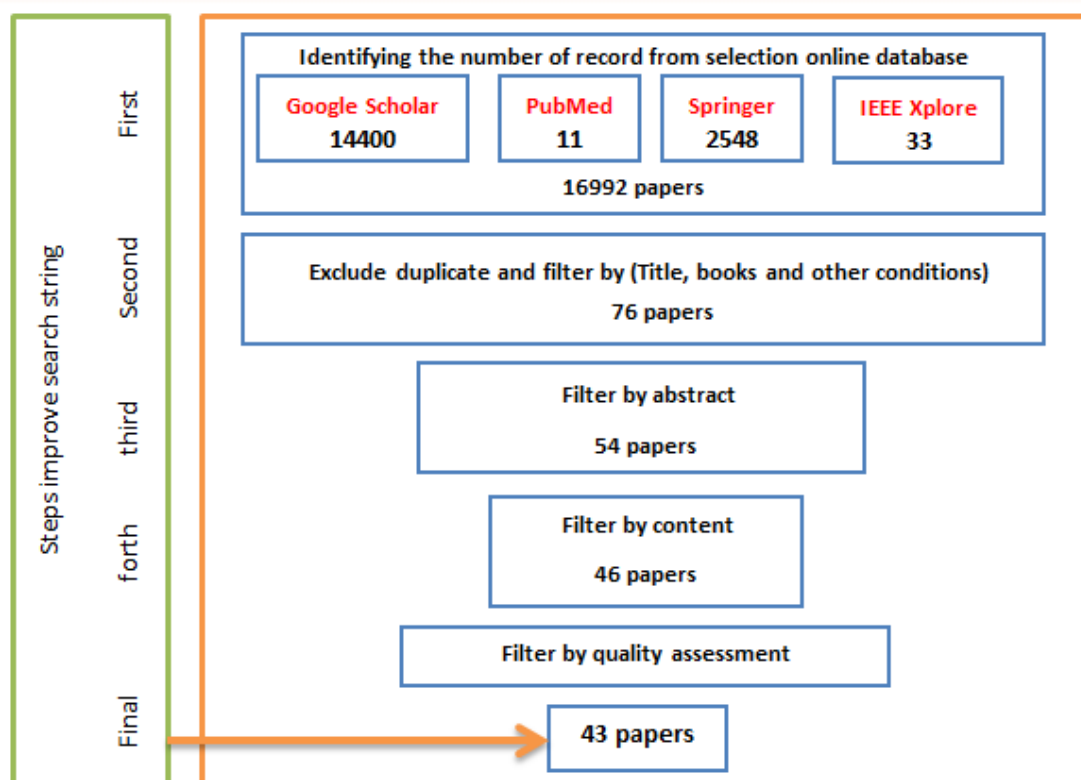


Figure 2: Search process strategy and filtering in this work

2.2 Research questions

In this work, a review of NoSQL methods used in the big data healthcare field has been addressed. Thus, the following sentences provide a summary of the main research questions:

Q1: Do datasets related to healthcare share the characteristics and features associated with big data? (answered in subsection 3.1)

Q2: What are the main big data types in healthcare applications? (answered in subsection 3.2)

Q3: Which NoSQL technologies are used in the academic literature to explicitly address the use of big data in healthcare? (answered in Section 4)

Q4: What challenges that have been noted by researchers in the literature when using big data in the healthcare industry? (Answered in Section 5).

3. Big Data Concepts

In the realm of information systems, the term "big data" has gained significant prominence in recent years. To the layperson, big data simply denotes massive volumes of data. However, within the field, big data encompasses more than just sheer volume—it refers to extensive and complex datasets resulting from the convergence of the well-known characteristics known as the Big Data V's. The term "big data" was initially coined in 1997 during a research study on the visualization of immense datasets presented at the Institute of Electrical and Electronics Engineers (IEEE) Conference on Display [13]. As shown in Figure 3, the phrase "big data" gained popularity in 2011. This chart displays Google



Trends searches [14] for Big Data modeling, which began to increase in intensity in 2011. Google Trends does not retain data prior to 2004, therefore searches made before then are not shown. In an attempt to better understand new paradigms for handling Big Data, academics have combined their efforts in recent years. In response to escalating demands, novel techniques for modeling and managing Big Data within databases have emerged.

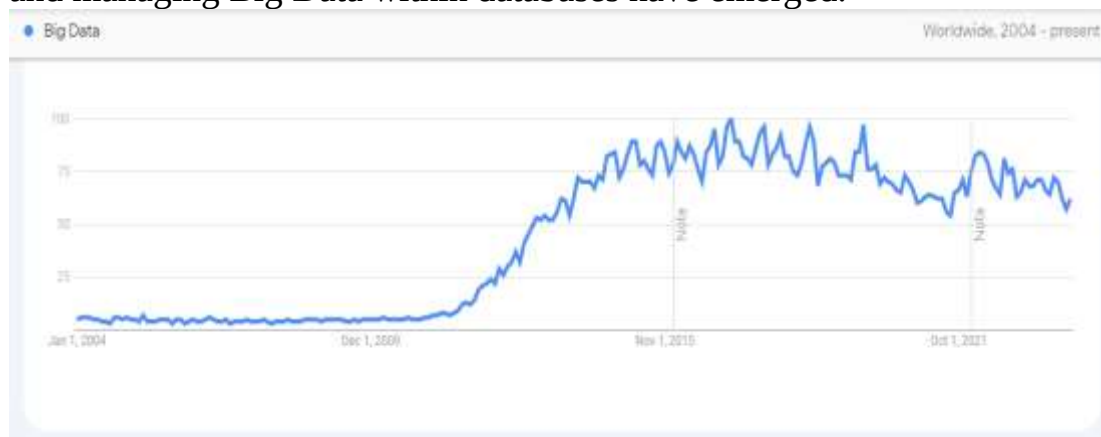


Figure 3: Big data term in Google trends

Healthcare data typically consists of various information types, such as electronic medical records (EMRs), which encompass patient medical histories, physician notes, clinical reports, biometric data, and other health-related details. The amalgamation of these diverse data sources results in what is commonly referred to as healthcare big data. The rise of healthcare big data represents a significant advancement that provides cost-effective solutions for both public and private healthcare sectors. Recent research endeavors highlight the importance of effective design and appropriate tool usage in managing large datasets within healthcare applications. This underscores the significance of utilizing big data analytics in healthcare systems [16].

