MapChain: A Blockchain-based Verifiable Healthcare Service Management in IoT-based Big Data Ecosystem

Umit Demirbaga, *Member, IEEE* and Gagangeet Singh Aujla, *Senior Member, IEEE*

***Abstract*—Internet of Things (IoT)-based Healthcare services, which are becoming more widespread today, continuously generate huge amounts of data which is often called big data. Due to the magnitude and intricacy of the data, it is difficult to find valuable information that can be used for decision-making and prediction. Big data systems take on a significant infrastructure service to better serve the purpose of IoT systems and support critical decision making. On the other hand, privacy preservation, data integrity, and identity verification are essential requirements in healthcare big data service management. To overcome these problems, this article offers a scalable computing system that provides verifiable data access mechanism for IoT-enabled health data analytics in the big data ecosystem. There are two primary sub-architectures in the proposed architecture, namely a big data analytics tracking system and a derived blockchain-based data storage/access system. This approach leverages big data systems and blockchain architecture to analyze, and securely store data from IoT-enabled devices and allow verified access to the stored data. The zero-knowledge protocol is used to ensure that no information is accessible to unauthenticated users alongside avoiding data linkability. The results demonstrate the effectiveness of the our method to solve the problems of big data analytics and privacy issues in healthcare.**

***Index Terms*—Big Data, Internet of Things (IoT), MapReduce, Blockchain, Healthcare**

I. Introduction

INFORMATION plays a key role in new developments and better decision making in service organisations. For this reason, data is of great importance for service industries. A well collected and analyzed data is used to predict current trends of certain parameters and future events [1]. As we see the benefit of this situation, new technological developments have been introduced to produce and collect more data in almost all areas of our lives, such as social activities, science, work, and health. This has caused the term ”big data” to become a part of our lives. Especially healthcare organisations are producing data at a tremendous speed, which brings with it challenges [2]. Healthcare big data refers to heterogeneous, multi-spectral, incomplete and uncertain observations (for example, diagnosis, illness, injury, treatment, physical and mental disorders, demography, and disease prevention) from primary sources in structured, semi-structured, and unstructured formats. Structured data includes ICD codes, phenotype, genotype, genomic information, while unstructured data includes medical imaging, notes, environmental, clinical notes, lifestyle, prescriptions, and health economics data [3]. Moreover, the advent of Internet of Things (IoT) has accelerated data-driven applications such as transportation, networking, smart cities, and healthcare [4]. The health industry, in particular, has been a significant area in terms of using various sensors and equipment to monitor a patient's health status [5].

The huge amounts of data generation have resulted from the advances in omics fields such as genomics, proteomics, and metabolomics [6]. The transition from paper medical records to Electronic Health Records (EHR) is also effective in the growth of data [28]. Physicians, epidemiologists and health policy specialists aim to improve population health and provide better patient care using such rich and large data [7]. Therefore, it is of great importance to develop tools, infrastructure and techniques in order to use the generated big data effectively [8]. **Early Warning Score (EWS)** systems, a timely surveillance system aimed at collecting information about patients' illnesses to trigger public health interventions, are a practical example of how data can be used [9]. In the UK, the **National Early Warning Score (NEWS)**, was introduced in 2012, aimed at detecting and responding to clinical deterioration in patients with acute illness and providing a standard platform for initial assessment of acute illness severity [10].

NEWS process overview: Seven different physiological parameters, which are *respiration rate, oxygen saturation, temperature, systolic blood pressure, heart rate, consciousness,* and *air or oxygen*, are obtained from patients with acute illness to detect and respond to clinical impairment during clinical evaluation. A score is given that reflects the effect of each parameter. The obtained individual parameter scores are summed up to form the patient's NEWS which takes between 1 and 8 to indicate the severity of the illness. The following presents the steps in the sequence of calculating the NEWS for each patient. **Step 1:** Obtaining the score for each of the seven physiological parameters from the patient. **Step 2:** Calculating the NEWS by aggregating all the physiological parameter scores. **Step 3:** Checking if the trigger threshold has been reached for a single parameter.

It is aimed to use the NEWS data to perform the specific tasks above mentioned. To be able to carry them out, the NEWS data need to be analysed, and complex queries should be built. This situation causes a big data problem due to the volume and complexity of the data in terms of data

U. Demirbaga is with University of Cambridge, United Kingdom, European Bioinformatics Institute (EMBL-EBI), United Kingdom, and Bartin University, Türkiye, E-mail: ud220@cam.ac.uk.

GS Aujla is with the Department of Computer Science, Durham University, United Kingdom. E-mail: gagangeet.s.aujla@durham.ac.uk.

Manuscript received April 2022; August 2022.

analytics. In addition, although this data is considered to be a key to improving health outcomes, it is worth noting that, while obtaining valuable information and reducing costs, data privacy and integrity issues are so overwhelming that the healthcare sector cannot fully benefit from the resources available. This is another challenge of NEWS data. For that reason, it becomes essential to ensure data privacy and verification of data integrity through effective mechanisms and strategies to fill the deficiencies regarding these issues. In the UK, South Tees Hospitals NHS Foundation Trust[1] is responsible for the management of two North East hospitals, which have approximately five million patient data (patients-anonymous health data sets) in relational databases, each of which consists of 20 tables with about 50 features about patients. However, the execution of the queries on these data might take days or weeks because of the huge volume and complexity of data as well as the user-related problems while defining the user-defined functions (UDFs).

The biggest challenge of big healthcare data analytics is dealing with heterogeneous data to provide insights into better healthcare for millions of patients. For the analysis of such large and complex healthcare data, big data systems help solve the problems of healthcare delivery systems. Big data processing systems such as Hadoop[2] and Spark[3] often run in large-scale, highly concurrent and multi-tenant environments; this can easily cause hardware and software failures and hence lead to performance degradation [11]. Complex queries created when working with complex data cause unnecessary delays in result retrieval. Data transfer from wearable devices and sensors to servers generates a large in-house healthcare IoT monitoring network. The inherent problems of big data and different data privacy (storage, access, retention, immutability of sensitive data) and data integrity (accuracy, completeness, consistency, and safety of data in accordance to General Data Protection Regulation (GDPR)) challenges occur as a result of the increasing number of sensors and real-time data collecting, processing and storage [12]. The healthcare IoT ecosystem generates the most sensitive big data that is further analysed using large-scale analytical platforms. This big data is processed and stored on high-end servers wherein any inaccuracy, inconsistency or incompleteness can lead to serious integrity concerns as it can lead to wrong diagnoses and end up in life-threatening situations. In addition, privacy protection is an issue that requires attention when working with large health data with sensitive content. Cloud-based big data systems, especially public cloud computing systems, are vulnerable to cyber-attacks. According to McAfee, approximately 3.1 million cloud-based users' data have been hacked in one year [13]. This situation reveals the danger of health data processed in big data systems often built on centralized datacenters.

To avoid such situations, effective measures that can ensure the integrity and privacy of big data are highly desirable for healthcare systems. Thus, various cryptographic primitives have been employed during the past decade to protect healthcare data. However, the heavy and hard crypto-primitives do not resolve the entire healthcare service organisation's concerns. Explicit use of stringent security policies often hinders the performance of big data systems due to the 5V characteristics of big data and does not provide the required privacy protection, integrity verification, and identity hiding. To tackle such challenges, blockchain has been popular in various systems as it can ensure the privacy and integrity of data due to its tamper-proof architecture. Blockchain uses a distributed ledger system wherein the transactions (data) are recorded in a verifiable manner. Bitcoin, the famous cryptocurrency using blockchain technology, uses a distributed and public ledger system to encrypt, validate and record transactions [14]. Transactions here are encrypted in a way that can be verified, ensuring data integrity. In addition to cryptocurrencies, blockchain is widely used in many areas that require data integrity and verification, such as financial exchanges, lending, insurance, and voting [15]. In such systems, when new data is sent from any IoT devices or any sources, a unique block is created at a local end node in the blockchain after the data is verified [16].

However, looking into the amount of data generated in healthcare setup, the conventional blockchain architectures may fail to ensure the desired performance and efficiency. So, one way is to limit the data stored on the blockchain. Moreover, storing all the data on the blockchain may lead to further privacy concerns in line with the data protection regulations enforced by various countries. Thus, it becomes important to decide what data need to be stored on the blockchain in order to handle the scale and privacy regulations. For example, if the NEW score (computed on the basis of NEW data) is stored onchain, i.e., on the blockchain and the actual data used to compute the NEWS is stored offchain (on a central cloud server). To ensure data integrity and verification, the hash of NEW data can also be stored on the blockchain. This way it is possible to verify the healthcare data both ways, i.e., the onchain NEWS can verify the integrity of data stored on the cloud and alternatively the offchain data can be used to recompute the NEWS to verify the correctness of the data stored on the blockchain. Apart from these issues, healthcare data is often subjected to or exposed to a wide range of users (data owner, doctor, hospital staff, etc.), making it necessary to verify the users as legitimate before allowing them the requested data access. The conventional methods often use authentication methods that can bind (or link) the identity of the user with the data leading to data linkability and privacy concerns. This concern becomes even more important when the data is related to healthcare. Thus, the verification process adopted to verify the legitimacy of users must ensure that the identity of users is not linked to the data.

