MANAGING BIG DATA IN KUWAITI HEALTHCARE SYSTEM: TECHNIQUES AND CHALLENGES

Abdullah F Alenezi

Assistant Professor

College of business studies

Public Authority for applied education and training

Author Email: Af.alenezi@paaet.edu.kw

Abstract

This paper delves into the complexities of implementing and manipulating Big Data Analytics in Kuwaiti with a main emphasis on the healthcare sector. It underscores the significant potential that big data holds for the country and highlights the considerable challenges involved in its adoption and manipulation. It indicated that Kuwaiti healthcare sector has applied few practical tools like ETL processes, data visualization, optimization techniques and robust security measures as ways in data manipulation. Nevertheless, this study identified primary obstacles including a lack of knowledge about the technology, difficulties understanding diverse data sets, concerns over data storage, integration challenges from various sources and financial constraints. However, these challenges are not insurmountable. Proposed solutions range from extensive training programs and workshops to infrastructural investments and robust financial support. The paper emphasizes the importance of skilling up the current workforce providing them with essential resources and creating an

informed ecosystem. Furthermore, the government's commitment to technological advancement demonstrated through budget allocations and digital transformation initiatives is seen as a promising sign for the future. The recommendations provided aim to guide organizations and policymakers in harnessing big data effectively for the betterment of Kuwait's industries especially healthcare.

1. Introduction

Big data is commonly understood as an extensive collection of structured, semi-structured, and unstructured data. Organizations harness this data to derive meaningful insights, which are then applied in machine learning, predictive modelling, and other sophisticated analytical methods. Modern entities consider big data an essential asset for its role in evidence-based decision-making, particularly in areas like advertising, marketing, and strategic planning. By leveraging big data, businesses can improve customer service, craft personalized marketing strategies, and, as a result, potentially boost revenue. Those organizations adept at harnessing the power of big data often gain a competitive advantage. They have an enhanced understanding of market dynamics, from evolving consumer behaviors to prevailing market demands. In healthcare, big data facilitates the efficient management of patient records, hospital logistics, and information from various health technologies, with the stored data proving invaluable for future use. This paper seeks to highlight the importance and potential of big data in Kuwait. It will delve into the associated challenges of its adoption and administration, especially in the country's business and industrial sectors. Globally, big data's influence is evident with the market witnessing notable growth, increasing by 20% in 2018, valued at around $42 billion (Yeung, 2017). Kuwait has embarked on a journey of digital transformation, aiming for a shift towards intelligent e-services.

From recovering from wartime devastations to actively pursuing digitalization, the nation has made commendable strides. The Kuwaiti government's push for online platforms not only aids in digital transformation but also paves the way for a more connected and efficient administration. A notable change is evident in the industrial sector, which has benefited from big data analytics, enhancing its operational capacities. These endeavors align Kuwait with global standards, promoting its stature regionally and internationally (Vassakis et al., 2018). The onset of the COVID-19 pandemic accelerated the adoption of digital solutions, including big data analytics and artificial intelligence. These technologies have facilitated sectors like health and industry in their transformative journey. Governmental efforts in integrating AI and related digital technologies have played a pivotal role in this advancement (Singh & Singh, 2017). Globally, there is a marked increase in the adoption of big data within healthcare. Data suggests a notable traction from 2016, with forecasts for 2025 indicating a rise from $11.5 million to a staggering $70 billion.

Through the integration of big data into operational processes, companies can enhance customer service provisions and devise bespoke marketing campaigns, thus contributing to augmented revenue generation. Organizations employing this methodology are endowed with a distinct competitive edge, possessing comprehensive market intelligence, encompassing evolving consumer tastes and preferences as well as extant market demand. In the context of healthcare systems, big data implementation enables organizations to proficiently manage patient medical records, hospital information, and data procured from an array of health-related devices. Such data is subsequently retained for future utilization. The ensuing discussion will explicate the significance of big data utilization in Kuwait, in addition to examining the

challenges associated with its implementation and management within the nation's business and industrial sectors.

Use of Big Data in Kuwait

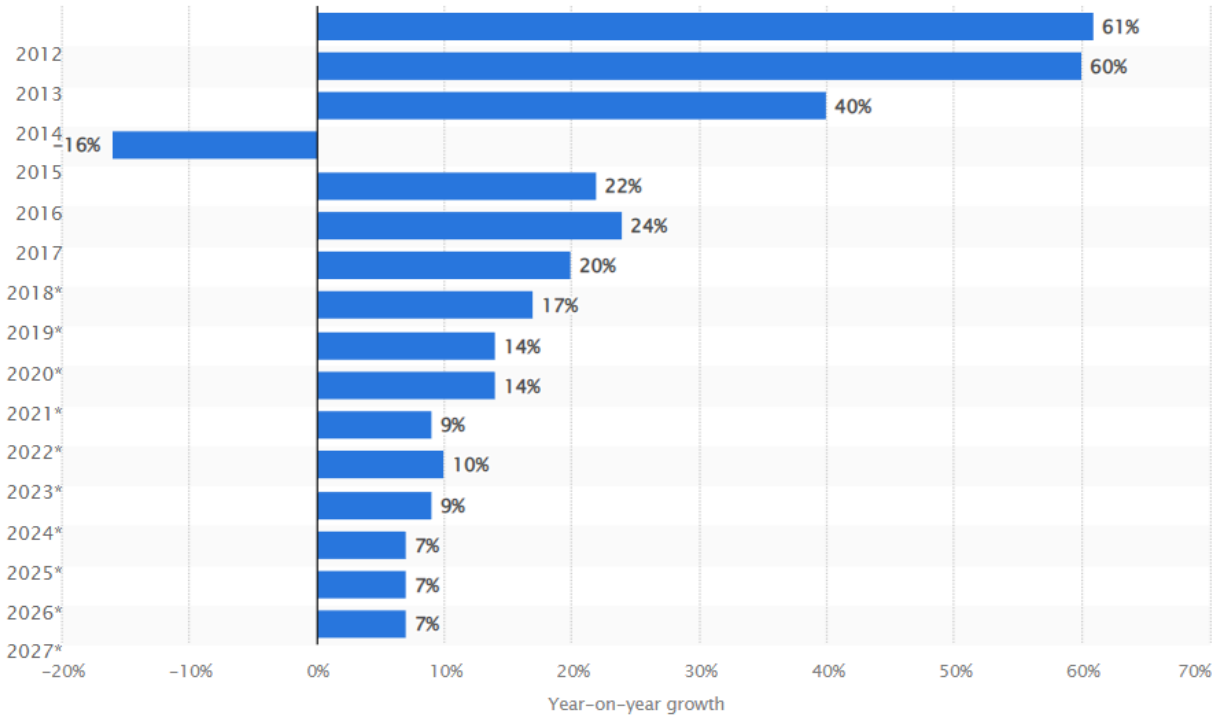

Figure 1: Global Growth of Big Data

(Source: Statista Research Department, 2022)

The utilization of big data offers a plethora of advantages for corporations, as it facilitates the generation of precise predictions, enabling organizations to devise strategic planning initiatives that bolster their operational capabilities. An examination of big data usage reveals that the global market experienced significant growth between 2012 and 2027, with predictions indicating a surge of over 30% in the forthcoming years. In 2018 alone, the market for big data witnessed a

20% increase, amounting to nearly $42 billion in value (Yeung, 2017). Kuwait has been endeavoring to actualize a digital transformation by creating a framework predicated on intelligent e-services. Over recent years, the nation has made considerable progress, transitioning from a severely war-affected state to a developing country through the implementation of digitalization initiatives. The Kuwaiti government has been steadfast in its efforts to establish online government platforms, which will facilitate the digital transformation process and the establishment of a smart government. It is worth noting that the industrial sector has already experienced significant changes due to the adoption of big data analytics, thereby providing improved operational support. This measure taken by the government is anticipated to enhance social welfare and align Kuwait with other nations on a global scale (Yeung, 2017). The integration of big data analytics and digital transformation will afford Kuwait the opportunity to vie for prominence both within the Gulf region and on an international stage (Vassakis et al., 2018). This will also address the nation's technological advancement, along with various legal and ethical challenges that have arisen over time. The advent of the COVID-19 pandemic saw the introduction of online services and contemporary technologies such as big data analytics and artificial intelligence, paving the way for the industrial and health sectors to gradually assimilate these advanced systems. The government has also taken the crucial step of implementing artificial intelligence and related technologies, with the aim of incorporating digitalization services throughout the country, thereby facilitating the industrial and healthcare sectors in their optimal utilization of these services (Singh & Singh, 2017).