3.1 Big Data Characterization

Initially, Big Data was defined not solely by its high volume but also by its velocity and variety, collectively known as the three Vs. of Big Data. As research progressed, additional characteristics such as value and veracity were recognized, expanding the concept. Presently, some authors propose models with 7 Vs. [17] or even 10 Vs. [18] to fully encapsulate the complexities of Big Data. For our study, we will focus on the primary 5 Vs. of Big Data, as depicted in Figure 4.



Figure 4: The 5 vs. of big data characteristics

In the same context, the 5Vs of Healthcare Big Data Characteristics could be are briefly stated as:

Volume: Healthcare data serves as a quintessential illustration within the realm of big data. The term "volume" denotes the sheer magnitude of data, which is expanding daily at an exponential pace. Projections indicate that the volume of big data is set to reach 181 zettabytes by 2025. The healthcare sector generates vast quantities of data, encompassing electronic medical records (EMRs), biometric data, clinical records, radiological images, genomic data, and more, surpassing many other industries in terms of data production. These diverse datasets collectively constitute big data in healthcare [19]. Consequently, healthcare organizations are increasingly turning to platforms like Hadoop and MongoDB due to their ability to store and analyze extensive datasets [20].

Velocity: Velocity is the pace at which data are produced and obtained from different healthcare systems [21].

Variety: Diversity and heterogeneity in data are referred to as variety. The pace at which data is generated and gathered by the healthcare sector from many sources, including social media, sensors, cameras, and cellphones, is astounding [22]. Nonetheless, the healthcare data might be semi-structured, unstructured, or structured in any combination [23]. Clinical data is one example of structured data; unstructured or semi structured data includes things like photographs, doctors' notes, mobile data, social media data, and radiological films [24, 25].

Veracity. In the context of healthcare data, the veracity characteristics pertain to the data's trustworthiness, which is synonymous with quality assurance. It provides a level of legitimacy about medical expertise [26, 27].



Value. Among the five virtues of big data in healthcare, value stands out as the most crucial and distinctive characteristic. The capability to convert healthcare data into actionable insights and valuable information is paramount [28]. The healthcare sector is concentrating on improving healthcare processes and operational efficiency as a result. By using more effective methods for service delivery, data analysis, administration, and integration, the latter seeks to reduce fraud, waste, and expenses. The former seeks to develop innovative methods for treating patients while effectively allocating healthcare resources [29, 30].

3.2 Big Data Types in healthcare application

The process of continuously monitoring a patient's health status generates a plethora of data. This medical data can be categorized into structured, semi-structured, and unstructured formats. Structured data includes standard Electronic Health Records (EHRs) [31, 32], while semi-structured data originates from various medical devices [33, 34]. Unstructured data, on the other hand, may consist of information such as biomedical photography [35].

Electronic Healthcare Records: EHRs are comprehensive digital repositories containing various aspects of a patient's medical history. These records encompass demographics, prescription details, diagnoses, laboratory findings, physician's notes, radiology reports, clinical data, and payment records. Together, these data describe the health status of the patient and make EHRs an invaluable resource for healthcare analytics. Additionally, EHRs facilitate data sharing amongst healthcare professionals [31,32].

Biomedical Images: When it comes to diagnosing illnesses and providing treatment, biomedical imaging is seen to be a very useful tool. However, handling these pictures is a difficulty as they include noisy information that needs to be removed to enable doctors to make precise conclusions [35].

Social networks: Data collection from social media, such as social networking sites, is necessary for social network analysis. The next stage is to extract information, including the identification of infectious diseases that may have an impact on healthcare predictive analysis. Social network data is often characterized by uncertainty, which makes using it to create prediction models dangerous [36,37].

Mobile Phone: Currently, mobile phone is one of the most widely used techniques products worldwide. In contrast to its early versions, mobile phones have evolved from being simple tools for communication to sophisticated gadgets with a wide range of functionality. Now days, they have many sensors installed, including cameras, accelerometers, and satellite location services [38]. The widespread usage and diverse features of mobile phones make them ideal for collecting health data. This capability has led to the development of numerous effective healthcare apps, including those for heart rate monitoring, pregnancy tracking, and child nutrition management [39, 40].

Sensing Data: Healthcare monitoring solutions rely on a range of sensors to track various aspects of a patient's well-being. These devices play a crucial role in monitoring health by measuring essential medical parameters such as the temperature of body, the pressure of blood, and heart rate [41]. To ensure effective health monitoring, a patient's environment may become filled with devices like microphones, pressure sensors, and surveillance cameras. Consequently, the amount of data generated by health monitoring systems tends to rise substantially, requiring the implementation of advanced processing techniques during data processing [39].

4. Types of NoSQL

NoSQL databases offer much functionality such as handling diverse data types, supporting horizontal scalability, processing data concurrently across multiple servers, and accommodating extensive data storage needs. The rise of big data and cloud computing has brought about significant changes in data storage requirements in recent times. The rapid expansion of structured, semi-structured, and unstructured data highlights the necessity for innovative storage solutions in the era of big data. The emergence of big data has posed new challenges in terms of data gathering and storage. NoSQL databases set themselves apart from relational databases primarily based on their data model. NoSQL encompasses various categories, including document stores, column stores, key-value stores, and graph stores, as illustrated in the Figure 5 [17].