To this end, we want to investigate the following two main research questions (*RQ*):

- ***(RQ1): How do we track the query efficiency on big data in healthcare service organisations?*** Some queries might take hours or days if a problem happens with any processes or machines. That is why monitoring the query phases plays a crucial role in finding the root of the

[1] https://www.southtees.nhs.uk/news/
[2] https://hadoop.apache.org/
[3] https://spark.apache.org/

problem to accelerate data processing. Monitoring the queries on big data in real-time poses a big data complexity problem as it needs the capture the dependencies of each component from multiple processes. For example, Apache Hive, a data warehouse used for providing data query and analysis, achieves good performance by scaling to large computing clusters, yet this might result in a very time-consuming when an error happens during query processing such as defining wrong syntax or failures in any data-nodes.

- ***(RQ2): How do we eliminate the data integrity, data linkability and privacy concerns related to healthcare big data?*** Big data systems can run on physical or virtualized servers and as well as in public clouds. Public clouds have many benefits, such as being quite scalable and affordable. However, their physical infrastructure is shared and offers no real control over the integrity or availability of data, which makes a big problem for healthcare data. The conventional primitives are often hard but complex to resolve the concerns of healthcare service management. Explicit use of stringent security policies often hinders the performance of big data systems and does not provide the desired privacy protection, integrity verification, and user identity hiding.

To the best of our knowledge, there is no previous study to answer the questions mentioned above regarding both tracking the performance problems (system or user-related) and eliminating data privacy and integrity problems in IoT-based big data systems in healthcare service organisations. Based on this fact, the contributions of this paper to solve the research questions mentioned above are as follows:

- To address *RQ1*, we propose an intelligent monitoring and visualization system that collects job execution information, including the process and status of each specific task, along with computing resource utilization in real-time. At the same time, the web-based visualization component modelled as directed acyclic graphs (DAGs) visualizes these streaming data and the query results executed in big data systems in a user-friendly interface.
- To address *RQ2*, we develop a blockchain-based system integrated into big data systems that extract the outputs of the data analytics, such as query results or NEWS. For this purpose, a derived blockchain solution, as suggested in our previous works [17]–[19], can be deployed in healthcare big data systems. The system gives authority and trust to a decentralized virtual network that shares the query results with predefined users (ensuring legitimate access avoiding data linkability). After being approved, the results are stored in the blockchain system in a secure way using a derived blockchain mechanism (on-chain and off-chain) that ensure two-way verification of data ensuring integrity. The stored results can be accessed only by a legitimate or authorised entity after verification of their identity and credentials (the ZKP technique is used to hide user identity under a temporary key).

A. Organisation

The paper is organized as follows. The background is outlined in §II. While §III presents the proposed method, the results are discussed in §IV. Before drawing a conclusion in §VI, we discuss the related work in §V.

II. Background

The background of the big data analytics process is described in the following sections.

A. The Process of Big Data Analytics in Healthcare

Big data analytics in healthcare helps analyze large datasets from millions of patients, identify correlations between datasets, and uncover complex relationships between different features (i.e., drugs, prescription). Fig. 1 demonstrates the processes of big data analytics in healthcare.

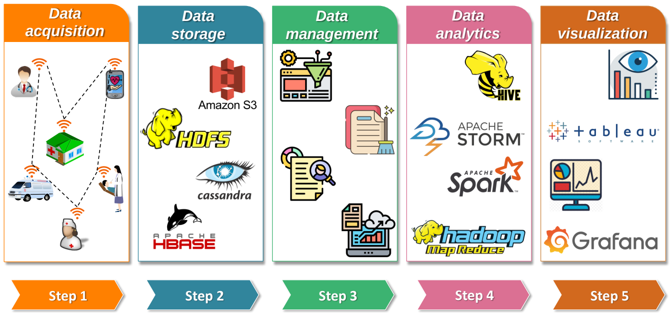


Fig. 1. Big data analytics in healthcare

1) Data acquisition: Healthcare data can be in any format such as structured, semi-structured or unstructured [20]. Furthermore, electronic health records, government sources, Computerized Physician Order Entry (CPOE), Health Maintenance Organization (HMO), social media, and smartphones can be seen as a source of big data in healthcare [21].

2) Data storage: As the size of the data in the healthcare area is dramatically increasing, it is needed a large storage platform to be able to store this data. At this point, clouds enable storage and managing of such data easily. As stated in [22], the structure of cloud provides an flexible platform to handle and analyse big data. Some important the cloud platforms are Google, Amazon, IBM, Cloudera, etc,. Such systems provide numerous advantages for healthcare in terms of data sharing between doctors, patients, and other health institutions [23]. However, there are several medicolegal concerns with this storage strategy, such as protection and confidentiality of information. Although cloud computing service providers claim to offer their services in a secure and reliable way, in reality, their distribution is not as secure and reliable as they claim [24]. To meet this deficit, blockchain is the right method of storing data in a decentralised network that makes use of idle hard drive space from people all around the world. The decentralised architecture is a viable alternative to centralised cloud storage and can address many issues that plague centralised systems.

3) Data management: Data management plays a crucial role in healthcare as it is one of the main factors when performing a risk assessment of patients. It contains data governance, cleaning, checking the data whether they have missing or unnecessary, and some process of extracting usable data in large datasets called data retrieval [25]. It is very important to maintain the confidentiality of individual patient records in healthcare management. Data governance, managed by GDPR, provides information about the basic legislation relating to health regulations and government regulations aimed at addressing the confidentiality of health data and refers to the overall management of the privacy, integrity, and availability of the data intended for use by an organization [26].

4) Data analytics: The stages in the process of converting raw data into valuable information are called data analysis which consists of four different types, Descriptive, diagnostic, predictive, and prescriptive analytics [27]. Descriptive analytics, also known as unsupervised learning, conducts a past performance evaluation depending on historical data. Diagnostic analysis forecasts the main reason for the problem by using historical data. Predictive analytics, also known as supervised learning, predicts what might happen in the future using historical data and real-time data. It is not used to forecast the future. Prescriptive analytics is used to get different possible outcomes by analysing the big data automatically before taking a decision.

5) Data visualisation: It is a method to present the result of healthcare data analytics to understand the complex healthcare data and help to make a better decision in a pictorial or graphical format. Data visualisation is also used to find out the pattern and correlation among the data.

B. Apache Hadoop for Big Data

Apache Hadoop is the most widely used open-source software platform for distributed analysis and storage of large datasets based on the MapReduce paradigm. It is built to handle petabytes of structured, semi-structured, and unstructured data, and it works as a cluster of ordered computers [28]. Hadoop basically consists of three main components: the first one is MapReduce, used for data processing; the second one is Hadoop Distributed File System (HDFS), used for data storage; and the last one is YARN (Yet Another Resource Negotiator), the resource management and job scheduling component. HDFS is a distributed file system designed to store and analyse large amounts of data for high-bandwidth applications [29]. The Hadoop framework's MapReduce paradigm makes it simple to convert and explore extensive data collections. Hadoop's most notable feature is distributing data and calculations across thousands of machines and performing the executions in parallel [30]. The five essential qualities of scalability, cost-effectiveness, flexibility, speed, and failure resistance are taken into consideration when designing Hadoop architecture [31]. MapReduce is a distributed computing system that uses a parallel and distributed method to process massive datasets designed to work with replicated files on HDFS [32]. It comprises the map and reduce functions and is carried out in three steps: map, shuffle, and reduce. A worker node analyses the data and generates key-value pairs, using the map function to break the incoming data into numerous little pieces. In the Shuffle stage, these pairings are divided by key. The responsible reducer receives each group. Finally, each reducer generates a new set of output stored in HDFS. After the mapping and reduction threads are completed, the resulting data is reassembled by Hadoop and presented to the user as a single output [33]. YARN performs all processing activities by allocating resources and scheduling tasks. The main idea behind YARN is to isolate the resource management and task scheduling/monitoring functions into independent daemons. It has two components: ResourceManager and NodeManager.

III. Proposed Method: MapChain

The architecture overview of the proposed MapChain is presented in Fig. 2. The system architecture consists of multiple components, including a comprehensive big data monitoring system, a visualization system, and a blockchain-based storage and verification system.

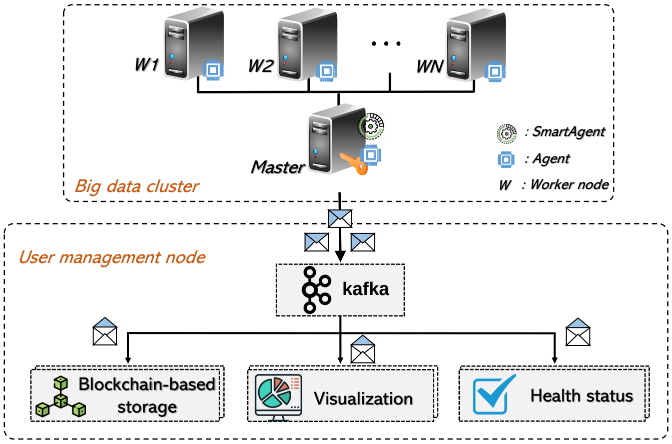


Fig. 2. The key design idea of the proposed system

The institution aims to perform the following specific tasks:

1) Linking the electronic recording of observations and NEWS data to other data held about the patient:
   - Patient demographics,
   - Hospital spell information (the type of admission, dates, speciality etc.),
   - Diagnosis and procedure coding,
   - Outcome: mortality, cardiac arrest and unplanned use of critical care.
2) Developing a way of handling the large volumes of data involved,
3) Modelling the sensitivity and specificity of NEWS and observations (i.e. what cut-offs best predict outcome),
4) Doing trajectories of scores (instead of individual scores) to improve sensitivity and specificity.