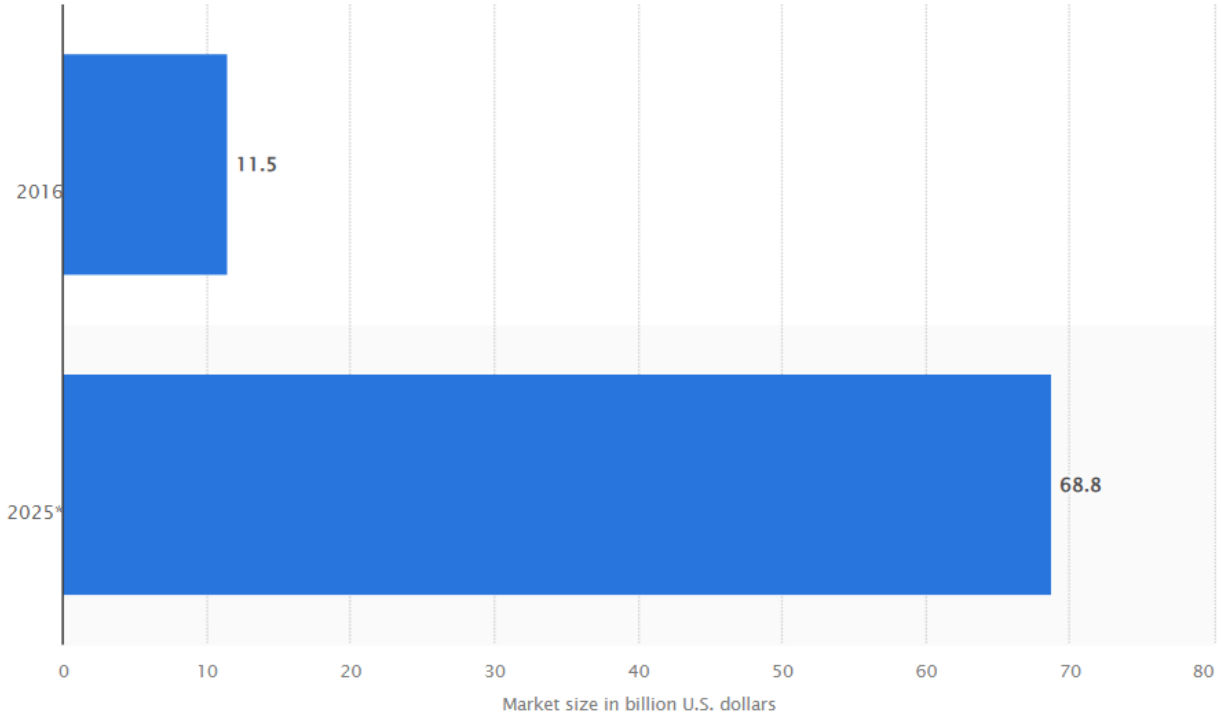

Figure 2: Growth of Big Data in Healthcare Sector globally

(Source: Stewart, 2021)

Empirical evidence demonstrates a substantial global upsurge in the employment of big data within the healthcare industry. A meticulous examination of the pertinent data reveals that the healthcare sector's big data industry gained considerable traction in 2016. Projections for 2025 suggest a significant potential for growth in big data implementation within the healthcare industry, with an estimated increase from approximately 11.5 million USD to nearly 70 billion USD. This underscores the prominence accorded to this sector, predicated on its capacity to bolster healthcare support and augment overall efficiency. In the context of Kuwait, the incorporation of big data analytics into the healthcare system promises to fortify operational support by enhancing surveillance capabilities for the detection of

communicable and non-communicable diseases. This innovation will also facilitate the development of targeted strategies and enable the execution of interventions through which health promotion and prevention can be efficaciously managed. Furthermore, big data analytics will streamline the accurate identification of at-risk populations, thereby enabling the healthcare sector to acquire a comprehensive understanding of various diseases and fluctuations in human health. The implementation of big data analytics will foster a more profound comprehension of the interplay between genetic and environmental determinants of health, thus paving the way for the discovery of mitigating strategies to optimally manage the population's health (Yeung, 2017).

Within Kuwait, the integration of big data in healthcare stands to revolutionize surveillance of diseases, both communicable and non-communicable. It will also enable targeted health strategies and interventions, ensuring better health promotion and prevention measures. Moreover, with the aid of big data analytics, healthcare professionals can identify at-risk groups with precision, gaining a deeper understanding of various health conditions. This deep dive into data will enhance the understanding of the interrelation between genetic and environmental health factors, potentially leading to better health management strategies (Yeung, 2017).

Big Data and Its Mechanism of Operation

Big data refers to extremely large data sets that are beyond the capacity of traditional database systems to capture, store, manage and analyze (Kar, and Dwivedi, 2020). This data often emanates from a plethora of sources including sensors, social media platforms, digital images and transaction records (Upadhyay, and Domadiya, 2023). Typically characterized by its volume, velocity and variety (the three Vs), big data could further include dimensions of veracity and value (Sun, and Scanlon, 2019). At its core, the operational mechanism of big data hinges on distributed storage and parallel processing (Ajah, and Nweke, 2019). Rather than being stored on a single machine, the data is distributed across a cluster of machines (Cui, et al., 2020). Tools such as Apache Hadoop and Spark have become pivotal in this domain. Hadoop's Distributed File System (HDFS) allows for the resilient storage of vast quantities of data across multiple nodes (Belov, et al., 2021). Its MapReduce component facilitates parallel processing; moreover, breaking down tasks into smaller sub-tasks to be processed concurrently (Abuín, et al., 2020). On the other hand, Spark offers in-memory processing, which significantly accelerates data processing tasks. These frameworks ensure that big data analytics processes are scalable, fault-tolerant and reliable (Bhagat, and VM, 2022). Furthermore, machine learning algorithms and advanced analytical techniques could then be deployed on this distributed architecture to derive insights, patterns and predictions from the data (Sun, and Scanlon, 2019). As data continues to grow in both complexity and volume, the tools and methodologies surrounding big data are continually evolving to meet the challenges and opportunities it presents (Li, Chen, and Shang, 2022).

Practical Ways to Manipulate Big Data in Kuwaiti Healthcare System

Manipulating big data entails a series of processes involving data collection, storage, processing, analysis, and visualization (Al-Alwan, et al., 2022) (. At each stage, it might take place in multifarious manners as follows:

Data Collection and Ingestion

This can be done through Tools like Fluentd or Logstash which can collect logs and events (Kratzke, 2022). Moreover, Apache Kafka and Apache Pulsar are utilized for real-time data ingestion and streaming (Fennell, 2022). Regarding data collection and ingestion in the Kuwaiti Healthcare Sector Log Collection includes Fluentd and Logstash (Boros, Lehotay-Kéry, and Kiss,). Regarding Fluentd, within the healthcare ecosystem of Kuwait, diverse systems like Electronic Health Records (EHR), Hospital Information Systems (HIS) and medical imaging devices continuously generate logs (Tiwari, et al., 2021). Fluentd can serve as an integrative layer, capturing these logs in a consistent format (Djiken, et al., 2021). For instance, every access or modification to a patient's EHR can be logged, ensuring transparency and traceability. Fluentd's flexibility with various plugins can allow it to interface with proprietary healthcare systems and databases common in medical environments (Bhanage, et al., 2021).

For Logstash, given the varied nature of medical data, from textual notes to binary image data, Logstash can be configured to ingest this heterogeneous data efficiently (Tisbeni, et al., 2021). Its powerful filtering and transformation capabilities can be harnessed to structure data, like parsing medical codes or categorizing log events, before stashing them into an analytics-ready environment such as Elasticsearch (Bautista, et al., 2019).

Apart from log collection, stream Processing can be processed through Apache Kafka and Apache Pulsar (Kjerrumgaard, 2021). In case of Apache Kafka, in real-time healthcare scenarios, like emergency patient monitoring, immediate data availability can be crucial (Debauche, et al., 2022). Devices such as heart rate monitors, respiratory devices, or ICU equipment can be configured as producers pushing real-time patient metrics into Kafka topics (Farahani, et al., 2018). Healthcare professionals or systems can then act as consumers, accessing this data stream for immediate insight or triggering alarms for patient health anomalies (Arumugam, et al., 2019). Kafka's persistent storage ensures that even high-frequency data from these devices is reliably stored for retrospective analysis or audit (Farahani, et al., 2018).