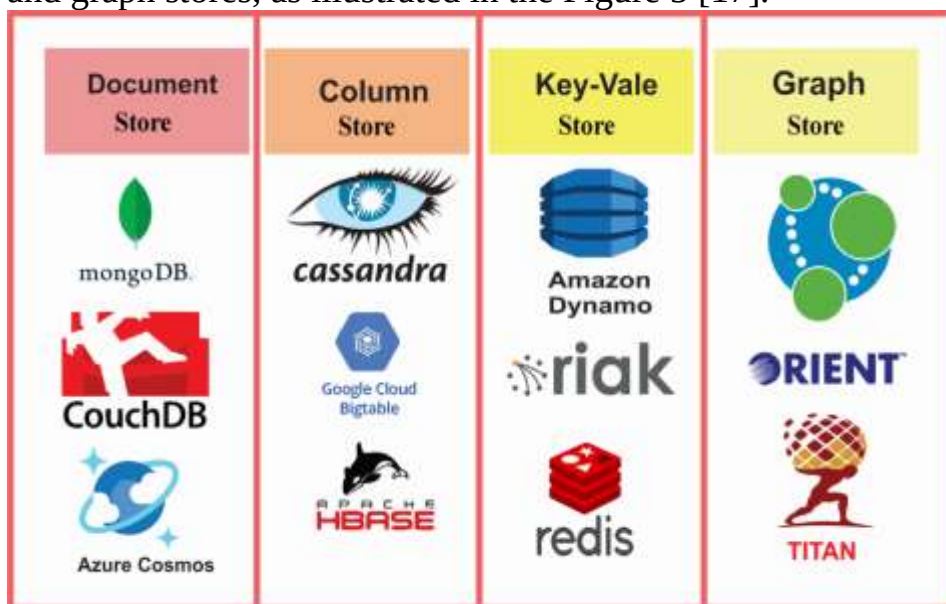


Figure 5: Types of NoSQL database [17]

A summary of NoSQL storage tools is described in Table 2. The table highlights summary with respect to the advantages and disadvantages of the systems and the main applications of NoSQL databases for big data storage technologies [42-44].



Table 2: Summary of NoSQL storage tools

Data Model	Systems & Tools	Advantages	Disadvantages	Applications
Document Store	MongoDB	Document-oriented data storage, scalability, flexibility.	Not suitable for complex transactions requiring multi-document transactions	Content management, mobile applications, real-time big data processing
	Couchbase	High performance, in-memory processing, JSON document storage	Limited support for complex queries	Caching, real-time analytics, mobile applications
Column Store	Cassandra	High write and read throughput, fault tolerance, linear scalability	Complex configuration, may not be ideal for small-scale deployments	Time-series data, IoT applications, real-time analytics
	HBase	Scalability, fault tolerance, handling large datasets	Limited support for ad-hoc queries	Time-series data, data warehousing, analytics
Key value	Redis	In-memory data storage, low latency, high-throughput	Limited support for complex queries, persistence can be configured	Caching, session storage, message queuing.
	Riak	High availability, fault tolerance, distributed systems	May have a steeper learning curve for some users	Distributed systems, fault-tolerant applications, real-time analytics
Graph Store	Neo4j	Graph database, relationships between data entities	May not be ideal for non-graph data models	Social networks, fraud detection, recommendation engines
	HyperGraphDB	Hyper graph representation, handling complex relationships, versatile data modeling	May have a learning curve due to its unique data model	Complex application relationships, semantic data modeling, knowledge representation, and scenarios where relationships have attributes

4.1 Document Store



All records and the data they are related with are kept in a single document in a document-oriented data store. A document database example is shown in Figure 6. Data is saved in a variety of forms, making this kind of data repository flexible. It analyzes vast amounts of data, facilitates horizontal scaling, and offers more visible execution. Standard formats including XML, PDF, JSON, and others are supported by the document-oriented data storage [44]. Regarding the capability of healthcare data, the authors in [21] provide a document-oriented approach in light of NoSQL databases. Using the dispersed attribute, the data was appropriated on the database. The model produced findings for the amount of time needed to retrieve records under standalone and distributed databases, as well as the quantity of records obtained, based on the observation. The model was developed in Java, and the B-tree and hashing algorithms were used to verify the retrieval efficiency. The outcome demonstrates that the model performs well in terms of the amount of time needed to get records for standalone databases. For deployment on cloud servers that are only partially trustworthy, the researchers in [30] suggested a novel searchable system for encrypted personal health records (PHRs) on MongoDB. Nearly all query operations provided by plaintext MongoDB are supported by this technique, with multi-dimensional, multi-keyword searches with range queries being particularly well-supported. The plan may accomplish safe and useful searchable encryption for PHR systems, according to experimental findings based on the time of range query. A technique for creating a NoSQL database's schema is proposed in [45] using a relevant ontology, a sample query set, some statistical data about the queries, and the query set's performance requirements. The authors' core healthcare domain ontology served as the basis for the creation of a schema intended for a MongoDB data store. The usefulness of the suggested MongoDB technique is shown by contrasting the performance of the suggested architecture with a relational model created for the same primary healthcare data. In their study [46], researchers introduce an innovative technique known as the Event-level Inverted Index (ELII), which relies on document stores. ELII is designed to optimize the time trade-offs between batch preprocessing and ongoing, user-defined temporal searches. The researchers implemented an experimental temporal query engine using the ELII approach within a NoSQL database. Notably, this approach demonstrated near-real-time performance when applied to critical Covid-19 electronic health records (EHRs) data, with 3.76 billion records and million unique patients 1.3 inclusive. The ELII measures performance in three types of questions: absolute time, relative time, and classical (non-time). In contrast to the conventional method without the use of ELII, the experimental findings show considerable improvement. ELII accomplished query completion within seconds, exhibiting average speed improvements of 26.8 times for relative temporal queries, 88.6 times for absolute temporal queries, and a remarkable 1037.6 times for conventional queries. Additionally, another investigation detailed in [47] centers on creating