To this end, first, the NEWS data stored in MySQL database servers is transferred to HDFS in the Hadoop ecosystem for distributed processing. Fig. 3 depicts the data flow from the Galera cluster, a multi-master database cluster used for data replication synchronously for MySQL, to the big data system via Apache Sqoop. HAProxy, standing for High Availability Proxy, is used for high availability load balancing and application delivery controller in the cloud for TCP and HTTP-based applications. Apache Sqoop[4] is a command-line interface application developed by the Apache Software Foundation to transfer structured datasets stored in database systems such as MySQL to Apache Hadoop. The big data cluster, namely the Hadoop cluster, consists of a network of a single master and multiple worker nodes (W1, W2, ..., WN) that coordinates and carries out numerous tasks across the HDFS.

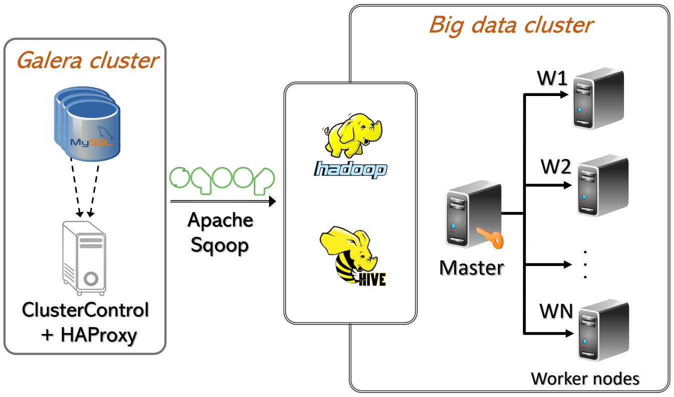


Fig. 3. Data migration from RDBMS to Apache Hadoop cluster

The overall system proposed in this paper consists of two main tasks: *Tracking the query and system health in Hadoop ecosystem* and *Blockchain-based storage and verification system for NEWS* as shown in Fig. 4. Here, we first explain the high-level system architecture of the proposed method and then describe the details of each component.

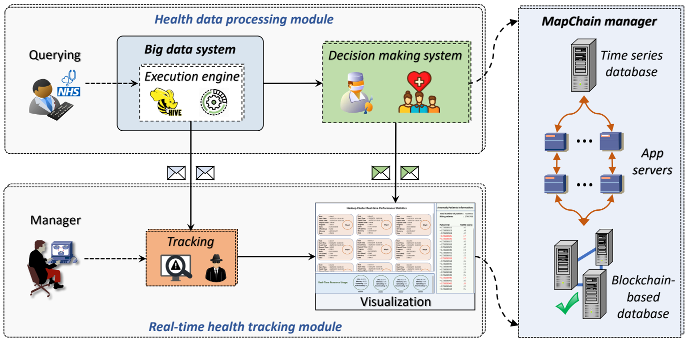


Fig. 4. The proposed system for healthcare data analytics in big data and blockchain environments

A. Tracking the query and system health in Hadoop ecosystem

The monitoring system has two main parts, namely *SmartAgent* and *Agents*. An agent, SIGAR library plugged, is deployed on each worker node to collect the resource utilization of the node, such as CPU/memory utilization, bandwidth usage, and I/O status. Using Yarn APIs, the *SmartAgent* is responsible for collecting tasks, applications and cluster information from worker nodes via TaskTrackers. *SmartAgent* is integrated into Apache Hive to listen to the Hive queries. This enables *SmartAgent* to track the tasks (mapper/reducer) related to queries, namely UDFs, as well as extract these queries results. The proposed framework shown in Fig. 4 demonstrates the processes of analyzing NEWS data in Apache Hive and monitoring the status of UDFs along with the performance of the big data system. The monitoring system extracts the query results after the query is completed and then transfers them to the *Decision-making system* integrated into the system through *SmartAgent*.

Query tracking. The queries defined by the users, namely UDFs, are tracked by *SmartAgent*. After the query syntax is verified, then the referenced objects in the query are validated whether they exist or not. After starting the execution of the query in Apache Hive, the MapReduce process will be monitored in real-time by *SmartAgent* that runs on top of SmartMonit [34], real-time big data monitoring system, which reports any failures or performance reduction in the cluster if it happens. *SmartAgent* finds the stragglers, which are the tasks that take longer to be completed than equivalent tasks running on the cluster synchronously. When a system exception, which is generated by the common language runtime, occurs during this process, *SmartMonit* uses counters, the proper channels, to gather statistics about the MapReduce job. SmartMonit agents track the *MapInputRecords*, *MapOutRecords*, and, *ShuffleErrors* counters to monitor progress inside UDFs. To report system exceptions to the user, the value of the relevant counter is increased and sent to the visualization system. Furthermore, *SmartAgent* obtains the query results and sends them to the *Decision-making system* to define the health status of the patients by calculating the NEWS. The collected time series data is also simultaneously injected into a time series database InfluxDB[5].

System health tracking. *SmartAgent* collects all the information related to the query tasks such as mapper and reducer, and *agents* collect the infrastructure information, such as CPU, memory, I/O, network, etc., in real-time. Yarn APIs are plugged into the *SmartAgent* while *agents* plugged SIGAR API. The visualization component is one of the parts of *System health tracking* that provides a user-friendly interface that enables users to see all the interactions of the system.

Decision-making system for NEWS. The main aim of NEWS is to standardise the assessment of acute illness severity in the NHS. The physiological parameters, which are *respiration rate, oxygen saturation, temperature, systolic blood pressure, heart rate, consciousness,* and *air or oxygen* are the key factors for calculating the NEWS. Fig. 5 shows these parameters and the thresholds and trigger values. An aggregate NEWS of 5 or 6 is a critical threshold that should prompt an urgent clinical review; a NEWS of 7 or higher should prompt a high-level clinical alert, i.e., an emergency clinical review. Moreover, as shown in Fig. 6, the final decision about patients is identified by the severity and NEW score. The clinical risk assessment is used in the visualization section. Algorithm 1 demonstrates the calculation of NEWS for each patient. NEWS is calculated for each patient using *Calc* function (see

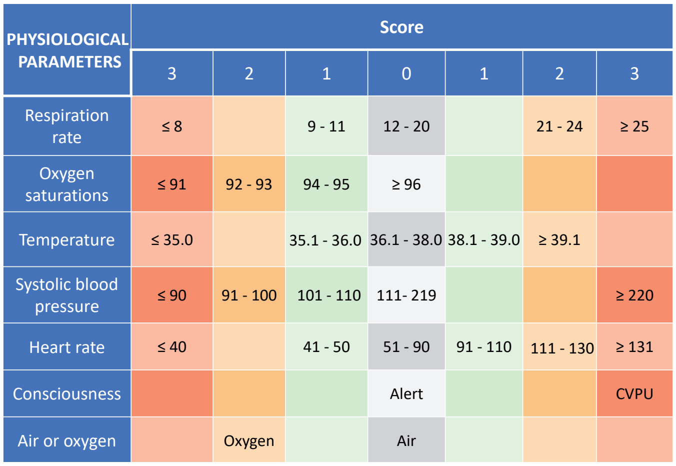
[4] https://sqoop.apache.org/

[5] https://www.influxdata.com/

Algorithm 1 NEWS calculation for each patient

```
Input:  $R_r$ - respiration rate,
        $O_s$ - oxygen saturations,
        $\mathcal{T}$ - temperature,
        $S_{bp}$ - Systolic blood pressure,
        $H_r$ - heart rate,
        $C_s$ - consciousness,
        $A/O$ - air or oxygen.
Output: $N_s$ - new score,
        $C_r$ - clinical risk.
1  // Calculate $N_s$ for each input
2  Function Calc($R_r$, $O_s$, $\mathcal{T}$, $S_{bp}$, $H_r$, $C_s$, $A/O$):
3      // Calculate, update, and return the $N_s$
4      $N_s \leftarrow$ Update ($N_s^{new}$, $N_s$)
5  return
6  for each patient do
7      // Clear the $N_s$, and $C_r$
8      $N_s \leftarrow$ Clear ($N_s^{new}$, $N_s$)
9      $C_r \leftarrow$ Clear ($C_r^{new}$, $C_r$)
10     // Call the Calc function
11     Calc()
12     if ($N_s >= 5$) && ($N_s <= 6$) then
13         // Prompt an urgent clinical review
14         $C_r \leftarrow$ Prompt Urgent response
15     end
16     if ($N_s >= 7$) then
17         // Prompt a high-level clinical alert
18         $C_r \leftarrow$ Prompt Emergency response
19     end
20     $C_r \leftarrow$ Update ($C_r^{new}$, $C_r$)
21 end
```

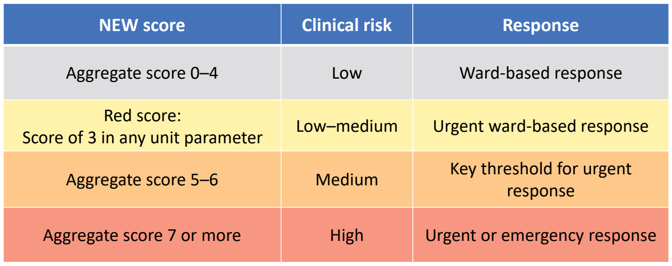
Algorithm 1, Line 4). Then, the patients status are identified based on the thresholds and triggers specified in Fig. 6 based on their NEW scores (see Algorithm 1, between Line 13 and Line 18). Finally, NEWS is updated to be sent to blockchain-based system (see Algorithm 1, Line 20).