Regarding Apache Pulsar, taking into consideration the expansion of telemedicine and inter-hospital coordination in modern healthcare, a system like Pulsar can be advantageous for the Kuwaiti healthcare domain (Özgüven, and Eken, 2023). By leveraging Pulsar's native support for multi-data center replication, patient data can be streamed across facilities in real-time, ensuring immediate availability irrespective of the point of care (Macey, 2021). Its segment-centric storage can aid in consistently high performance, essential for critical healthcare applications (Kjerrumgaard, 2021).

To manipulate big data in the Kuwaiti healthcare sector, it is paramount to consider data security and patient privacy (Al-Otaibi, et al., 2022). Hence, regulations might necessitate special configurations and protections to ensure the confidentiality, integrity, and availability of patient data (Al Kuwaiti, et al., 2023). Proper encryption, access controls, and audit trails should be integral components of the data collection and ingestion processes (Duggineni, 2023).

Data Storage

Manipulation of big data in storage phase could be done during Distributed Storage or/and NoSQL Databases (Boudriki Semlali, and Freitag, 2021). For Distributed Storage, big data generally leverages distributed storage systems like Hadoop's Distributed File System (HDFS) or cloud storage solutions such as Amazon S3 (Shorfuzzaman, and Masud, 2019). In regard to NoSQL Databases, for unstructured or semi-structured data, databases like MongoDB, Cassandra and HBase are preferred due to their scalability and schema flexibility (Sen, and Mukherjee, 2023). When it comes to Kuwaiti Healthcare sector, manipulated of big data in distributed storage will be through Hadoop's Distributed File System (HDFS) and Amazon S3 (Ahmed, et al., 2021).

Regarding Hadoop's Distributed File System (HDFS), acknowledging the vast and ever-growing amounts of patient data, laboratory results, medical imaging and other healthcare-related datasets, HDFS would be an efficient solution for the Kuwaiti healthcare infrastructure (Shorfuzzaman, and Masud, 2019). By distributing data across multiple nodes, HDFS ensures high fault tolerance, making data loss due to any single point of failure highly unlikely. HDFS can be tuned to store large sequences of patient records and medical images, facilitating quick access and ensuring data durability (Ahmed, et al., 2021). Backup nodes within the HDFS cluster can be set up in geographically distinct data centers in Kuwait, ensuring data redundancy and availability even during unforeseen disasters (Saadoon, et al., 2022).

Moreover, for healthcare organizations in Kuwait that opt for cloud-based storage solutions, Amazon S3 provides scalable object storage (Farid, et al., 2021). acknowledging its ability to integrate seamlessly with various analytics tools, it is ideal for storing patient data, genomic

sequences, or medical imaging (Rahamathulla, 2020). With features like data versioning, lifecycle policies and server-side encryption, S3 ensures that the healthcare data remains both secure and compliant with any regional data protection regulations (Alazzam, et al., 2021).

Moreover, experts in Kuwaiti healthcare sector can practically manipulate big data through

NoSQL Databases through MongoDB, Cassandra, and HBase (Topcu, and Rmis, 2020).

MongoDB: With the diverse nature of healthcare data, MongoDB's document-oriented nature allows for flexible schemas (PRIOR, 2020). It could store textual patient records, prescriptions, and clinical notes without requiring a strict structure. Its horizontal scaling ensures that as the data volume in Kuwaiti healthcare institutions grows, the database may scale seamlessly (Truică, et al., 2021).

Cassandra: taking into account its high availability and fault tolerance, Cassandra is particularly suited for critical applications such as real-time patient monitoring or telemedicine platforms (KEKEVİ, and AYDIN, 2022). Its distributed nature means that even with large influxes of patient data, read and write operations remain fast. In the Kuwaiti healthcare context, where timely data access might be crucial, Cassandra's promise of low-latency operations could be invaluable (Topcu, and Rmis, 2020).

HBase: Operating over HDFS, HBase brings the power of real-time read/write access to big data (Ahmed, et al., 2021). For scenarios in the Kuwaiti healthcare sector where quick, random access to vast datasets is essential such as querying a patient's longitudinal health record across years or accessing patterns in epidemics, HBase can offer a robust solution (Sen, and Mukherjee, 2023).

For the effective manipulation of big data in Kuwait's healthcare sector, it is imperative to prioritize data security and integrity (Kumi, 2023).

Specialized configurations including encryption at rest and in transit, regular backups and role-based access controls should be established. Due to the sensitive nature of healthcare data, adhering to any Kuwaiti regulations regarding data protection and patient privacy is crucial during storage implementation and management (Al Kuwaiti, et al., 2023).

Data Processing

This can be done through Batch Processing in which Apache Hadoop's MapReduce allows for distributed processing of large data sets across clusters of computers (Ahmed, et al., 2020). In the healthcare environment of Kuwait, periodic analyses, like monthly disease prevalence reports or annual epidemiological studies, require processing vast amounts of data (Escamilla, 2020). MapReduce, a core component of Hadoop, is adept at handling such large-scale, batch-oriented tasks (Sultan, and Abdullah, 2023). Patient records, diagnostic results, and other data residing in HDFS can be processed using MapReduce jobs. The Map phase can parse and filter relevant data, and the Reduce phase can aggregate, summarize or apply more complex computations (Bharti, et al., 2019). For instance, to identify disease trends, MapReduce could sift through millions of patient records, map out diagnosed diseases, and then reduce them to a prevalence count (Escamilla, 2020).

Moreover, Real-time Processing could be implemented as Apache Spark and Apache Flink could handle real-time data processing tasks (Nazari, et al., 2019). Spark's in-memory capability ensures faster data processing compared to MapReduce. For pache Spark, the dynamic nature of healthcare data in Kuwait, like real-time patient vitals or emergency room statuses, necessitates faster processing methods than batch processing can offer (Fernandes, et al., 2020). Apache Spark,

with its in-memory computing capabilities, would rapidly process these real-time datasets. It may seamlessly integrate with HDFS or other storage systems, providing a versatile platform for varied tasks from simple data filtering to complex machine learning models (Shorfuzzaman, and Masud, 2019). For example, Spark Streaming could continuously monitor patient vital signs, and Spark MLlib would be employed for predictive analytics such as forecasting patient deterioration based on incoming data streams. In regarding to Apache Flink, Spark manages both batch and stream processing; nevertheless, ApacheFlink is designed primarily for stream processing, making it extremely effective for real-time healthcare scenarios in Kuwait (Nazari, et al., 2019). It provides accurate, event-time-based computations and supports complex event processing. In a telemedicine context, for instance, Flink could be utilized to monitor and analyze video stream data, vital signs, and patient feedback concurrently so as to offer healthcare professionals insights in real time (Davidson, and Ravindran, 2021).

Incorporating these data processing tools into the Kuwaiti healthcare sector involves setting up the necessary computational infrastructure, optimizing configurations for healthcare-specific workloads and ensure data privacy and security measures are robustly implemented (Alazzam, et al., 2021). Taking the criticality of healthcare data, fault tolerance, recovery mechanisms and consistent monitoring should be integral to any data processing setup (Saadoon, et al., 2022).

Data Analysis

SQL-like Querying: Tools like Apache Hive and Apache Impala allow for SQL-like querying on top of HDFS which makes it easier for analysts to work with big data (Venkatesh, et al., 2019). As such, healthcare

analysts in Kuwait which are familiar with relational databases and SQL could trigger seamlessly transition to big data analytics using Apache Hive (Loris, et al., 2022). This tool provides an abstraction over HDFS which would allow structured querying using a language called HiveQL (Venkatesh, et al., 2019). For instance, querying trends in patient diagnoses, drug prescriptions or even treatment outcomes becomes straightforward as Hive translates these HiveQL queries into MapReduce jobs that run atop HDFS (Sultan, and Abdullah, 2023). Partitioning and bucketing features of Hive can optimize query performance, essential for massive datasets such as those from nation-wide health records or genome studies (Mustafa, et al., 2023). Moreover, for those situations in the Kuwaiti healthcare domain requiring quicker, near real-time insights, Apache Impala shines. Unlike Hive, which converts queries into MapReduce jobs, Impala directly queries the data in HDFS, ensuring lower latencies (Yu, and Zhou, 2019). Such capabilities could be specially useful for interactive dashboards displaying real-time hospital metrics or querying the immediate impact of health campaigns or interventions (Abuqabita, et al., 2019).