an Integrated Data Repository of Genetic Disorders Data (GENE2D). This project aims to create GENE2D by developing and deploying a NoSQL-based integration framework. The objective of this framework is to centralize data from diverse genetic clinics and research institutions throughout Saudi Arabia. The backend database utilized by the IDR is MongoDB, which drives a NoSQL document store. The GENE2D technology exhibits potential for expanding and reinforcing Saudi Arabia's national genetic diseases database.

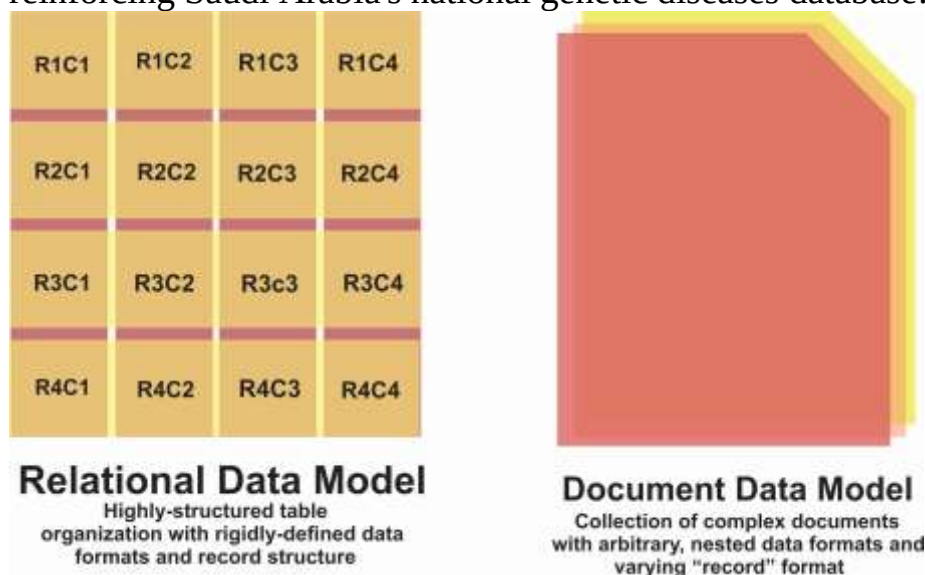


Figure 6: An example of document database [17]

4.2 Column Store

With column-store systems, a database is divided entirely vertically into a group of distinct columns that are kept apart for storage as shown in Figure 7. These column-based systems allow queries to read just the properties they want, as opposed to reading full rows from disk and discarding unnecessary attributes once they are in memory, since each column is stored independently on disk. A comparable advantage occurs for data transfers from main memory to CPU registers, resulting in better overall memory and I/O bandwidth utilization. Overall, there are a lot of novel database structures that can be created by pushing the column-oriented method to its extreme [6]. The researchers in [48] discussed the properties of big data in the healthcare system, with a special emphasis on NoSQL databases. Using HBase as a NoSql on top of the HADOOP platform, they presented an architectural model and demonstrated how query execution speed varies depending on the amount of data stored in HBase. The findings indicate that although there is not a significant difference in performance between the distributed and standalone or pseudo-distributed modes, the distributed mode offers more storage support. Furthermore, the authors of [49] present a sophisticated proof of concept for a healthcare project in Canada. The project involves loading 30 terabytes of healthcare data into an HBase cluster, with MapReduce responsible for handling all fundamental processing tasks. The authors illustrate the significant limitations of the MapReduce method, highlighting that data ingestion and loading using HBase



are time-consuming processes, taking approximately a month to complete. Additionally, they emphasize the challenges associated with creating a schema for complex healthcare data.

Column family of location			Column family of vitals					
MRN	City	Prov.	BP	Temp	Doctor	LT	Surg.	Satis.
456-2	KHI	SDH			John Dave	T1, T2	S1	
224-3			119/78	100				High
123-9	LHR	PJB				T65		Low
874-3			110/70	99				
546-3					Jack Sleeve	T43		
								Med.
445-7	ISB	Punjab	156/99				S3	
678-3			178/90		Sara		S4	
998-4								Low
387-5						T1, T2, T3		

Figure 7: column-oriented database example [23].

4.3 Key-Value Store

Key-value pairs are used to store data in this most basic kind of data storage. Figure 8 shows an example of key-value store, each record has a unique key, and the value corresponds to the data that is actually saved for that key. A string type variable is used to hold the key, and any kind of data may be used for the value. This kind of NoSQL is really basic in its approach to data management and is mostly similar to relational databases in terms of functionality. The most important advantage is that key-value storage is available which is flexible, schema-free, and easy to understand. The most useful databases are available to us when we need to retrieve information such as a user's favorites or to store a user's time or shopping cart [7]. Researchers in [50] are working on a fix to the problem with centralized storage. They think localized data might be the answer. So far, they've been able to use OrbitDB and the Interplanetary File System (IPFS) to store electronic patient health records in a distributed format. But they didn't stop there. On top of that, they built a blockchain network using Hyperledger Fabric and brought in Hyperledger Composer so that it can manage data hashes and access control upon receipt. The researchers believe this platform could make healthcare systems more resilient while also addressing smart healthcare's inherent security issues that come with high performance. Through benchmarks made with Hyperledger Caliper, they were able to show off its increased privacy capabilities too. In [51], we talked about how a clinical management system was created so FM data queries work faster while also visualizing FM performance indicators. Using Big Data Analytics (BDA), Building Information Modeling (BIM), and Not Only SQL (NoSQL) databases, the system was designed using the Design Science Research methodology. This

system aims to streamline healthcare facility management processes by leveraging advanced analytics and database technologies. Results demonstrate that the system is useful and efficient for practitioners in healthcare facility management to undertake FM data retrieval and analysis. The use of Dynamo with BIM will help to increase Building Information Modeling BIM used in FM.