PHYSIOLOGICAL PARAMETERS	Score						
	3	2	1	0	1	2	3
Respiration rate	≤ 8		9 - 11	12 - 20		21 - 24	≥ 25
Oxygen saturations	≤ 91	92 - 93	94 - 95	≥ 96			
Temperature	≤ 35.0		35.1 - 36.0	36.1 - 38.0	38.1 - 39.0	≥ 39.1	
Systolic blood pressure	≤ 90	91 - 100	101 - 110	111- 219			≥ 220
Heart rate	≤ 40		41 - 50	51 - 90	91 - 110	111 - 130	≥ 131
Consciousness				Alert			CVPU
Air or oxygen		Oxygen		Air			

Fig. 5. The NEWS scoring system [9]

Visualization. All the collected data including the system and the results of the query is sent to the database and the visualization system through Apache Kafka[6], used for building a real-time streaming data pipeline in a secure way. Basically, the *Visualization* component has two different parts, namely the *query engine* and the *user interface*. The database which stores all the collected data is queried by the query engine. The web-based user interface is fed by the data provided by the query engine at a pre-defined time interval. The user interface shows this data along with the NEWS scores provided by *SmartAgent* after the query is completed. All the data is shown to allow users to quickly comprehend the anomalies in both the system and the patients through this user-friendly interface. The values representing different information are represented by various colours in the user interface. While task status and stragglers are shown in orange, the resource utilization of the master node and each worker node is represented in blue. While task status and stragglers are marked in orange, the resource utilization of the master node and each worker node is marked in blue. Moreover, all NEWS scores are shown on a green background. Patients with normal values are shown in black, while the information of patients in danger is indicated in red.



NEW score	Clinical risk	Response
Aggregate score 0–4	Low	Ward-based response
Red score: Score of 3 in any unit parameter	Low–medium	Urgent ward-based response
Aggregate score 5–6	Medium	Key threshold for urgent response
Aggregate score 7 or more	High	Urgent or emergency response

Fig. 6. NEWS thresholds and triggers [9]

We created the execution graph using different technologies to take the collected information from the database and visualize it in real-time. These technologies are HTML, CSS, and PHP, which help improve the graph's functionality and efficiency. The basic structure of our execution graph is created using HTML, such as the positions of the circles symbolizing the headers. We have also structured parts of our execution graph such as sections, paragraphs, headings, and links using HTML. We used CSS to color the fields and highlight the arrows indicating the data for each mapper and reducer and the communication between the mappers and reducers. PHP has the critical role of fetching all the metrics stored in the InfluxDB database within a specific time interval to visualize it in our execution graph.

B. Derived Blockchain Framework for NEWS

Following the advent of Bitcoin, blockchain has been accepted by various groups and has been used in a variety of applications. Various implementations are available now that provide a platform for conducting business with simplicity. However, it has been noted that blockchain scalability is still a concern. Furthermore, none of the frameworks can claim to be able to manage big data and enable analytics, which is a

[6] https://kafka.apache.org/

significant and vital aspect of today's corporate sector. That is why, instead of using a blockchain system as a storage, it would be more beneficial to leverage the immutable and verifiable nature of blockchain. Due to this reason, in this work, we use a decoupled blockchain architecture used in our previous works [17]–[19]. In this architecture, only the hash of the NEWS (score) is stored in the block, and the rest of the sensitive data is stored in off-chain storage. NEWS for each patient generated as a result of the execution of the queries will be stored on a blockchain-based database by building over various distributed databases like MongoDB in the public cloud as it allows access to data and blockchain-enabled applications. By this means, blockchain can offer distributed retention of encrypted data in the public cloud. So, integrating blockchain and public cloud technologies will ensure consistent performance and data integrity. Even this will help make NEWS data transparent and even ensure backward (as well as forward) traceability to verify the data stored in the off-chain storage. The steps performed in this system are listed below as simulated in Fig. 7.

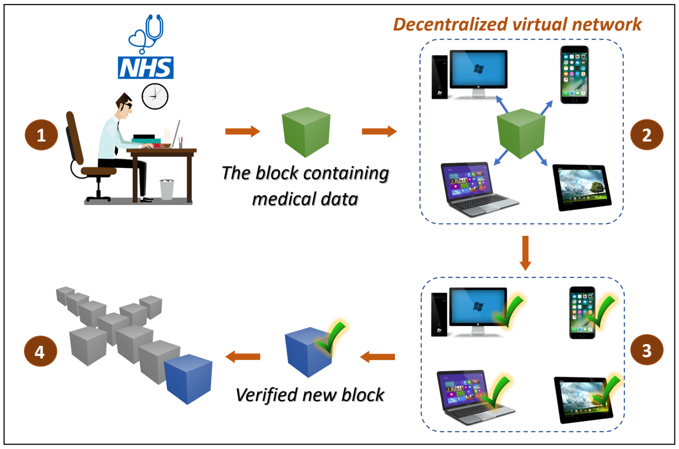

NHS
Decentralized virtual network
1
The block containing medical data
2
3
4
Verified new block

Fig. 7. A conceptual blockchain-based ecosystem

- **Step 1:** Healthcare personnel perform data analytics on the patient's EHR and generates NEWS that is stored on the blockchain.
- **Step 2:** The block is delivered to all the peers in the patient's network, such as the patient itself, doctors, researchers, and hospitals in a decentralized virtual network.
- **Step 3:** The block is verified and approved by all the nodes in the network.
- **Step 4:** The block is inserted in the chain and linked with the previous blocks to provide a permanent and transparent record.

The proposed derived blockchain process works in two phases, i.e., identity registration and verification phase and block generation phase. All these phases are explained below.

1) Registration and Verification Phase: Allowing a legitimate person to store and access the data on the blockchain is one of the key challenges in the healthcare sector. To ensure the same, we have proposed a mechanism based on zero knowledge protocol (ZKP). This step of the blockchain process is to register and verify the legitimate users (or patients) ($\mathbf{U}_i$) whenever they want to register on the MapChain or access the NEWS score on the blockchain.

- Initially, the $\mathbf{U}_i$ sends a request ($\mathbf{RQ}_i$) to the MapChain Manager ($\mathbf{M}_j$) to register and access the blockchain. For this purpose, a temporary key ($\mathbf{TK}_i$) is generated for the requesting $\mathbf{U}_i$. Here, $\mathbf{TK}_i$ is generated based on $\mathbf{U}_i$ and device credentials ($\mathbf{DC}_i$) (such as device ID, MAC address). Let us say, $\mathbf{DC}_i$ consists of $\mathbf{DID}_i$ and $\mathbf{DMAC}_i$. These attributes are used to generate $\mathbf{DC}_i$ as shown below.
$$\mathbf{DC}_i = \mathbf{H}\big[\mathbf{DID}_i, (\mathbf{DMAC}_i \bigoplus \mathbf{RN})\big] \tag{1}$$
where, $\mathbf{RN}$ is a pseudo-random number used to hide the MAC address of the sensor or device associated with $\mathbf{U}_i$. The timestamp ($\mathbf{TS}_i$) at which $\mathbf{TK}_i$ is generated is also recorded to streamline the sequence of request generation. The $\mathbf{TK}_i$ along with $\mathbf{RQ}_i$ is sent to $\mathbf{M}_j$ over a secure SSL/TLS channel for further processing.
- Once $\mathbf{M}_j$ receives $\mathbf{RQ}_i$, it extracts $\mathbf{TK}_i$ and check $\mathbf{RQ}_i$ for any format or script errors. After $\mathbf{RQ}_i$ is checked, an unverified identity ($\mathbf{ID}_i$) is generated. This is a unverified identity issued to the user request and it is verified only after the success of ZKP process.
- In the next step, the signal flag ($\mathbf{S_{ZKP}}$) is sent to $\mathbf{U}_i$ indicating to start the ZKP process.
- The ZKP process is initialised wherein $\mathbf{U}_i$ is the prover and $\mathbf{M}_j$ is the verifier (challenger). ZKP is used to authenticate the legitimacy of the two parties without revealing any secret information to the verifier. In this process, we compute the value of y based on a large prime number p and a generator g as shown below.
$$y = g^{\mathbf{TS}_i} mod\, p \tag{2}$$
$\mathbf{U}_i$ has to prove that it has the knowledge of $\mathbf{TS}_i$ based on the value of y but can reveal its value during the whole process. Next, $\mathbf{U}_i$ computes the value of r, i.e., a random value used to further computed d as shown below.
$$r = \mathbf{TS}_i + \mathbf{RQ}_i \tag{3}$$
$$d = g^r mod\, p \tag{4}$$
In the next step, d is sent to $\mathbf{M}_j$ over a secure channel.
- The $\mathbf{M}_j$ extracts d and selects two queries (questions) as shown below.
$$Q_1 \rightarrow \mathbf{TS}_{i+\mathbf{RQ}_i} \tag{5}$$
$$Q_2 \rightarrow \mathbf{TS}_i + \mathbf{TS}_{i+\mathbf{RQ}_i} mod(p-1) \tag{6}$$
The question is transmitted to $\mathbf{U}_i$ over a secure channel.
- On the receipt of the question, $\mathbf{U}_i$ selects the appropriate answer (A). This answer is transmitted to $\mathbf{M}_j$.
- In the final step, the answer is verified by $\mathbf{M}_j$ based on the following conditions.
$$g(^{\mathbf{TS}_{i+\mathbf{RQ}_i}})modp == x \tag{7}$$
$$g(^{\mathbf{TS}_{i+\mathbf{RQ}_i} mod(p-1)})mod\, p = x.u\, mod\, p \tag{8}$$