Machine Learning Libraries: Libraries such as MLLib (for Spark) or platforms like TensorFlow and PyTorch facilitate advanced data analyses and the development of predictive models (Rosendo, et al., 2022). As such, for MLLib, within the Apache Spark ecosystem, MLLib provides a suite of scalable machine learning algorithms optimized for big data. The Kuwaiti healthcare sector harness this library for tasks ranging from clustering patients based on symptoms, classifying disease types or even regression models predicting patient outcomes (Al Kuwaiti, et al., 2023). Its integration with Spark ensures that data pre-processing, model training, and validation occur within the same environment, streamlining the analytics pipeline. In terms of TensorFlow

&PyTorch, for deep learning applications especially in areas like medical imaging (MRI, CT scans) or genomics, TensorFlow and PyTorch are indispensable (Pati, et al., 2021). Kuwaiti healthcare institutions might leverage these platforms to develop convolutional neural networks for image analyses such as tumor detection or organ segmentation (Nazari, et al., 2019). Additionally, recurrent neural networks can be constructed for sequence data like DNA or time-series patient data. These platforms offer extensive flexibility, support for GPUs (crucial for large-scale training) and extensive communities providing pre-trained models and tools tailored for healthcare applications (Tiwari, et al., 2021).

Incorporating these data analysis tools into Kuwait's healthcare framework requires careful consideration of infrastructure demands especially for machine learning applications (Al-Otaibi, et al., 2022). Additionally, because of the critical and sensitive nature of healthcare data, rigorous validation and verification processes should be in place to ensure the accuracy and reliability of analytic outcomes. Furthermore, appropriate measures should be implemented to protect patient confidentiality during the analytics process (Alazzam, et al., 2021).

Data Cleaning and Transformation (ETL Processes)

Data cleaning and transformation particularly the ETL processes are vital for ensuring the quality and accuracy of analyses mainly in critical sectors like healthcare. ETL Tools including Apache NiFi, Apache Beam, and Talend can extract, transform, and load (ETL) data which renders it suitable for analysis. In further details, Apache NiFi: Deployed as a data flow automation tool, Apache NiFi is integral for sourcing, transforming, and distributing large amounts of data within the Kuwaiti healthcare system (Mira, 2023). Designed with a web-based user interface, NiFi provides real-time control that ensures transparent flow

of data between systems (Manickam, and Rajasekaran Indra, 2023). In a healthcare context, it could be utilized to pull patient records from disparate systems, transform the data into a unified format and then load into centralized databases or data lakes. Its provenance feature ensures traceability of every piece of data which is a crucial aspect in healthcare to maintain data lineage and integrity (Rahamathulla, 2020).

Apache Beam: This tool offers a unified model for defining both batch and streaming data-parallel processing pipelines. For the Kuwaiti healthcare sector which might deal with both real-time data (like patient monitors) and historical data (like past medical records), Apache Beam would abstract the complexity of distinguishing between these processing types (Kimaina, et al., 2022). Using Beam, data from different healthcare systems, regardless of their tempo could be cleaned (e.g., handle missing values, correct anomalies) and transformed (e.g., normalization, encoding) before analysis.

Talend: Talend, an open-source integration platform, supports ETL processes for big data. In the Kuwaiti healthcare landscape, data often resides in varied formats and sources, from traditional databases to modern cloud-based storage (Manickam, and Rajasekaran Indra, 2023). Talend could be configured to extract from these myriad sources, transform the data to adhere to healthcare standards (like HL7 or FHIR) and load into target analytics platforms. Its rich set of connectors and its drag-and-drop interface simplify the otherwise complex task of healthcare data integration (Rosendo, et al., 2022) .

Data Cleaning in Healthcare: as there is critical nature of healthcare decisions, data quality is paramount. Tools like NiFi, Beam, and Talend might automate the cleaning processes through handling missing values (e.g., imputation), outlier detection and correction, deduplication of patient records and standardizing data formats (Mira, 2023). For example, patients' records from different departments might represent

data differently. To tackle such cacophony, a unified representation can be achieved through these ETL tools (Mira, 2023).

Data Transformation ensures that the data is in the right shape and format for downstream analyses. This might involve encoding categorical variables, normalization of quantitative metrics (like lab results), feature engineering for machine learning or restructuring datasets for specific analytical tools (Kimaina, et al., 2022).

Implementing these ETL tools in the Kuwaiti healthcare sector will require a thorough understanding of the data, regular audits for data quality and consistent monitoring of ETL pipelines to ensure they are functioning as intended (Al Kuwaiti, et al., 2023). Proper documentation and logging mechanisms are also essential for traceability and troubleshooting.

Data Visualization

Data visualization has role in the healthcare sector as it provides an intuitive representation of complex datasets and facilitating informed decision-making.

BI Tools like Tableau, Power BI, and Apache Zeppelin would be integrated with big data systems for data visualization and dashboards. These tools allow technical users to craft visual representations of complex data sets (Manickam, and Rajasekaran Indra, 2023).

Tableau: Tableau stands as a premier data visualization tool with its ability to connect seamlessly to various data sources that ranges from traditional databases to distributed big data systems like Hadoop (Belov, et al., 2021). In the Kuwaiti healthcare environment, Tableau would be utilized to craft interactive dashboards displaying patient demographics, trends in disease prevalence or the efficacy of treatments (Al-Otaibi, et al., 2022). Its drag-and-drop interface enables

swift data exploration, while its powerful engine could manage the rendering of intricate visualizations from voluminous datasets. For example, epidemic tracking would be visualized via heat maps or patient recovery rates through time-series analyses (Ahmed, et al., 2020).

Power BI: Microsoft's Power BI provides a holistic suite of business analytics tools tailored for data integration, transformation and visualization. Its deep integration with Azure which is Microsoft's cloud platform makes it a fitting choice for Kuwaiti healthcare establishments employing Azure-based big data infrastructures (Al-Otaibi, et al., 2022). Power BI's Q&A feature allows users to query data using natural language which is a boon for non-technical healthcare professionals (Alghamdi, and Al-Baity, 2022). Its DAX (Data Analysis Expressions) scripting provides flexibility in designing custom metrics such as patient readmission rates or hospital resource utilization.

Apache Zeppelin: Designed specifically for big data and integrated with data processing backends like Apache Spark, Apache Zeppelin offers the Kuwaiti healthcare sector a web-based notebook for data exploration, visualization and collaborative analytics (Nazari, et al., 2019). Zeppelin's interpreter concept allows direct connectivity to various data sources and its dynamic forms make it easier to parameterize visualizations. For instance, genomic data analyses or deep dives into electronic health records could be visualized in Zeppelin notebooks and shared with professionals for collaborative decision-making (Kumi, 2023).

Visualization Paradigms for Healthcare include Temporal Analyses Tools which would be configured to display trends over time, crucial for monitoring the progression of diseases, patient recovery rates or the impact of healthcare interventions (Alghamdi, and Al-Baity, 2022). Moreover, Geospatial Visualization in which Mapping tools integrated

within these BI platforms offer geographical representations which is essential for tracking disease outbreaks or visualizing the distribution of healthcare facilities across Kuwait (Al-Otaibi, et al., 2022). Moreover, Hierarchical Data in healthcare are generally nested like patients within hospitals or symptoms within diseases. Treemaps or nested pie charts would visualize such hierarchies, giving thoughts into data distributions at various levels (Gamal, et al., 2021). In addition, Correlation and Dependency in which Scatter plots, correlation matrices or even complex network graphs could be employed to understand relationships such as correlations between health metrics or dependency of symptoms in complex diseases (Manickam, and Rajasekaran Indra, 2023).

Deploying these visualization tools in the Kuwaiti healthcare sector necessitates adherence to data privacy standards so as to assure that sensitive patient information is adequately protected (Al-Otaibi, et al., 2022). Moreover, it is imperative that visualizations are accurate, comprehensible and designed with the target audience whether technical professionals or medical practitioners in mind (Kimaina, et al., 2022).

Optimization and Manipulation Tools

DataFrames: Abstractions like Pandas (in Python) and DataFrames in Spark facilitate data manipulation using a tabular structure. As such, Pandas in Python provides DataFrame structures for efficiently manipulating structured data (Abuín, et al., 2020). In the healthcare context of Kuwait, Pandas could be leveraged to preprocess and cleanse patient data, conduct exploratory data analysis and restructure datasets for downstream analytics (Alnashmi, et al., 2022). Its suite of built-in functions allows for tasks like handling missing data,

aggregating metrics or filtering records based on specific criteria which is essential operations for maintaining data quality in healthcare records (Kratzke, 2022)

Spark DataFrames: Operating in distributed environments, Apache Spark's DataFrame API provides a higher-level abstraction for manipulating big datasets (Fernandes, et al., 2020). It denotes the flexibility to scale operations across multiple nodes in a cluster. Within the Kuwaiti healthcare context, Spark DataFrames could be employed to process vast quantities of medical records, genomic data, or real-time patient monitoring data (Ahmed, et al., 2020). The lazy evaluation model of Spark ensures optimizations are made before execution to enable efficient processing. Moreover, built-in functionalities like SQL querying and seamless integration with machine learning libraries make it apt for diverse healthcare analytics tasks (Nazari, et al., 2019).