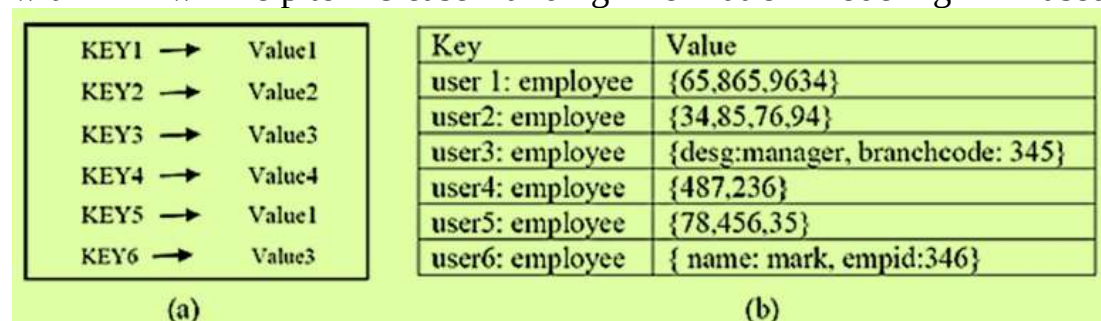


Figure 8: (a) Key--value store (KVS) data model, (b) Data in KVS [5].

4.4 Graph Store

The graph data model is the fundamental algorithm of Graph Database. The relationships between things may have more significance in some applications than the actual entities themselves. Both dynamic and fixed relationships are possible. Airlines and the oil and gas industry are just two of the numerous sectors using graph data models. Titan, HyperGraphDB, and Neo4J databases are examples of those constructed using graph data models. An example of a graph-based database that demonstrates how the data are connected and correlated is shown in Figure 9. The relationship between the data is what matters most in graph databases. For storing unstructured and semi-structured data, it provides efficient, schema-free storage. The primary challenge with graph databases is to accomplish sharing [15]. A multilayer graph has been used in [52] to represent the patient-doctor connection. Graph database Neo4j (NoSQL) is used to evaluate the multilayer patient-doctor graph data model. A relational model is also included in this study and contrasted with the multilayer graph model. The findings demonstrate that, in terms of data access, the graph data model performs better than the relational model because of the intricacy and irregularity of the patient-doctor interactions. Researchers in [53] compared the relational database PostgreSQL with the NoSQL Neo4j to see how well graph databases could handle health data. The investigators used a publicly accessible dataset from an American hospital. The collection includes information on hospital admissions, diagnosis, test results, and medication information for people with diabetes, among other topics. This two-dimensional tabular information is first modeled using an Entity-Relationship Diagram (ERD), which is constructed on a relational database. The Entity-Relationship Diagram (ERD) is subsequently transformed into a schema suitable for a graph database and deployed on a NoSQL graph database platform. The main findings show that Neo4j works better than PostgreSQL, yet there were some minor data input errors. In the sane context, Platforms for Healthcare Social Networks



(HSNs) are starting to appear in a reasonable environment, with the aim of improving patient care and education in [54]. The goal of this study is to demonstrate that managing large amounts of data with a vast number of relationships can be handled by a NoSQL graph database management system. The outcomes of the experiments demonstrate that Neo4j makes managing HSN data easier while also ensuring that findings are appropriate from the standpoint of future social science research.

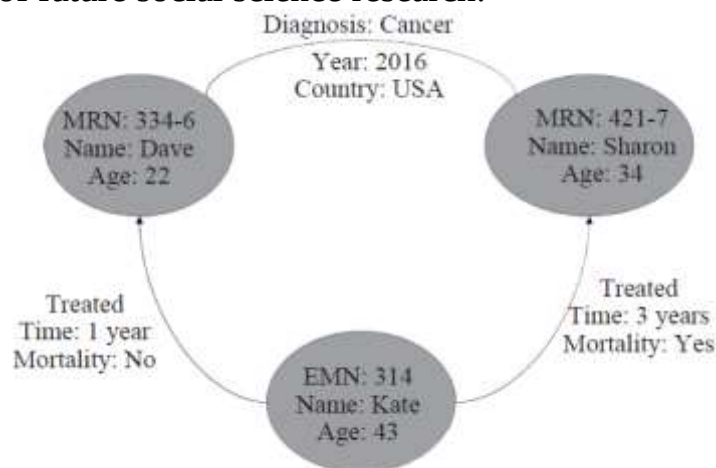


Figure 9: Graph store example [23]

5. Challenges of Big Data in Healthcare Domain

We identified five problems that the healthcare sector faces when using big data applications. We characterize them as follows:

Confidentiality: The patient may have problems with their job and/or insurance coverage as a result of the theft of their medical records [21,55]. Healthcare data is highly sensitive, containing personal and often identifiable information. Maintaining patient privacy and ensuring data security are paramount concerns, especially with the increasing number of cyber threats. Breaches can lead to severe consequences for patients and healthcare organizations, including financial loss and reputational damage. A more adaptable framework that might be useful in the adoption of big data technologies could be provided by legislative reforms pertaining to data security and confidentiality. Nevertheless, with the adoption of BDA in the data-driven health industry, analytics on sensitive patient care data remains a concern [56].