After this, the identity of $\mathbf{U}_i$ is verified and assigned $\mathbf{ID}_i$, and it can use it to store or access the NEWS data on the blockchain. The entire workflow of this phase is shown in Fig. 8.

User (Patient) $\mathbf{U}_i$	MapChain Manager $\mathbf{M}_j$
INITIATE: $\mathbf{RQ}_i$ $\mathbf{TK}_i \leftarrow [\mathbf{U}_i, \mathbf{DC}_i]$ RECORD: $\mathbf{TS}_i$	
$\xrightarrow[\text{(SSL/TLS)}]{\langle \mathbf{TK}_1(\mathbf{RQ}_i) \rangle}$	
	EXTRACT: $\mathbf{TK}_i(\mathbf{RQ}_i)$ CHECK: $\mathbf{RQ}_i$ GENERATE: $\mathbf{ID}_i$ [unverified] SIGNAL: $\mathbf{ID}_i$ for **ZKP**
	$\xleftarrow[\text{(SSL/TLS)}]{\langle \mathrm{S}_{\mathrm{ZKP}} \rangle}$
INITIALISE ZKP COMPUTE: $y = g^{\mathbf{TS}_i} mod\, p$ COMPUTE: $r \leftarrow \mathbf{TS}_i + \mathbf{RQ}_i$ COMPUTE: $d = g^r mod\, p$	
$\xrightarrow[\text{(SSL/TLS)}]{\langle \mathbf{d} \rangle}$	
	EXTRACT: $d \rightarrow g^r mod\, p$ QUESTIONS: Q_1 or Q_2 (Eqs. 5 and 6) SEND: $Q = (Q_1/Q_2)$
	$\xleftarrow[\text{(SSL/TLS)}]{\langle Q \rangle}$
EXTRACT: Q IF Q == Q_1 **SEND** A_1 ELSE **SEND** A_2 $A = (A_1/A_2)$	
$\xrightarrow[\text{(SSL/TLS)}]{\langle A \rangle}$	
	EXTRACT: A IF A == A_1 VERIFY: Eq. 7 ELSE COMPUTE: Eq. 8 VERIFIED REGISTER $\mathbf{ID}_i$ and ALLOW ACCESS

Fig. 8. Registration and Verification Process

2) Block Creation and Update Phase: This phase comprises the process of block creation and thereafter updation once a new transaction is generated. After the $\mathbf{U}_i$ is registered with $\mathbf{M}_j$, the following steps are performed.

- The $\mathbf{U}_i$ generates a key pair ($\mathbf{PUK}_i$, $\mathbf{PRK}_i$) comprising public and private keys. These keys are generated for a limited session to ensure identity de-linking and a time-out tag ($\mathbf{T}_{TO}$) is attached to them. The keys get invalided once the $\mathbf{T}_{TO}$ expires. Now, $\mathbf{U}_i$ send a request ($\mathbf{R}_{T_i}$) to $\mathbf{M}_j$ for initiating the transaction ($\mathbf{T}_i$). Now, $\mathbf{R}_{T_i}$ is signed using $\mathbf{PRK}_i$ and sent to $\mathbf{M}_j$. On receipt, $\mathbf{M}_j$ authenticates the request using $\mathbf{PUK}_i$. Once, the request is verified, a new block ($\mathbf{B}_i$) is generated by $\mathbf{M}_j$ for requesting users.
- After this the NEWS score (N_{S_i}) generated for each $\mathbf{U}_i$ is generated as a transaction ($\mathbf{T}_i$). This transaction has to be added to the blockchain. But the complete data (N_{D_i}) related to the computation of the NEWS score is stored in off-chain mode. However, the hash of the data is appended with the NEWS score and recorded on the blockchain for verification at any later stage. This is done to ensure data integrity and at the same time ensure optimal usage of storage on the blockchain.
- The NEWS score and related hashed data is signed by $\mathbf{U}_i$ using its $\mathbf{PRK}_i$ as shown below.

$$\mathbf{SIGN}_i = \left[N_{S_i}, \mathbf{H}(N_{D_i}), \mathbf{PRK}_i\right] \qquad (9)$$

Now, the $\mathbf{T}_i$ is generated as shown below.

$$\mathbf{T}_i = \left[N_{S_i}, \mathbf{H}(N_{D_i}), \mathbf{PBK}_i, \mathbf{ID}_i, \mathbf{SIGN}_i\right] \qquad (10)$$

- After this, $\mathbf{T}_i$ is sent to $\mathbf{M}_j$ for validation. $\mathbf{M}_j$ checks the $\mathbf{PBK}_i$, and if they are verified then $\mathbf{T}_i$ is added to $\mathbf{B}_i$. Similarly, all the transactions received from various users are added to the block.
- The block is finally sent to the blockchain network for consensus and validation. Once validated by the peers, it is appended to the blockchain, and the whole blockchain is synchronised in the network.

The block creation and update process in the proposed method is shown in the Fig. 9.

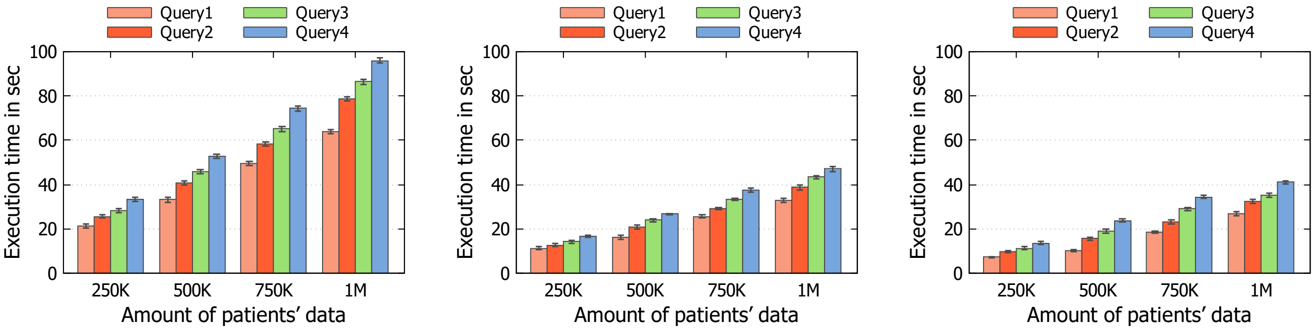
User (Patient) $\mathbf{U}_i$	MapChain Manager $\mathbf{M}_j$	Blockchain network
GENERATE: ($\mathbf{PUK}_i, \mathbf{PRK}_i$) SET: ($\mathbf{T}_{TO}$) START: ($\mathbf{R}_{T_i}$) SGN: ($\mathbf{R}_{T_i}$, $\mathbf{PRK}_i$)		
$\xrightarrow[\text{(SSL/TLS)}]{\langle \mathbf{R}_{T_i}, \mathbf{PRK}_i \rangle}$		
	AUTHENTICATE: $\mathbf{R}_{T_i}$ using $\mathbf{PUK}_i$ CREATE: $\mathbf{B}_i$ GENERATE: $\mathbf{SGN}_j$	
	$\xleftarrow[\text{(SSL/TLS)}]{\langle \mathbf{SGN}_j \rangle}$	
VALIDATE: ($\mathbf{SGN}_j$) IF ($\mathbf{SGN}_j$ == VALID) THEN READ: N_{D_i} GENERATE: N_{S_i} using Algorithm 1 GENERATE: $\mathbf{H}(N_{D_i})$ SIGN: $\mathbf{SIGN}_i = \left[N_{S_i}, \mathbf{H}(N_{D_i}), \mathbf{PRK}_i\right]$ GENERATE: $\mathbf{T}_i$ using Eq. 10 SEND: ($\mathbf{T}_i$)		
$\xrightarrow[\text{(SSL/TLS)}]{\langle \mathbb{T}_i \rangle}$		
	VALIDATE: $\mathbf{ID}_i, \mathbf{PBK}_i$ CHECK: SGN_i VALIDATE: SGN_i APPEND: $\mathbf{T}_i$ to $\mathbf{B}_i$ BROADCAST: $\mathbf{B}_i$	
	$\xleftarrow[\text{(SSL/TLS)}]{\langle \mathbf{G}_i^{\mathrm{TAG}} \rangle}$	
		Validate: $\mathbf{B}_i$ Synchronize BC