Distributed Algorithms: Algorithms are generally adapted to run in distributed environments to manipulate data across several nodes. These algorithms ensure the effective partitioning, replication, and computation on subsets of the data (Farahani, et al., 2018). Partitioning: Effective partitioning is essential when dealing with massive datasets in distributed environments. In the healthcare sector, data might be partitioned based on attributes like patient IDs, hospital branches or disease types. Proper partitioning ensures localized computation, reducing data shuffle and, thereby, improving performance (Alazzam, et al., 2021) . For instance, patient records would be partitioned based on hospitals to enable all data related to a specific hospital resides on a single node and optimize data access and computation.

Replicating data across nodes ensures fault tolerance and high availability. In critical sectors like healthcare, where data loss might have dire consequences, replication ensures that even if a node fails, the data remains accessible from other nodes (Al Kuwaiti, et al., 2023).

However, striking a balance between redundancy for fault tolerance and storage costs is vital. In addition, Distributed Algorithms are designed to process data that resides across different nodes efficiently. By breaking a task into smaller sub-tasks and distributing them to nodes holding the relevant data, distributed algorithms reduce data movement and accelerate computation (Escamilla, 2020). In the Kuwaiti healthcare context, this might involve distributed analyses of patient demographics, parallel processing of medical image datasets or concurrent computations to derive health insights from wearable device data (Al-Otaibi, et al., 2023).

Implementing these tools and strategies within the Kuwaiti healthcare sector necessitates a deep understanding of the data distribution and inherent characteristics of the healthcare domain (Nazari, et al., 2019). Moreover, compliance with data security and privacy regulations is paramount due to the sensitive nature of medical data. Proper benchmarking and performance tuning are essential to ensure these tools and algorithms deliver optimal results in real-world healthcare scenarios (Tiwari, et al., 2021).

Security and Governance

Security and governance are paramount concerns in any healthcare setting given the sensitive nature of the data. Authentication and Authorization could be realised through applying Apache Ranger and Apache Knox which provide security features for big data environments (Bhathal, and Singh, 2019)

Apache Ranger: Apache Ranger offers a centralized platform for the management of security policies across the Hadoop ecosystem (Bharti, et al., 2019). For the Kuwaiti healthcare sector, this tool is pivotal in ensuring that only authenticated personnel gain access to patient data and medical records (Ahmed, et al., 2020). Ranger allows for fine-

grained access control to empower administrators to specify which users or groups would access specific resources and the type of operations they can perform (Rahamathulla, 2020). Role-based access controls would ensure for example that only oncologists can access cancer patient data and only pharmacists can alter medication records.

Apache Knox: Serving as a REST API Gateway for interacting with the Hadoop ecosystem, Apache Knox provides another layer of security. It stands guard at the periphery of the Hadoop cluster to assureonly authenticated requests pass through (Belov, et al., 2021). This is mainly relevant for healthcare applications in Kuwait that might be sourcing data from remote locations or medical devices. Knox ensures that any API calls made to retrieve or update data are first authenticated, and potentially encrypted before they reach the core big data infrastructure (Farid, et al., 2021).

Data Lineage and Metadata Management could be achieved through tools like Apache Atlas help trace data origins and modifications to maintain data integrity and compliance (Kar and Dwivedi, 2020). Apache Atlas is a critical tool for data governance that provides a scalable framework for metadata management and data lineage. Within the context of the Kuwaiti healthcare system, understanding where data originated and how it has been transformed or manipulated over time is vital for maintaining data integrity and ensuring patient safety (Manickam, and Rajasekaran Indra, 2023).

With Apache Atlas, the process of data Lineage would enable medical professionals from tracing back the origins of any piece of data. For instance, if an anomaly is detected in a patient's lab results, the lineage feature could be used to track back all processes, transformations and sources that contributed to that result (Alazzam, et al., 2021).

Moreover, with comprehensive metadata tagging, healthcare datasets could be classified, indexed and annotated. This facilitates quicker data discovery and ensures compliance by flagging sensitive information like personally identifiable patient data or critical medical records (Escamilla, 2020). For Auditing and Compliance, Apache Atlas provides a robust auditing mechanism which assures stringent regulatory requirements surrounding healthcare data. Any access or changes to the data are logged to assure a transparent record for compliance checks (Djiken, et al., 2021). This feature becomes even more essential considering the regional and global regulations the Kuwaiti healthcare sector might need to adhere to assure patient privacy and data protection (Alnashmi, et al., 2022).

Integrating these tools into the Kuwaiti healthcare sector requires a meticulous approach. Setting them up correctly ensures the security and privacy of patient data and the integrity and reliability of the vast amounts of medical data being handled (Alnashmi, et al., 2022). Due to the grave implications of breaches or inaccuracies in the healthcare domain, a rigorous testing and validation regime should accompany the implementation of these tools (Mira, 2023).

4. Challenges : Difficulties of adopting big data and how to solve problems related to such challenges

There are several challenges associated with the implementation of Big Data Analytics and efficiently using it across all major operations in a country (Senthilkumar et al., 2018). Especially in the case of the Healthcare sector, the challenges are message and require the implementation of a Change management system through which complete changes can be made possible across all major operations by upgrading the present system (Janssen *et al.* 2017; Martin, 2020). The following section outlines the major challenges that are present followed

by how to find solutions to these issues so that big data can be used efficiently in all major industries in the country.

Lack of knowledge and information regarding the technology

One of the major issues that arises regarding big data is a lack of complete knowledge regarding the technology which includes how to use it followed by the benefits it provides by implementing it in business operations (Lee, 2017). It is important that companies using this method have skilled data professionals, including data scientists, analysts and engineers who can manage large data sets and efficiently analyze the data to find the effective outcome (Mohammadi and Al-Fuqaha, 2018). The lack of proper uneducated professionals in these fields makes it difficult to handle such complicated tools and as a result it gets difficult to implement these modern technologies across business operations. In the case of the health sector (Househ *et al*. 2017), The lack of knowledge makes it difficult to enhance the efficiency of the diagnosis for interpreting the results to provide the necessary medication (Rong et al., 2019).

Solution: The effective solution that can be outlined in this case is for organizations including health Institutions to invest more capital and recruit professionals who are skilled in the management and handling of these processes (Kaur *et al*. 2018). training and development programs can also be organized which will help in enhancing the knowledge of the existing employees and make them more efficient when it comes to managing Data Analytics (Goldstein *et al*. 2021). Distributing information and knowledge will enhance the effectiveness of handling such processes.

Lack of understanding of data sets

Another major issue that arises regarding the management and implementation of Big Data Analytics is insufficient understanding regarding the different kinds of data that is present and how to cleanse the data so that efficient analysis can be conducted (Dai *et al*. 2020). The lack of knowledge and understanding on what particular is required followed by how to store and process it, and the sources from where correct data can be found followed by how to clean the entire data to make it an effective mixture difficult to use this process (Kamilaris *et al*. 2017). The lack of employee knowledge thereby is a major hindrance as it may cause the increase in efficiency where employees may lose sensitive data without having any backup. it may also lead to conducting analysis without cleansing the data thereby giving rise to incorrect results (Dai *et al*. 2020). Increase of the health sector such lack of understanding may lead to wrong predictive analysis thereby leading to wrong implementation of strategies.

Solution: One of the best solutions that can be outlined in this case is organizing workshops and seminars that will provide the chance to the Employees to learn more about big data and understand the basic difference that is present between different kinds of data (Pépin et al., 2020). undertaking training programs may also prove to be successful making it easier to develop understanding leading to less mistakes (Chen *et al.* 2017).

Issues with Data storage and growth

Rapid increase in digitalization has brought forward one of the most pressing challenges associated with big data is regarding the storage of the data sets. In most cases there has been a rapid increase in use of modern technology that has induced companies to become concerned regarding how to store the data sets for an infinite period of time making

it difficult for them to handle it (Lv and Qiao, 2020). In most cases the data sets are unstructured and are built from documents and audio files, videos and text files which makes it difficult to further simplify the search while managing these storage locations (Benhlima, 2018). In such cases this raises concern especially for the Healthcare sector as they have to store large amounts of patient data which can be required at any number of times. In this case not having proper structure aur growth present within the systems raises concern and causes health officials to lose valuable time. Hence data storage is one of the major issues that is faced when implementing big data technology in organizations.