Granular Access Control : The duties, privileges, rights, and roles of hospital medical users and patients may be defined with granular access control so that hospital-affiliated users are only granted access to the relevant data or system functional areas [57]. One problem with BDA applications for healthcare that is often mentioned is ensuring a high degree of accessibility and security to access pertinent data [58].

Interoperability: Achieving interoperability remains a significant challenge in healthcare, as disparate systems often fail to communicate and exchange data effectively. Lack of interoperability leads to fragmentation of patient information, duplication of tests, and inefficiencies in care delivery. Initiatives



such as Fast Healthcare Interoperability Resources (FHIR) plan to encourage smooth communication across healthcare systems and standardize the formats for data sharing [59].

Economic Challenges: During clinical visits, the facilities in medical sector that connect healthcare and patients providers are dependent on payment. The development of technology related to this procedure thus burdens the medical community and has an unwarranted negative effect on the staff about such unpaid services [60].

Resource Constraints: Limited resources, including budgetary constraints, shortage of skilled personnel, and technological barriers, pose challenges to leveraging big data effectively in healthcare. Healthcare organizations must invest in infrastructure, training, and partnerships to overcome resource constraints and harness the full potential of big data analytics for improving patient outcomes and operational efficiency [58,61].

6. Discussion

In this section, we will explore the findings related to our research questions. Addressing the first question, it is evident that healthcare datasets possess the fundamental attributes of big data, encompassing volume, velocity, variety, veracity, and value. The examination of these datasets emphasizes the significance of employing data analytics within contemporary healthcare practices. Through the efficient utilization of big data, healthcare institutions have the opportunity to refine decision-making processes, streamline operations, and ultimately elevate patient care outcomes. For the second question, there are plenty of sources when it comes to big data in healthcare included EHRs, biomedical images, social networks, mobile phones and sensing data. Answering the third question is all about the types of NoSQL that have been used in big data healthcare applications and we found document storage MongoDB database and spreadsheet storage Neo4j as the most commonly used systems in healthcare applications. Lastly, the response to the fourth question enumerates the difficulties encountered when analyzing substantial amounts of health data. These include worries about patient privacy and data security, the need of fine-grained access controls to guarantee that the right people have access to the right data, and the continuous battle to make disparate systems work together. These problems are made worse by financial constraints, such as the strain that new technology place on hospitals and employees. Furthermore, a shortage of skilled staff and inadequate funding prevent big data in healthcare from reaching its full potential.

7. Conclusion

To sum up, when we look at health-related data, we can see that it lines up with all of the usual traits of big data; it has volume, velocity, variety, veracity and value characteristics. The paper also identifies the main categories of big data that are often used in healthcare applications, including biological pictures,



social network data, mobile phone data, electronic health records, and sensor data. An examination of academic literature also identifies several NoSQL technologies and offers insights into their efficacy in addressing big data concerns in the healthcare industry. Lastly, a review of the difficulties identified by the researchers emphasizes the difficulties in using big data in the healthcare sector, including problems with privacy, accessibility, coordination, budgetary limitations, and resource scarcity. To overcome these obstacles and achieve the potential of big data to revolutionize health care delivery and enhance patient outcomes at the appropriate location, coordinated efforts, technical advancements, and legislative reforms are needed.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- Khan, Afreen, and B. M. Ubaidullah. "Critical Review on Threat Model of Various NoSQL Databases." In V: 2017 4th International Conference on Computing for Sustainable Global Development, vol. 40353.
- Petroc Taylor "Amount of data created, consumed, and stored 2010-2020, with forecasts to 2025" Nov 16, 2023 online available <https://www.statista.com/statistics/871513/worldwide-data-created/>
- Pramanik, Pijush Kanti Dutta, Saurabh Pal, and Moutan Mukhopadhyay. "Healthcare big data: A comprehensive overview." Research anthology on big data analytics, architectures, and applications (2022): 119-147.
- Agarwal, S., Rajan, K.S., "Analyzing the performance of NoSQL vs. SQL databases for Spatial and Aggregate queries", Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings. 2017, p. 4. Available at <https://scholarworks.umass.edu/foss4g/vol17/iss1/4/>;
- Ramzan, Shabana, Imran Sarwar Bajwa, Rafaqut Kazmi, and Amna. "Challenges in NoSQL-based distributed data storage: a systematic literature review." Electronics 8, no. 5 (2019): 488.
- Strozzi, C. "Nosql-a relational database management system". Lainattu 1998, 5, 2014.
- Moniruzzaman, A.; Hossain, S.A. "Nosql database: New era of databases for big data analytics-classification", characteristics and comparison. arXiv 2013, arXiv:1307.0191.
- M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," IEEE Access, vol. 5, pp. 8869–8879, 2017.
- T. U. Mane, "Smart heart disease prediction system using improved K-means and ID3 on big data," in *Proceedings of the 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, IEEE, Pune, India, pp. 239–245, February 2017.
- K. Rahimi, D. Bennett, N. Conrad et al., "Risk prediction in patients with heart failure," JACC: Heart Failure, vol. 2, no. 5, pp. 440–446, 2014.