Fig. 9. Block creation and validation process

IV. Results and Discussions

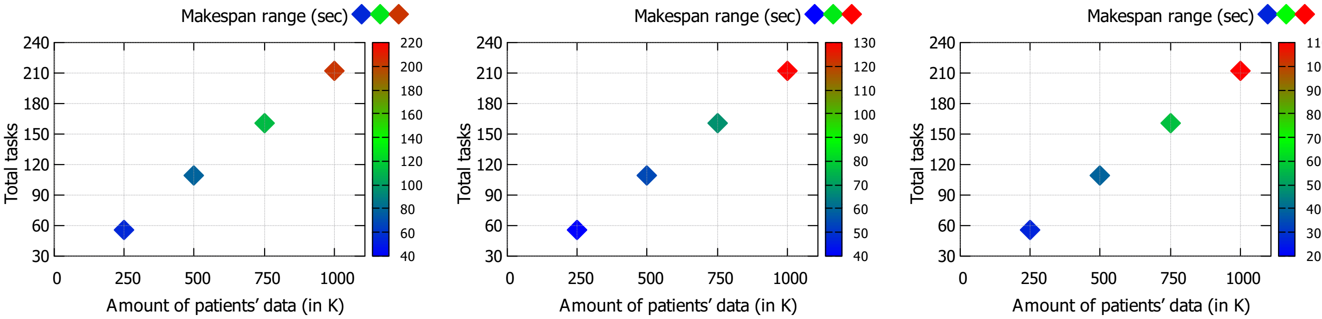
This section demonstrates the experimental results of MapChain to test its efficiency and applicability, as well as its resource usage and overheads, in big data-driven IoT for healthcare analytics. To verify the efficiency and reliability of the derived blockchain-based verifiable big data ecosystem for healthcare IoT, we use the NEWS data[7], generated based on the structure and values taken from South Tees Hospitals NHS Foundation Trust, consisting of blocks of different patient numbers, such as 250K, 500K, 750K, 1M. The dataset contains 52 different features regarding patients' health status, including pulse, pain score, nausea, vomiting, weight, the dates of admission, duration of hospital stay, etc., along with the seven different physiological parameters.

[7] https://github.com/umitdemirbaga/NEWS



(a) Cluster Type 1 (b) Cluster Type 2 (c) Cluster Type 3

Fig. 10. Comparison of the execution time of the tasks based on the number of patients and the types of clusters



(a) Cluster Type 1 (b) Cluster Type 2 (c) Cluster Type 3

Fig. 11. Comparison of makespan and number of tasks based on the number of patients and types of clusters

A. *Evaluation of the health status of the system*

Experimental setup. To perform the performance evaluation of the monitoring system, we deployed three different Hadoop YARN clusters on AWS: **Type 1:** 1 master and 3 workers, **Type 2:** 1 master and 6 workers, and finally **Type 3:** 1 master and 9 workers. We chose Ubuntu Server 20.04 LTS (HVM), SSD Volume Type as the operating system for all the nodes and the Hadoop version 3.3.2 and the Hive version 3.1.2. All the nodes have the same configuration (i.e., 4 cores and 16 GB of memory).

Query performance evaluation. We tested the performance of four different complex Hive queries on NEWS datasets. Fig. 10(a), Fig. 10(b), and Fig. 10(c) demonstrate the total execution time of mapper and reducer tasks on four different datasets while executing the queries on different clusters. The results show that the job completion time decreases as the number of machines increases. The average and standard deviation are reported for each query after being repeated five times. All the information is gathered through *SmartAgent* and *Agents* and then is sent to the visualization component.

Makespan and the number of tasks evaluation. We compared the makespan (the total time taken to complete a job) and the number of tasks based on the different datasets. Fig. 11(a), Fig. 11(b), and Fig. 11(c) show the number of the tasks launched relying on four different datasets, namely 250K, 500K, 750K, and 1000K. These figures also demonstrate the makespan regarding the datasets. The main difference between the figures is the makespan as the total number of tasks does not change according to the types of the clusters.

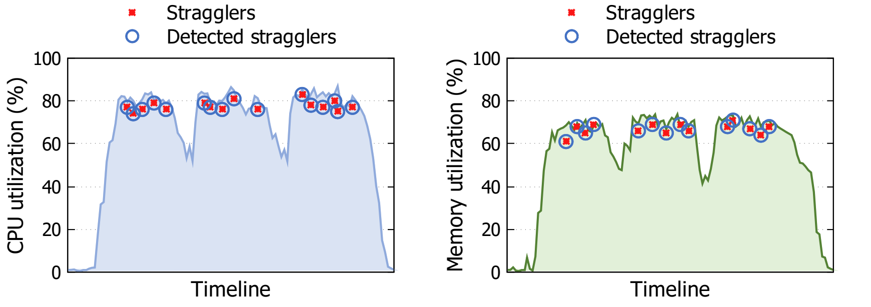
Tasks status evaluation. We also evaluated the straggler detection during NEWS data processing on the big data system. The results shown in Fig. 12 show the straggler detection under high resource utilization conditions. Fig. 12(a) demonstrates the results of straggler detection while CPU utilization is high. Similarly, Fig. 12(b) shows the results while memory usage is high. *SmartAgent* simultaneously processes the collected data to track the health status of the big data processing system in real-time that detects the stragglers. Both results were obtained from the analysis of data of 500K patients.

B. *Evaluation of the Decision-making system for NEWS*

Fig. 13 shows the effect of different physiological parameters on NEWS. For example, Fig. 13(a) demonstrates the impact of respiration rate and oxygen saturations on the severity of NEWS, while Fig. 13(b) presents the effect of temperature and systolic blood pressure. We also examine the relationship between respiration rate and heart rate in Fig. 13(c) for NEWS. As indicated in Fig. 6, NEWS values above 7 represent the *"high clinical risk"* and the response is *"Urgent or emergency response"*. All the values in the figures represent different patients' data, and these results were obtained by analyzing data from 500 patients.

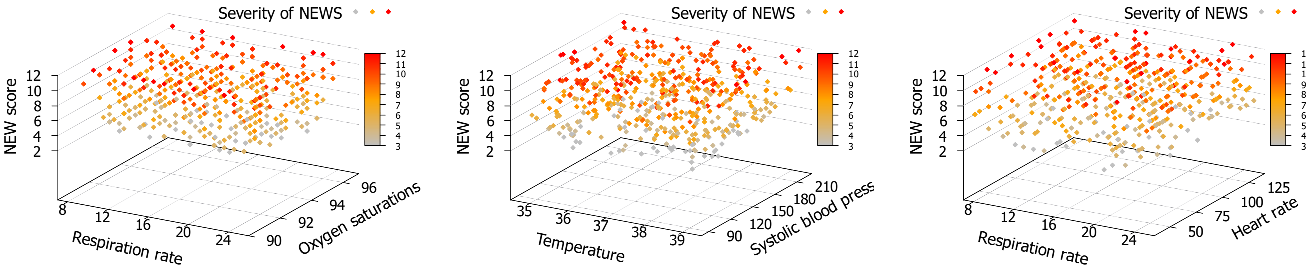
C. *Validation of Blockchain system*

The blockchain experiments are performed to evaluate the cost incurred on the transactions when they are deployed and executed. For this purpose, we develop a prototype based



(a) Detected stragglers while high CPU usage (b) Detected stragglers while high memory usage

Fig. 12. Task tracking evaluation under high resource utilization



(a) Correlation 1 (b) Correlation 2 (c) Correlation 3

Fig. 13. Correlation between NEWS score and other values

on Ropsten[8], i.e., a public test network for blockchain. The performance of the proposed smart contracts is measured based on the amount of *gas consumed* during the execution of the transactions. Further, we investigate the *mining time* for the executed transactions. To realize the above experiment, we implement the smart contracts using Solidity[9], that were complied using Remix compiler[10]solidity working on Ethereum virtual machine.

The proposed system has been validated based on the gas consumed during the execution of the deployed smart contracts. We have considered two cases, i.e., 100 and 200 users, respectively, for comparison of experimental results. The results were averaged after performing the experiments five times. Table I shows the results obtained from the experiments. The average cost for storing the NEW score (and hashed data) on the blockchain is shown in Table I. The gas consumption increases with the number of users clearly depicting the impact of the amount of data generated on the performance. The increase in the number of users is aligned with an increase in the NEWS data generated by the big data system. This tends to generate additional overhead in terms of gas consumption.

Similarly, the gas consumed for the verification process and transaction generation show a similar kind of trend. The main reason for the difference in gas consumption is concerned with the increase in the number of users. The increase in the number of users interacting with the big data systems tends to incorporate additional overheads in terms of gas costs. The percentage increase in the verification process also acknowledges the role of the ZKP process used to authenticate the users. Finally, we investigate the mining time for the blockchain process. The mining time for 100 users is 289 secs as compared to 358 secs for 200 users. This again justifies the impact of an increase in the number of users with respect to the performance of the proposed MapChain study.