Solution: In most cases the effective solution which can be implemented using modern technologies through which it will become possible to compress, tiering and deduplication of these large data sets making it easier to reduce the total file size and store it efficiently (Baker, 2017). Through the use of the data tiering process it becomes easier to store each kind of data under appropriate names and spaces mostly on public clouds are private clouds along with flash storage that highlights the information regarding what kind of data is being stored and how important it is. The common tools that can be used include Hadoop, and NoSQL (Mehta and Pandit, 2018).

Integration of Data from different sources

Another major challenge that arises when managing Big Data Analytics is the different sources from which data is collected which can include social media, customer login financial reports, emails, chats among several other such aspects. Taking into consideration all these sources and combining the data to develop a strong report is a challenging task. In such cases it is important that every individual

working in this field have complete knowledge regarding the different kinds of data present and how to outline which sources are authentic and can be effectively used in the process (Al-Mesad, 2018). The process of data integration is crucial for conducting efficient analysis and interpretation so that the strategy developed from it is perfect and can be effectively implemented (Rialti et al., 2019). In the case of Healthcare systems it is to the efficient analysis of the provided data that Healthcare professionals decide on what kind of medications or health interventions to use that will provide efficient results (Rubinfeld & Gal, 2017).

Solution: One of the best methods that can be implemented to solve this issue bhai is by using proper tools that will help in efficiently integrating the data from different sources and create correct data sheets free from bias. This will ensure that efficient analysis is conducted to outline conclusive results (Alharthi *et al.* 2017). This includes tools such as centerprise integrity, talend data integration, Informatica powercenter among others.

Funding

The substantial health expenditure in Kuwait has played a pivotal role in bolstering the well-being of its citizens by ensuring the provision of superior healthcare resources. As reported by the World Bank in 2014, the total health expenditure constituted 48.68% of the nation's budget, based on various development indicators. This substantial allocation demonstrates the government's unwavering commitment to enhancing the support provided to its citizens, with per capita expenditure increasing considerably in recent years. However, the healthcare sector's inability to harness modern technology has resulted in extended working hours and escalated complications. It is posited that

the incorporation of big data could address these issues and bolster the nation's healthcare system. Despite the myriad benefits offered by the implementation of big data analytics in the healthcare sector and other industries in Kuwait, several challenges must be confronted and surmounted (Xiao et al. 2020).

Among these challenges are the lack of requisite infrastructure, substantial investment requirements, data integration complexities, and a deficiency of knowledge and skills to manage the ensuing analyses. Financial resources represent a significant challenge for a country grappling with numerous economic and legal hurdles, impeding the allocation of substantial investments in this domain (Shilo et al. 2020). Nonetheless, the government has initiated programs entailing the creation of a dedicated budget to invest in the implementation of these technologies, alongside artificial intelligence, to enhance the efficacy of the nation's systems. Furthermore, financial constraints hamper the establishment of the necessary infrastructure to facilitate the adoption of contemporary technology. The paucity of requisite skill sets constitutes another salient challenge, stemming from the unavailability of educational and training courses that equip the populace with the knowledge necessary to implement these technologies (Shayaa et al. 2018).

Solution: the advent of online education has afforded the population the opportunity to acquire expertise in these technologies and their implementation, thereby bolstering their skill sets. Skilled personnel with comprehensive knowledge of various methods of data integration are essential to address the complexities inherent in this process, facilitating the development of exhaustive analyses upon which decisions can be predicated. Consequently, the Kuwaiti industry is confronted with a plethora of challenges pertaining to the utilization and

implementation of big data analytics. The government's initiatives to foster digital transformation within the country are poised to yield substantial benefits for the populace by enhancing their skills and technological proficiency (Shamim et al. 2019).

5. Recommendations

The following recommendations which have been outlined will work towards providing efficient guidance and support to organizations who are intending to implement big data Technologies in their operations based on which strategic decisions will be undertaken.

Organizing workshops and training: As it has been discussed before, one of the major challenges involved with the management of big data technology is a lack of skilled professionals. One can implement these Technologies and develop correct interpretation of the data being provided. It is important that employees in organizations tasked with the management of Big Data Analytics are completely informed regarding how to manage these operations and the steps that are to be followed to efficiently develop correct interpretation from the data sheet given. Providing training to the Employees will ensure that they have proper working knowledge on how to implement this technology and efficiently use it (Alexandru *et al*. 2018). Moreover, conducting workshops periodically will provide them with advanced knowledge about this technology and prepare them for different aspects and advantages that are provided through the use of Big Data Analytics.

Providing more knowledge: The lack of complete knowledge regarding big data and its usage is another major challenge which can be rectified by conducting seminars at offices and Health Care Centers for employees through which they will be able to enhance their knowledge and learn about it.It is important for information to be

provided to to employees and workers across Industries regarding the different kinds of data that are present and how to clean them before conducting analysis using different analytical tools (Aldaihani *et al*. 2021). Information also needs to be provided regarding the strategies or tools that can be implemented to collect and group data together from different sources, so that an efficient data analysis process can be conducted.

Need for Financial resources: Financial backup is another major aspect that needs to be outlined in this case where organizations need to find sponsors who can provide them with financial resources through which Big Data Analytics can be implemented. The government needs to develop policies through which it can provide resources to Healthcare institutions in the country so that Big Data Analytics can be implemented within the operations (Alansari *et al*. 2018). The financial resources present will not only help in implementing these technologies in the operations but also provide sufficient resources to provide training to the employee so that they can effectively manage it or hire employees who have skills in the management of these Technologies.

References

- Abuín, J.M., Lopes, N., Ferreira, L., Pena, T.F. and Schmidt, B., 2020. Big data in metagenomics: Apache spark vs MPI. *Plos one*, *15*(10), p.e0239741.
- Abuqabita, F., Al-Omoush, R. and Alwidian, J., 2019. A comparative study on big data analytics frameworks, data resources and challenges. *Modern Applied Science*, *13*(7), pp.1-14.
- Ahmed, N., Barczak, A.L., Susnjak, T. and Rashid, M.A., 2020. A comprehensive performance analysis of Apache Hadoop and

Apache Spark for large scale data sets using HiBench. *Journal of Big Data*, *7*(1), pp.1-18.

- Ahmed, S.T., Basha, S.M., Arumugam, S.R. and Patil, K.K., 2021. *Big Data Analytics and Cloud Computing: A Beginner's Guide*. MileStone Research Publications.
- Ajah, I.A. and Nweke, H.F., 2019. Big data and business analytics: Trends, platforms, success factors and applications. *Big Data and Cognitive Computing*, *3*(2), p.32.
- Al Kuwaiti, A., Nazer, K., Al-Reedy, A., Al-Shehri, S., Al-Muhanna, A., Subbarayalu, A.V., Al Muhanna, D. and Al-Muhanna, F.A., 2023. A Review of the Role of Artificial Intelligence in Healthcare. *Journal of Personalized Medicine*, *13*(6), p.951.
- Al-Alwan, M., Al-Nawafah, S., Al-Shorman, H., Khrisat, F., Alathamneh, F. and Al-Hawary, S., 2022. The effect of big data on decision quality: Evidence from telecommunication industry. *International Journal of Data and Network Science*, *6*(3), pp.693-702.
- Al-Mesad, A., 2018. The preparation of big data phenomena in the public sector in Kuwait.
- Al-Otaibi, J., Tolma, E., Alali, W., Alhuwail, D. and Aljunid, S.M., 2022. The Factors Contributing to Physicians' Current Use of and Satisfaction With Electronic Health Records in Kuwait's Public Health Care: Cross-sectional Questionnaire Study. *JMIR Medical Informatics*, *10*(10), p.e36313.
- Alansari, Z., Anuar, N.B., Kamsin, A., Soomro, S., Belgaum, M.R., Miraz, M.H. and Alshaer, J., 2018, August. Challenges of internet of things and big data integration. In *International*

Conference for Emerging Technologies in Computing (pp. 47-55). Springer, Cham.