- Y. Wang, L. Kung, and T. A. Byrd, "Big data analytics: understanding its capabilities and potential benefits for healthcare organizations," *Technological Forecasting and Social Change*, vol. 126, pp. 3–13, 2018.
- N. Mehta, A. Pandit, and S. Shukla, "Transforming healthcare with big data analytics and artificial intelligence: a systematic mapping study," *Journal of Biomedical Informatics*, vol. 100, p. 103311, 2019.
- Cox, M.; Ellsworth, D. "Application-controlled demand paging for out-of-core visualization". In *Proceedings of the 8th IEEE Conference on Visualization*, Phoenix, AZ, USA, 24 October 1997; pp. 235–244.
- Google. Google Trends. Available online: <https://trends.google.es/trends/explore?date=all&q=%22big%20data%22> (accessed on 23 January 2024).
- M. S. Hadi, A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Patient-centric cellular networks optimization using big data analytics," *IEEE Access*, vol. 7, pp. 49279–49296, 2019.
- A. Jindal, A. Dua, N. Kumar, A. K. Das, A. V. Vasilakos, and J. J. P. C. Rodrigues, "Providing healthcare-as-a-service using fuzzy rule based big data analytics in cloud computing," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1605–1618, 2018.
- Ali, Aqib, Samreen Naeem, Sania Anam, and Muhammad M. Ahmed. "A state of art survey for big data processing and nosql database architecture." *International Journal of Computing and Digital Systems* 14, no. 1 (2023): 1-1.
- Martinez-Mosquera, Diana, Rosa Navarrete, and Sergio Lujan-Mora. "Modeling and management big data in databases—A systematic literature review." *Sustainability* 12, no. 2 (2020): 634.
- A. S. Panayides, M. S. Pattichis, S. Leandrou, C. Pitris, A. Constantinidou, and C. S. Pattichis, "Radiogenomics for precision medicine with a big data analytics perspective," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 5, pp. 2063–2079, 2018.
- Dubey, Sonia, and Aditya Saxena. "Evaluation Of Performance of NoSQL Database Management System:-MongoDB And Relational Database Management System:-PostgreSQL for Efficient Management Of Healthcare Data." *Journal of Survey in Fisheries Sciences* 10, no. 1S (2023): 7153-7157.
- Gayathiri, N. R., D. David Jaspher, and A. M. Natarajan. "Big health data processing with document-based Nosql database." *Journal of Computational and Theoretical Nanoscience* 15, no. 5 (2018): 1649-1655.
- Benhlila, Laila. "Big data management for healthcare systems: architecture, requirements, and implementation." *Advances in bioinformatics* 2018 (2018).
- Imran, Sohail, Tariq Mahmood, Ahsan Morshed, and Timos Sellis. "Big data analytics in healthcare— A systematic literature review and roadmap for practical implementation." *IEEE/CAA Journal of Automatica Sinica* 8, no. 1 (2020): 1-22.



Tomar, Dimpal, Jai Prakash Bhati, Pradeep Tomar, and Gurjit Kaur. "Migration of healthcare relational database to NoSQL cloud database for healthcare analytics and management." In Healthcare Data Analytics and Management, pp. 59-87. Academic Press, 2019.

Ranchal, Rohit, Paul Bastide, Xu Wang, Aris Gkoulalas-Divanis, Maneesh Mehra, Senthil Bakthavachalam, Hui Lei, and Ajay Mohindra. "Disrupting healthcare silos: Addressing data volume, velocity and variety with a cloud-native healthcare data ingestion service." IEEE Journal of Biomedical and Health Informatics 24, no. 11 (2020): 3182-3188.

Ambigavathi, M., and D. Sridharan. "Big data analytics in healthcare." In 2018 tenth international conference on advanced computing (ICoAC), pp. 269-276. IEEE, 2018.

Kumari, Sandhya. "Big data analytics for healthcare system." In 2018 IADS international conference on computing, communications & data engineering (CCODE), pp. 7-8. 2018.

Kumar, Sunil, and Maninder Singh. "Big data analytics for healthcare industry: impact, applications, and tools." Big data mining and analytics 2, no. 1 (2018): 48-57.

Mostajabi, Faezeh, Ali Asghar Safaei, and Amir Sahafi. "A Systematic Review of Data Models for the Big Data Problem." IEEE Access 9 (2021): 128889-128904.

Chen, Lanxiang, Nan Zhang, Hung-Min Sun, Chin-Chen Chang, Shui Yu, and Kim-Kwang Raymond Choo. "Secure search for encrypted personal health records from big data NoSQL databases in cloud." Computing 102 (2020): 1521-1545.

Dash, Sabyasachi, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. "Big data in healthcare: management, analysis and future prospects." Journal of big data 6, no. 1 (2019): 1-25.

Wang, Yichuan, LeeAnn Kung, William Yu Chung Wang, and Casey G. Cegielski. "An integrated big data analytics-enabled transformation model: Application to health care." Information & Management 55, no. 1 (2018): 64-79.

Mehta, Nishita, and Anil Pandit. "Concurrence of big data analytics and healthcare: A systematic review." International journal of medical informatics 114 (2018): 57-65.

Ristevski, Blagoj, and Ming Chen. "Big data analytics in medicine and healthcare." Journal of integrative bioinformatics 15, no. 3 (2018): 20170030.

Tobore, Igbe, Jingzhen Li, Liu Yuhang, Yousef Al-Handarish, Abhishek Kandwal, Zedong Nie, and Lei Wang. "Deep learning intervention for health care challenges: some biomedical domain considerations." JMIR mHealth and uHealth 7, no. 8 (2019): e11966.