TABLE I
EVALUATION RESULTS

Number of users	100	200
Storage (wei)	88756	133214
Verification (wei)	804531	1896788
Transaction generation (wei)	487532	890504
Mining time (seconds)	289	358

V. Related Work

Many researchers have investigated health-related solutions by utilizing IoT networks. The authors in [35] highlighted methods related to cloud computing and big analytics, as well as the use of smart devices for the analysis of health data collected in IoT networks. The authors of [36] proposed a paradigm for structural health monitoring (SHM) that is based on IoT technologies and includes intelligent and consistent observation. Furthermore, the data routing plan is given with

[8] https://github.com/ethereum/ropsten
[9] https://docs.soliditylang.org/en/v0.8.13/
[10] https://remix.ethereum.org

the help of big data analytics. In terms of scalability and low latency, the suggested framework increased performance. Oueida *et al.* [37] employed the edge computing paradigm to describe time-independent resources to manage resources in a smart healthcare framework in a more effective manner. The authors of [38] aimed to alleviate the burden of some tasks such as local data storage and real-time data processing from the cloud to edge gateways by utilizing gateways at the edge of healthcare IoT networks. These authors argued that using these methods can address different challenges in the healthcare environment. However, they did not consider that these edge gateways are vulnerable to privacy concerns and what actions could be taken to improve the integrity of the data.

Blockchain technology, yet another promising technology solution for securely transferring data from edge devices to the cloud, has been used by a number of researchers to ensure privacy and integrity in the underlying network in a range of applications. The authors of [39] leveraged edge-as-a-service to enable blockchain-based energy trading in smart grid applications. Clauson *et al.* [40] employed blockchain in the field of healthcare to improve supply chain management, reducing fraud and mistakes while also increasing confidence in the supply chain process. Nguyen *et al.* [41] developed a data offloading and data-sharing model for healthcare systems between edge and cloud using blockchain technology. To this end, first, they deployed a data offloading model for IoT health devices with privacy awareness. After that, a blockchain-based data-sharing model is deployed for enabling data exchange between users. The authors in [42] proposed a hierarchical blockchain system based on edge-cloud infrastructure to reduce the response time for massive-scale IoT devices between edge nodes and cloud environments. Then, they stored the blockchain data in various distributed clouds and edge nodes to increase scalability. While these articles focus on data sharing among the users and increasing the usefulness of their systems, they do not provide authentication to access and modify data as we did in our study.

Besides this, blockchain is commonly utilised in healthcare to ensure access control and authentication requirements and enable non-repudiation and interoperability. Survey studies presented in [43], [44], [45] gave insight into how blockchain technology is used to address data integrity and verification concerns by various researchers. While [44] indicated the potential of blockchain technology in the health sector and drew attention to the potential future application of this rapidly emerging field, the authors in [45] emphasized the difficulties with blockchain in processing big health data, especially the scalability issues. The scalability issue arises with the increasing number of intermediary devices and transactions in blockchain while the resources are limited. To overcome this issue, a decoupled blockchain approach is offered in [46], [47]. The authors in [46] proposed separate storage locations for block headers and block ledgers for smart city services running on multiple IoT devices. [47], the authors proposed a public blockchain of a two-chain structure to overcome the scalability issue in fog and IoT computing environments. Shukla *et al.* [48] proposed a solution for data validation in a decentralized environment using blockchain technology in the fog computing environment. This system, in which Advanced Signature Based Encryption (ASE) algorithm is applied and built on Fog computing, consists of an analytical model, mathematical framework and IoT device identification layers. The proposed solution allows users to transmit real-time data collected via IoT devices with better authentication.

In [49], the authors have proposed a data sharing and integration mechanism based on blockchain. This is achieved through a web interface where the user can share the EHR. The metadata is stored on-chain, whereas the actual EHR is stored off-chain in encrypted form. However, this work has not considered ZKP or any other method to authenticate or hide the user identity. Moreover, it doesn't consider the big data management systems wherein the actual decision-making is performed on big healthcare data. In another work [50], the authors proposed a national blockchain scheme for managing access control and funds concerning the EHR. This work aims to provide a transparent insurance claim mechanism alongside ensuring the audit trail of events through smart contracts. ZKP protocol was used to ensure user identity authentication. However, it has not considered any big data processing systems. Our proposed scheme incorporated blockchain alongside a big data analytical system comprising decision-making and monitoring systems as crucial components. Although ideas on how to use blockchain technology to address privacy and integrity concerns have been suggested by various researchers, there is no study on how to eliminate data integrity and privacy concerns related to big data using the ZKP that we recommend in our MapChain study.

VI. Conclusion

Cloud-based big data analytics systems bring advantages to business and management community services regarding cost savings, new product development, real-time decision making, and time reduction. At the same time, blockchain technology eliminates the privacy and trust issues in the healthcare and finance industries. In our work, we emphasized the importance of analysing the complex and huge amount of healthcare data and storing this information in inaccessible storage in an authorized blockchain-based system. We identified a secure and scalable healthcare data analytics framework by combining big data processing technologies with the blockchain paradigm to increase the service quality in the healthcare industry, which requires the collaboration of computer scientists and health science experts.

Scalability and privacy are two important issues in healthcare-based big data analytics in IoT environments. MapChain overwhelms these problems by proposing a robust, scalable, and user authenticator data sharing and storage system. In the future, we will integrate advanced communication technologies, such as 5G, satellite, navigation systems, and software-defined networks to increase the performance of underlying network services.

Acknowledgements

This research is funded by the Republic of Türkiye Ministry of National Education and partially funded through Durham

University Start Up Grant (090614) and Durham University SeedCorn Grant (RQ090002). Umit Demirbaga is fully funded by the University of Cambridge, UK.

References

[1] S. M. Idrees, M. A. Alam, and P. Agarwal, “A prediction approach for stock market volatility based on time series data,” *IEEE Access*, vol. 7, pp. 17 287–17 298, 2019.

[2] N. Khan, M. Alsaqer, H. Shah, G. Badsha, A. A. Abbasi, and S. Salehian, “The 10 vs, issues and challenges of big data,” in *Proceedings of the 2018 international conference on big data and education*, 2018, pp. 52–56.

[3] B. Cyganek, M. Graña, B. Krawczyk, A. Kasprzak, P. Porwik, K. Walkowiak, and M. Woźniak, “A survey of big data issues in electronic health record analysis,” *Applied Artificial Intelligence*, vol. 30, no. 6, pp. 497–520, 2016.

[4] K. Alwasel, D. N. Jha, F. Habeeb, U. Demirbaga, O. Rana, T. Baker, S. Dustdar, M. Villari, P. James, E. Solaiman *et al.*, “Iotsim-osmosis: A framework for modeling and simulating iot applications over an edge-cloud continuum,” *Journal of Systems Architecture*, vol. 116, p. 101956, 2021.

[5] A. Jindal, A. Dua, N. Kumar, A. V. Vasilakos, and J. J. Rodrigues, “An efficient fuzzy rule-based big data analytics scheme for providing healthcare-as-a-service,” in *2017 IEEE international conference on communications (ICC)*. IEEE, 2017, pp. 1–6.

[6] S. Li, L. Kang, and X.-M. Zhao, “A survey on evolutionary algorithm based hybrid intelligence in bioinformatics,” *BioMed research international*, vol. 2014, 2014.

[7] M. Ross, W. Wei, and L. Ohno-Machado, ““big data” and the electronic health record,” *Yearbook of medical informatics*, vol. 23, no. 01, pp. 97–104, 2014.

[8] W. Raghupathi and V. Raghupathi, “An overview of health analytics,” *J Health Med Informat*, vol. 4, no. 132, p. 2, 2013.

[9] Partnering for health early warning systems. World Health Organization. [Online]. Available: https://public.wmo.int/en/bulletin/partnering-health-early-warning-systems

[10] A. McGinley and R. M. Pearse, “A national early warning score for acutely ill patients: A new standard should help identify patients in need of critical care,” *BMJ: British Medical Journal*, vol. 345, no. 7869, pp. 9–9, 2012.

[11] U. Demirbaga, Z. Wen, A. Noor, K. Mitra, K. Alwasel, S. Garg, A. Zomaya, and R. Ranjan, “Autodiagn: An automated real-time diagnosis framework for big data systems,” *IEEE Transactions on Computers*, 2021.

[12] I. Stellios, P. Kotzanikolaou, M. Psarakis, C. Alcaraz, and J. Lopez, “A survey of iot-enabled cyberattacks: Assessing attack paths to critical infrastructures and services,” *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3453–3495, 2018.

[13] M. A. Amanullah, R. A. A. Habeeb, F. H. Nasaruddin, A. Gani, E. Ahmed, A. S. M. Nainar, N. M. Akim, and M. Imran, “Deep learning and big data technologies for iot security,” *Computer Communications*, vol. 151, pp. 495–517, 2020.

[14] M. P. Singh, G. S. Aujla, and R. S. Bali, “An unorthodox security framework using adapted blockchain architecture for internet of drones,” in *2021 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2021, pp. 1–6.