- Alazzam, M.B., Alassery, F. and Almulihi, A., 2021. A novel smart healthcare monitoring system using machine learning and the Internet of Things. *Wireless Communications and Mobile Computing*, *2021*, pp.1-7.
- Aldaihani, S.G., Aljadi, U.Y. and Alali, A., 2021. The Ability of Educational Leaders to Investment the Big Data in Activating Educational Policy in Kuwait. *Journal of Educational and Psychological Studies [JEPS]*, *15*(2), pp.184-200.
- Alexandru, A.G., Radu, I.M. and Bizon, M.L., 2018. Big Data in Healthcare-Opportunities and Challenges. *Informatica economica*, *22*(2).
- Alghamdi, N.A. and Al-Baity, H.H., 2022. Augmented Analytics Driven by AI: A Digital Transformation beyond Business Intelligence. *Sensors*, *22*(20), p.8071.
- Alharthi, A., Krotov, V. and Bowman, M., 2017. Addressing barriers to big data. *Business Horizons*, *60*(3), pp.285-292.
- Alnashmi, M., Salman, A., AlHumaidi, H., Yunis, M. and Al-Enezi, N., 2022. Exploring the health information management system of Kuwait: lessons and opportunities. *Applied System Innovation*, *5*(1), p.25.
- Arumugam, S., Boulanger, K., Parker, E. and Shamaly, C., 2019. VSight: Analytics as a Service for Patient Monitoring.
- Baker, R., 2017. Big Data. *Total survey error in practice*.
- Bautista, E., Romanus, M., Davis, T., Whitney, C. and Kubaska, T., 2019, August. Collecting, monitoring, and analyzing facility and systems data at the national energy research scientific

computing center. In *Workshop Proceedings of the 48th International Conference on Parallel Processing* (pp. 1-9).

- Belov, V., Tatarintsev, A. and Nikulchev, E., 2021. Choosing a data storage format in the apache hadoop system based on experimental evaluation using apache spark. *Symmetry*, *13*(2), p.195.
- Benhlima, L., 2018. Big data management for healthcare systems: architecture, requirements, and implementation. *Advances in bioinformatics*, *2018*.
- Bhagat, V. and VM, T., 2022. Review on Big Data Analysis Tools. *Available at SSRN 4031901*.
- Bhanage, D.A., Pawar, A.V. and Kotecha, K., 2021. IT infrastructure anomaly detection and failure handling: a systematic literature review focusing on datasets, log preprocessing, machine & deep learning approaches and automated tool. *IEEE Access*, *9*, pp.156392-156421.
- Bharti, U., Bajaj, D., Goel, A. and Gupta, S.C., 2019. Identifying requirements for Big Data analytics and mapping to Hadoop tools. *International Journal of Recent Technology and Engineering*, *8*(3), pp.4384-4392.
- Bhathal, G.S. and Singh, A., 2019. Big Data: Hadoop framework vulnerabilities, security issues and attacks. *Array*, *1*, p.100002.
- Boros, A.P., Lehotay-Kéry, P. and Kiss, A., Performance impact of network encryption on log processing with Spark.
- Boudriki Semlali, B.E. and Freitag, F., 2021. SAT-hadoop-processor: a distributed remote sensing big data processing software for earth observation applications. *Applied Sciences*, *11*(22), p.10610.

- Chen, M., Hao, Y., Hwang, K., Wang, L. and Wang, L., 2017. Disease prediction by machine learning over big data from healthcare communities. *Ieee Access*, *5*, pp.8869-8879.
- Cui, Y., Kara, S. and Chan, K.C., 2020. Manufacturing big data ecosystem: A systematic literature review. *Robotics and computer-integrated Manufacturing*, *62*, p.101861.
- Dai, H.N., Wang, H., Xu, G., Wan, J. and Imran, M., 2020. Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies. *Enterprise Information Systems*, *14*(9-10), pp.1279-1303.
- Davidson, G.P. and Ravindran, D.D., 2021. Technical Review Of Apache Flink For Big Data. *Int. J. of Aquatic Science*, *12*(2), pp.3340-3346.
- Debauche, O., Nkamla Penka, J.B., Mahmoudi, S., Lessage, X., Hani, M., Manneback, P., Lufuluabu, U.K., Bert, N., Messaoudi, D. and Guttadauria, A., 2022. RAMi: A new real-time Internet of Medical Things architecture for elderly patient monitoring. *Information*, *13*(9), p.423.
- Djiken, G.L., Mourlin, F. and Mahmoudi, C., 2021. Indexing Large Amount of Log Data for Predictive Maintenance. *International Journal on Advances in Networks and Services Volume 14, Number 1 & 2, 2021*.
- Duggineni, S., 2023. Impact of Controls on Data Integrity and Information Systems. *Science and Technology*, *13*(2), pp.29-35.
- Escamilla, A.K.G., 2020. *Big Data and machine learning to improve medical monitoring and remote monitoring* (Doctoral dissertation, Université Bourgogne Franche-Comté).
- Farahani, B., Firouzi, F., Chang, V., Badaroglu, M., Constant, N. and Mankodiya, K., 2018. Towards fog-driven IoT eHealth:

Promises and challenges of IoT in medicine and healthcare. *Future generation computer systems*, *78*, pp.659-676.

- Farid, F., Elkhodr, M., Sabrina, F., Ahamed, F. and Gide, E., 2021. A smart biometric identity management framework for personalised IoT and cloud computing-based healthcare services. *Sensors*, *21*(2), p.552.
- Fennell, C., 2022. A Communication Architecture for Transportation Digital Twins using Apache Kafka.
- Fernandes, E., Salgado, A.C. and Bernardino, J., 2020, July. Big Data Streaming Platforms to Support Real-time Analytics. In *ICSOFT* (pp. 426-433).
- Gamal, A., Barakat, S. and Rezk, A., 2021. Standardized electronic health record data modeling and persistence: A comparative review. *Journal of biomedical informatics*, *114*, p.103670.
- Goldstein, I., Spatt, C.S. and Ye, M., 2021. Big data in finance. *The Review of Financial Studies*, *34*(7), pp.3213-3225.
- Hassan, M.K., El Desouky, A.I., Elghamrawy, S.M. and Sarhan, A.M., 2019. Big data challenges and opportunities in healthcare informatics and smart hospitals. In *Security in smart cities: Models, applications, and challenges* (pp. 3-26). Springer, Cham.
- Househ, M.S., Aldosari, B., Alanazi, A., Kushniruk, A.W. and Borycki, E.M., 2017. Big Data, Big Problems: A Healthcare Perspective. *ICIMTH*, pp.36-39.
- Janssen, M., van der Voort, H. and Wahyudi, A., 2017. Factors influencing big data decision-making quality. *Journal of business research*, *70*, pp.338-345.

- Kamilaris, A., Kartakoullis, A. and Prenafeta-Boldú, F.X., 2017. A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, *143*, pp.23-37.
- Kar, A.K. and Dwivedi, Y.K., 2020. Theory building with big data-driven research–Moving away from the "What" towards the "Why". *International Journal of Information Management*, *54*, p.102205.
- Kaur, P., Sharma, M. and Mittal, M., 2018. Big data and machine learning based secure healthcare framework. *Procedia computer science*, *132*, pp.1049-1059.
- KEKEVİ, U. and AYDIN, A.A., 2022. Real-Time Big Data Processing and Analytics: Concepts, Technologies, and Domains. *Computer Science*, *7*(2), pp.111-123.
- Kimaina, A., Dick, J. and Sadjad, B., 2022, June. OpenMRS Analytics Engine: A FHIR Based Approach. In *MEDINFO 2021: One World, One Health—Global Partnership for Digital Innovation: Proceedings of the 18th World Congress on Medical and Health Informatics* (Vol. 290, p. 314). IOS Press.
- Kjerrumgaard, D., 2021. *Apache Pulsar in Action*. Simon and Schuster.
- Kratzke, N., 2022. Cloud-native observability: the many-faceted benefits of structured and unified logging—a multi-case study. *Future Internet*, *14*(10), p.274.
- Kumi, E.A., 2023. *Assessing the Acceptability of Blockchain Technology as a Way to Protect Healthcare Data: A Qualitative Study* (Doctoral dissertation, Northcentral University).
- Lee, I., 2017. Big data: Dimensions, evolution, impacts, and challenges. *Business horizons*, *60*(3), pp.293-303.

- Li, C., Chen, Y. and Shang, Y., 2022. A review of industrial big data for decision making in intelligent manufacturing. *Engineering Science and Technology, an International Journal*, *29*, p.101021.
- Loris, B., Riccardo, C., Fabrizio, M., Alessio, O., Domenico, T. and Paolo, T., 2022. Programming big data analysis: principles and solutions. *Journal of Big Data*, *9*(1).
- Lv, Z. and Qiao, L., 2020. Analysis of healthcare big data. *Future Generation Computer Systems*, *109*, pp.103-110.
- Macey, T., 2021. *97 Things Every Data Engineer Should Know*. " O'Reilly Media, Inc.".
- Manickam, V. and Rajasekaran Indra, M., 2023. Dynamic multi-variant relational scheme-based intelligent ETL framework for healthcare management. *Soft Computing*, *27*(1), pp.605-614.
- Martin, K.E., 2020. Ethical issues in the big data industry. In *Strategic Information Management* (pp. 450-471). Routledge.
- Mehta, N. and Pandit, A., 2018. Concurrence of big data analytics and healthcare: A systematic review. *International journal of medical informatics*, *114*, pp.57-65.
- Mikalef, P. and Krogstie, J., 2018, June. Big Data Governance and Dynamic Capabilities: The Moderating effect of Environmental Uncertainty. In *PACIS* (p. 206).
- Mikalef, P., Boura, M., Lekakos, G. and Krogstie, J., 2019. Big data analytics capabilities and innovation: the mediating role of dynamic capabilities and moderating effect of the environment. *British Journal of Management*, *30*(2), pp.272-298.
- Mira, M.P.F., 2023. *Implementing a data integration infrastructure for healthcare data: a case study* (Master's thesis).