Ali, Farman, Shaker El-Sappagh, SM Riazul Islam, Amjad Ali, Muhammad Attique, Muhammad Imran, and Kyung-Sup Kwak. "An intelligent healthcare



- monitoring framework using wearable sensors and social networking data." *Future Generation Computer Systems* 114 (2021): 23-43.
- Guo, Chonghui, and Jingfeng Chen. "Big data analytics in healthcare." In *Knowledge Technology and Systems: Toward Establishing Knowledge Systems Science*, pp. 27-70. Singapore: Springer Nature Singapore, 2023.
- Jagadeeswari, V., V. Subramaniaswamy, R. Logesh, and Varadarajan Vijayakumar. "A study on medical Internet of Things and Big Data in personalized healthcare system." *Health information science and systems* 6 (2018): 1-20.
- Ngiam, Kee Yuan, and Wei Khor. "Big data and machine learning algorithms for health-care delivery." *The Lancet Oncology* 20, no. 5 (2019): e262-e273.
- Manogaran, Gunasekaran, Ramachandran Varatharajan, Daphne Lopez, Priyan Malarvizhi Kumar, Revathi Sundarasekar, and Chandu Thota. "A new architecture of Internet of Things and big data ecosystem for secured smart healthcare monitoring and alerting system." *Future Generation Computer Systems* 82 (2018): 375-387.
- A. Kos and A. Umek, "Wearable sensor devices for prevention and rehabilitation in healthcare: swimming exercise with real-time therapist feedback," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1331–1341, 2018.
- G. Harerimana, B. Jang, J. W. Kim, and H. K. Park, "Health big data analytics: a technology survey," *IEEE Access*, vol. 6, pp. 65661–65678, 2018.
- S. Shafqat, S. Kishwer, R. U. Rasool, J. Qadir, T. Amjad, and H. F. Ahmad, "Big data analytics enhanced healthcare systems: a review," *7e Journal of Supercomputing*, vol. 76, no. 3, pp. 1754–1799, 2018.
- Sen, Poly Sil, and Nandini Mukherjee. "An ontology-based approach to designing a NoSQL database for semi-structured and unstructured health data." *Cluster Computing* (2023): 1-18.
- Huang, Yan, Xiaojin Li, and Guo-Qiang Zhang. "ELII: A novel inverted index for fast temporal query, with application to a large Covid-19 EHR dataset." *Journal of Biomedical Informatics* 117 (2021): 103744.
- Samra, Halima, Alice Li, and Ben Soh. "Gene2d: A nosql integrated data repository of genetic disorders data." In *Healthcare*, vol. 8, no. 3, p. 257. MDPI, 2020.
- Mondal, Sukhendu S., Somen Mondal, and Sudip Kumar Adhikari. "Performance Analysis of Healthcare Information in Big Data NoSql Platform." In *Doctoral Symposium on Intelligence Enabled Research*, pp. 235-247. Singapore: Springer Nature Singapore, 2022.
- Chrimes, Dillon, M. H. Kuo, A. W. Kushniruk, and B. Moa. "Interactive big data analytics platform for healthcare and clinical services." *Global Journal of Environmental Sciences* 1, no. 1 (2018).
- Zaabar, Bessem, Omar Cheikhrouhou, Faisal Jamil, Meryem Ammi, and Mohamed Abid. "HealthBlock: A secure blockchain-based healthcare data management system." *Computer Networks* 200 (2021): 108500.



Lee, DonHee. "Effects of key value co-creation elements in the healthcare system: focusing on technology applications." *Service Business* 13, no. 2 (2019): 389-417.

Mondal, Safikureshi, Anwesha Basu, and Nandini Mukherjee. "Building a trust-based doctor recommendation system on top of multilayer graph database." *Journal of Biomedical Informatics* 110 (2020): 103549.

Turhan, Sultan N. "Leveraging Graph Databases for Enhanced Healthcare Data Management: A Performance Comparison Study." In *2023 IEEE International Conference on Big Data (BigData)*, pp. 5007-5013. IEEE, 2023.

Celesti, Antonio, Alina Buzachis, Antonino Galletta, Giacomo Fiumara, Maria Fazio, and Massimo Villari. "Analysis of a NoSQL graph DBMS for a hospital social network." In *2018 IEEE symposium on computers and communications (ISCC)*, pp. 01298-01303. IEEE, 2018.

Wang, Xiaoming, Carolyn Williams, Zhen Hua Liu, and Joe Croghan. "Big data management challenges in health research—a literature review." *Briefings in bioinformatics* 20, no. 1 (2019): 156-167.

Hong, Liang, Mengqi Luo, Ruixue Wang, Peixin Lu, Wei Lu, and Long Lu. "Big data in health care: Applications and challenges." *Data and information management* 2, no. 3 (2018): 175-197.

Adam, Khalid, Mohammed Adam Ibrahim Fakharaldien, Jasni Mohamed Zain, Mazlina Abdul Majid, and Ahmad Noraziah. "Bigdata: Issues, challenges, technologies and methods." In *Proceedings of the International Conference on Data Engineering 2015 (DaEng-2015)*, pp. 541-550. Springer Singapore, 2019.

Arshad, Muhammad, M. Nawaz Brohi, Tariq Rahim Soomro, Taher M. Ghazal, Haitham M. Alzoubi, and Muhammad Alshurideh. "NoSQL: Future of BigData Analytics Characteristics and Comparison with RDBMS." In *The Effect of Information Technology on Business and Marketing Intelligence Systems*, pp. 1927-1951. Cham: Springer International Publishing, 2023.

Dullabh, P. M., Sondheimer, N. K., Katzan, I. L., Tierney, W. M., & Einstadter, D. (2019). Quantifying Health Information Exchange 'Hub' Costs in a Real-World Setting. *Applied Clinical Informatics*, 10(3), 446–454. <https://doi.org/10.1055/s-0039-1695765>

Shastri, Apoorva, and Mihir Deshpande. "A review of big data and its applications in healthcare and public sector." *Big data analytics in healthcare* (2020): 55-66.

Almassabi, Ahmed, Omar Bawazeer, and Salahadin Adam. "Top NewSQL databases and features classification." *International Journal of Database Management Systems* 10, no. 2 (2018): 11-31.

Cordeiro, João Rala, and Octavian Postolache. "Big data storage for a health predictive system." In *2018 International Symposium in Sensing and Instrumentation in IoT Era (ISSI)*, pp. 1-6. IEEE, 2018.