[15] I. Konstantinidis, G. Siaminos, C. Timplalexis, P. Zervas, V. Peristeras, and S. Decker, “Blockchain for business applications: A systematic literature review,” in *International conference on business information systems*. Springer, 2018, pp. 384–399.

[16] P. Zhang, X. Pang, N. Kumar, G. S. Aujla, and H. Cao, “A reliable data-transmission mechanism using blockchain in edge computing scenarios,” *IEEE Internet of Things Journal*, 2020.

[17] M. Singh, G. S. Aujla, and R. S. Bali, “Odob: One drone one block-based lightweight blockchain architecture for internet of drones,” in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2020, pp. 249–254.

[18] ——, “Derived blockchain architecture for security-conscious data dissemination in edge-envisioned internet of drones ecosystem,” *Cluster Computing*, pp. 1–22, 2022.

[19] G. S. Aujla and A. Jindal, “A decoupled blockchain approach for edge-envisioned iot-based healthcare monitoring,” *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 2, pp. 491–499, 2020.

[20] R. Nambiar, R. Bhardwaj, A. Sethi, and R. Vargheese, “A look at challenges and opportunities of big data analytics in healthcare,” in *2013 IEEE international conference on Big Data*. IEEE, 2013, pp. 17–22.

[21] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, “Efficient machine learning for big data: A review,” *Big Data Research*, vol. 2, no. 3, pp. 87–93, 2015.

[22] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, “The rise of “big data” on cloud computing: Review and open research issues,” *Information systems*, vol. 47, pp. 98–115, 2015.

[23] A. Noor, K. Mitra, E. Solaiman, A. Souza, D. N. Jha, U. Demirbaga, P. P. Jayaraman, N. Cacho, and R. Ranjan, “Cyber-physical application monitoring across multiple clouds,” *Computers & Electrical Engineering*, vol. 77, pp. 314–324, 2019.

[24] S. El Kafhali, I. El Mir, and M. Hanini, “Security threats, defense mechanisms, challenges, and future directions in cloud computing,” *Archives of Computational Methods in Engineering*, vol. 29, no. 1, pp. 223–246, 2022.

[25] J. Archenaa and E. M. Anita, “A survey of big data analytics in healthcare and government,” *Procedia Computer Science*, vol. 50, pp. 408–413, 2015.

[26] 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.

[27] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, “Big data: Issues and challenges moving forward,” in *2013 46th Hawaii international conference on system sciences*. IEEE, 2013, pp. 995–1004.

[28] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, “The hadoop distributed file system,” in *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*. Ieee, 2010, pp. 1–10.

[29] H. E. Ciritoglu, J. Murphy, and C. Thorpe, “Hard: a heterogeneity-aware replica deletion for hdfs,” *Journal of big data*, vol. 6, no. 1, pp. 1–21, 2019.

[30] U. Demirbaga, “Htwitt: a hadoop-based platform for analysis and visualization of streaming twitter data,” *Neural Computing and Applications*, pp. 1–16, 2021.

[31] L. Wang, J. Tao, R. Ranjan, H. Marten, A. Streit, J. Chen, and D. Chen, “G-hadoop: Mapreduce across distributed data centers for data-intensive computing,” *Future Generation Computer Systems*, vol. 29, no. 3, pp. 739–750, 2013.

[32] J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[33] U. Demirbaga and D. N. Jha, “Social media data analysis using mapreduce programming model and training a tweet classifier using apache mahout,” in *2018 IEEE 8th international symposium on cloud and service computing (SC2)*. IEEE, 2018, pp. 116–121.

[34] U. Demirbaga, A. Noor, Z. Wen, P. James, K. Mitra, and R. Ranjan, “Smartmonit: real-time big data monitoring system,” in *2019 38th symposium on reliable distributed systems (SRDS)*. IEEE, 2019, pp. 357–3572.

[35] Y. Yuehong, Y. Zeng, X. Chen, and Y. Fan, “The internet of things in healthcare: An overview,” *Journal of Industrial Information Integration*, vol. 1, pp. 3–13, 2016.

[36] C. A. Tokognon, B. Gao, G. Y. Tian, and Y. Yan, “Structural health monitoring framework based on internet of things: A survey,” *IEEE Internet of Things Journal*, vol. 4, no. 3, pp. 619–635, 2017.

[37] S. Oueida, Y. Kotb, M. Aloqaily, Y. Jararweh, and T. Baker, “An edge computing based smart healthcare framework for resource management,” *Sensors*, vol. 18, no. 12, p. 4307, 2018.

[38] A. M. Rahmani, T. N. Gia, B. Negash, A. Anzanpour, I. Azimi, M. Jiang, and P. Liljeberg, “Exploiting smart e-health gateways at the edge of healthcare internet-of-things: A fog computing approach,” *Future Generation Computer Systems*, vol. 78, pp. 641–658, 2018.

[39] A. Jindal, G. S. Aujla, and N. Kumar, “Survivor: A blockchain based edge-as-a-service framework for secure energy trading in sdn-enabled vehicle-to-grid environment,” *Computer Networks*, vol. 153, pp. 36–48, 2019.

[40] K. A. Clauson, E. A. Breeden, C. Davidson, and T. K. Mackey, “Leveraging blockchain technology to enhance supply chain management in healthcare:: An exploration of challenges and opportunities in the health supply chain,” *Blockchain in healthcare today*, 2018.

[41] D. C. Nguyen, P. N. Pathirana, M. Ding, and A. Seneviratne, “A cooperative architecture of data offloading and sharing for smart healthcare with blockchain,” in *2021 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*. IEEE, 2021, pp. 1–8.

[42] Y. Yu, S. Liu, P. L. Yeoh, B. Vucetic, and Y. Li, “Layerchain: a hierarchical edge-cloud blockchain for large-scale low-delay industrial internet

of things applications," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 5077–5086, 2020.

[43] M. Hölbl, M. Kompara, A. Kamišalić, and L. Nemec Zlatolas, "A systematic review of the use of blockchain in healthcare," *Symmetry*, vol. 10, no. 10, p. 470, 2018.

[44] S. Khezr, M. Moniruzzaman, A. Yassine, and R. Benlamri, "Blockchain technology in healthcare: A comprehensive review and directions for future research," *Applied sciences*, vol. 9, no. 9, p. 1736, 2019.

[45] T. McGhin, K.-K. R. Choo, C. Z. Liu, and D. He, "Blockchain in healthcare applications: Research challenges and opportunities," *Journal of Network and Computer Applications*, vol. 135, pp. 62–75, 2019.

[46] R. A. Michelin, A. Dorri, M. Steger, R. C. Lunardi, S. S. Kanhere, R. Jurdak, and A. F. Zorzo, "Speedychain: A framework for decoupling data from blockchain for smart cities," in *Proceedings of the 15th EAI international conference on mobile and ubiquitous systems: Computing, networking and services*, 2018, pp. 145–154.

[47] K. Lei, M. Du, J. Huang, and T. Jin, "Groupchain: Towards a scalable public blockchain in fog computing of iot services computing," *IEEE Transactions on Services Computing*, vol. 13, no. 2, pp. 252–262, 2020.

[48] S. Shukla, S. Thakur, S. Hussain, J. G. Breslin, and S. M. Jameel, "Identification and authentication in healthcare internet-of-things using integrated fog computing based blockchain model," *Internet of Things*, vol. 15, p. 100422, 2021.

[49] A. Dubovitskaya, F. Baig, Z. Xu, R. Shukla, P. S. Zambani, A. Swaminathan, M. M. Jahangir, K. Chowdhry, R. Lachhani, N. Idnani *et al.*, "Action-ehr: patient-centric blockchain-based electronic health record data management for cancer care," *Journal of medical Internet research*, vol. 22, no. 8, p. e13598, 2020.

[50] B. Sharma, R. Halder, and J. Singh, "Blockchain-based interoperable healthcare using zero-knowledge proofs and proxy re-encryption," in *2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)*. IEEE, 2020, pp. 1–6.

Umit Demirbaga is a Postdoctoral Research Associate in Health Data Science in the Department of Medicine at the University of Cambridge, UK. He is also a Visiting Postdoctoral Fellow at the European Bioinformatics Institute (EMBL-EBI), UK. He received MSc and PhD degrees in Computer Science from Newcastle University, UK, in 2017 and 2021, respectively. His research interests are mainly in the areas of big data analytics, cloud computing, distributed systems, and healthcare data analytics. He was awarded Outstanding Performance Award with Best Team Project Award in his MSc in 2017.

Gagangeet Singh Aujla is an assistant professor of computer science at Durham University, United Kingdom. Prior to this, he was a postdoctoral research associate with the School of Computing, Newcastle University, United Kingdom. He received his Ph.D. in computer science from Thapar University, India, in 2018. He received the 2018 IEEE TCSC Outstanding Ph.D. Dissertation Award and 2021 IEEE Systems Journal Best Paper Award, which recognized his leading expertise in the application of scalable and sustainable algorithms for cloud data centers, SDN, and smart grid. He is an Area Editor of Ad Hoc Networks (Elsevier) and Associate Editor of IET Smart Grid.