- Mohammadi, M. and Al-Fuqaha, A., 2018. Enabling cognitive smart cities using big data and machine learning: Approaches and challenges. *IEEE Communications Magazine*, *56*(2), pp.94-101.
- Mustafa, S.M.N., Farooq, M.U., Zehra, S.S. and Noronha, J.P.T., 2023, May. A Comparative Study of the Performance of Real time databases and Big data Analytics Frameworks. In *2023 7th International Multi-Topic ICT Conference (IMTIC)* (pp. 1-7). IEEE.
- Nazari, E., Shahriari, M.H. and Tabesh, H., 2019. BigData analysis in healthcare: apache hadoop, apache spark and apache flink. *Frontiers in Health Informatics*, *8*(1), p.14.
- Özgüven, Y.M. and Eken, S., 2023. Distributed messaging and light streaming system for combating pandemics: A case study on spatial analysis of COVID-19 Geo-tagged Twitter dataset. *Journal of Ambient Intelligence and Humanized Computing*, *14*(2), pp.773-787.
- Pati, S., Thakur, S.P., Bhalerao, M., Thermos, S., Baid, U., Gotkowski, K., Gonzalez, C., Guley, O., Hamamci, I.E., Er, S. and Grenko, C., 2021. Gandlf: A generally nuanced deep learning framework for scalable end-to-end clinical workflows in medical imaging. *arXiv preprint arXiv:2103.01006*.
- Pépin, J.L., Bailly, S. and Tamisier, R., 2020. Big Data in sleep apnoea: Opportunities and challenges. *Respirology*, *25*(5), pp.486-494.
- PRIOR, F., 2020. Document oriented graphical analysis and prediction. *Digital Personalized Health and Medicine: Proceedings of MIE 2020*, *270*, p.183.

- Rahamathulla, M.P., 2020. Cloud-based Healthcare data management Framework. *KSII transactions on Internet & information systems*, *14*(3).
- Rialti, R., Marzi, G., Ciappei, C. and Busso, D., 2019. Big data and dynamic capabilities: a bibliometric analysis and systematic literature review. *Management Decision*.
- Rong, M., Gong, D. and Gao, X., 2019. Feature selection and its use in big data: challenges, methods, and trends. *IEEE Access*, *7*, pp.19709-19725.
- Rosendo, D., Costan, A., Valduriez, P. and Antoniu, G., 2022. Distributed intelligence on the Edge-to-Cloud Continuum: A systematic literature review. *Journal of Parallel and Distributed Computing*, *166*, pp.71-94.
- Rubinfeld, D.L. and Gal, M.S., 2017. Access barriers to big data. *Ariz. L. Rev.*, *59*, p.339.
- Saadoon, M., Hamid, S.H.A., Sofian, H., Altarturi, H.H., Azizul, Z.H. and Nasuha, N., 2022. Fault tolerance in big data storage and processing systems: A review on challenges and solutions. *Ain Shams Engineering Journal*, *13*(2), p.101538.
- Sen, P.S. and Mukherjee, N., 2023. An ontology-based approach to designing a NoSQL database for semi-structured and unstructured health data. *Cluster Computing*, pp.1-18.
- Senthilkumar, S.A., Rai, B.K., Meshram, A.A., Gunasekaran, A. and Chandrakumarmangalam, S., 2018. Big data in healthcare management: a review of literature. *American Journal of Theoretical and Applied Business*, *4*(2), pp.57-69.
- Shamim, S., Zeng, J., Shariq, S.M. and Khan, Z., 2019. Role of big data management in enhancing big data decision-making

capability and quality among Chinese firms: A dynamic capabilities view. *Information & Management*, *56*(6), p.103135.

- Shayaa, S., Jaafar, N.I., Bahri, S., Sulaiman, A., Wai, P.S., Chung, Y.W., Piprani, A.Z. and Al-Garadi, M.A., 2018. Sentiment analysis of big data: Methods, applications, and open challenges. *IEEE Access*, *6*, pp.37807-37827.
- Shilo, S., Rossman, H. and Segal, E., 2020. Axes of a revolution: challenges and promises of big data in healthcare. *Nature medicine*, *26*(1), pp.29-38.
- Shorfuzzaman, M. and Masud, M., 2019. Leveraging a multi-objective approach to data replication in cloud computing environment to support big data applications. *International Journal of Advanced Computer Science and Applications*, *10*(3).
- Singh, D.S. and Singh, G., 2017. Big Data-A Review. *International Research Journal of Engineering and Technology*, *4*(04), pp.2395-0056.
- Stewart, C., (2021), Global big data in healthcare market size 2016 & 2025, Available at: https://www.statista.com/statistics/909654/global-big-data-in-healthcare-market-size/
- Statista Research Department., (2021), Global digital health market size 2019-2025 forecast, Available at: https://www.statista.com/statistics/1092869/global-digital-health-market-size-forecast/
- Statista Research Department., (2022), Forecast revenue big data market worldwide 2011-2027, Available at: https://www.statista.com/statistics/254266/global-big-data-market-forecast/

- Statista Research Department., (2022), Growth forecast: big data market 2012-2027, Available at: https://www.statista.com/statistics/255968/global-big-data-market-growth-forecast/
- Sultan, N.A. and Abdullah, D.B., 2023. A Comprehensive Study On Big Data Frameworks. *International Journal of Applied Sciences and Technology 5*(1), 35-48.
- Sun, A.Y. and Scanlon, B.R., 2019. How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. *Environmental Research Letters*, *14*(7), p.073001.
- Tisbeni, S.R., Cesini, D., Martelli, B., Carbone, A., Cavallaro, C., Duma, D.C., Falabella, A., Galletti, M., Gasparetto, J., Furlan, E. and Michelotto, D., 2021. A Big Data Platform for heterogeneous data collection and analysis in large-scale data centres. In *Proceedings of International Symposium on Grids & Clouds*.
- Tiwari, A., Dhiman, V., Iesa, M.A., Alsarhan, H., Mehbodniya, A. and Shabaz, M., 2021. Patient behavioral analysis with smart healthcare and IoT. *Behavioural Neurology*, *2021*.
- Topcu, A.E. and Rmis, A.M., 2020. Analysis and evaluation of the riak cluster environment in distributed databases. *Computer Standards & Interfaces*, *72*, p.103452.
- Truică, C.O., Apostol, E.S., Darmont, J. and Pedersen, T.B., 2021. The forgotten document-oriented database management systems: An overview and benchmark of native XML DODBMSes in comparison with JSON DODBMSes. *Big Data Research*, *25*, p.100205.
- Upadhyay, L.V. and Domadiya, D.H., 2023. The major boom of an Unstructured Big Data on Social Media.

- Vassakis, K., Petrakis, E. and Kopanakis, I., 2018. Big data analytics: applications, prospects and challenges. In *Mobile big data* (pp. 3-20). Springer, Cham.
- Venkatesh, K., Ali, M.J.S., Nithiyanandam, N. and Rajesh, M., 2019. Challenges and research disputes and tools in big data analytics. *International Journal of Engineering and Advanced Technology*, *6*, pp.1949-1952.
- Xiao, X., Tian, Q. and Mao, H., 2020. How the interaction of big data analytics capabilities and digital platform capabilities affects service innovation: A dynamic capabilities view. *IEEE Access*, *8*, pp.18778-18796.
- Yeung, K., 2017. 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society*, *20*(1), pp.118-136.
- Yu, J.H. and Zhou, Z.M., 2019. Components and development in Big Data system: A survey. *Journal of Electronic Science and Technology*, *17*(1), pp.51-72.
- Yu, Y., Li, M., Liu, L., Li, Y. and Wang, J., 2019. Clinical big data and deep learning: Applications, challenges, and future outlooks. *Big Data Mining and Analytics*, *2*(4), pp.288-305.