Big Data Analytics Framework for Childhood Infectious Disease Surveillance and Response System using Modified MapReduce Algorithm

A Case Study of Tanzania

Mr. Mdoe Mwamnyange[1], Dr. Edith Luhanga[2], Mr. Sanket R. Thodge[3]
School of Computation and Communication Science and Engineering (CoCSE)[1, 2]
The Nelson Mandela African Institution of Science and Technology (NM-AIST) Arusha, Tanzania[1, 2]
Pi R Square Digital Solutions Pvt Ltd, S B Road, Gokhale Nagar, Pune, India[3]

***Abstract*—Tanzania, like most East African countries, faces a great burden from the spread of preventable infectious childhood diseases. Diarrhea, acute respiratory infections (ARI), pneumonia, malnutrition, hepatitis, and measles are responsible for the majority of deaths amongst children aged 0-5 years. Infectious disease surveillance and response is the foundation of public healthcare practices, and it is increasingly being undertaken using information technology. Tanzania however, due to challenges in information technology infrastructure and public health resources, still relies on paper-based disease surveillance. Thus, only traditional clinical patient data is used. Nontraditional and pre-diagnostic infectious disease report case data are excluded. In this paper, the development of the Big Data Analytics Framework for Childhood Infectious Disease Surveillance and Response System is presented. The framework was designed to guide healthcare professionals to track, monitor, and analyze infectious disease report cases from sources such as social media for prevention and control of infectious diseases affecting children. The proposed framework was validated through use-cases scenario and performance-based comparison.**

Keywords—Big data analytics; childhood infectious diseases; infectious disease surveillance system; infectious disease report cases; framework; Hadoop; healthcare big data; map reduce

I. Introduction

In 2018, there were 5.3 million deaths of children under the age of 5 years around the world (World Health Organization report, 2020, Oct. 15) with most of the deaths taking place in the African region. Data from UNICEF, WHO, World Bank, and UN-DESA Population Division (2019, Oct. 18), show that almost half of these deaths (18 of every 1,000 births globally) occurs within the first 28 days of life. The probability of death lowers to 11 per 1,000 births between the ages of 1 month to 1 year and 10 per 1,000 births between the ages of 1 and 5. The main causes of mortality are malnutrition and preventable, childhood diseases including diarrhea, pneumonia, and malaria. Malaria, pneumonia, and diarrhea are the main causes of the majority of child deaths in Tanzania [1].

The United Nations Sustainable Development Goal no. 3 has set a childhood mortality rate of 25 deaths per 1,000 births as the target. A wide range of solutions to achieve this goal have been implemented across Sub-Saharan Africa (SSA). The solutions include: providing affordable maternal health care services and improved access to drinking water and sanitation services [2], providing broad-spectrum antibiotics, such as azithromycin to children [3], and effective coverage of available cost-effective interventions and technologies [4]. A strong functional infectious disease surveillance system has also been highlighted as a significant tool to provide healthcare information to support public health decision-making worldwide (WHO 2000). According to WHO, a disease surveillance system “is the continuous, systematic collection, analysis and interpretation of healthcare-related data needed for the planning, implementation, and evaluation of public health practices”. It serves as an early warning and alert system for unforeseen public health emergencies, supports medical practitioners to prepare infectious disease spread report cases for intervention, track disease spread advancement, and provides crucial information to the epidemiologist, policy, and decision-makers.

In 1998, the United Republic of Tanzania adopted an Integrated Disease Surveillance and Response Program (IDSR) for all disease surveillance activities in the country [5]. Thirty-four (34) notifiable diseases were included in the guidelines, as well as procedures for disease case detail, reporting, and actions to be taken for public health service levels to achieve timely infectious disease detection, investigation, and response to disease outbreaks. The IDSR has been facilitated with the nationwide implementation and functioning of the Health Management Information System (HMIS) and District Health Information System (DHIS2) systems. However, there was still weakness in data coverage and completeness, as not all infectious diseases were accommodated for surveillance, healthcare data from the laboratory were not linked with the IDSR system, and data on cases shared via the internet, web-based system, and short text message system (SMS) were not accommodated, even though no-care or home-care was sometimes practiced [6].

Due to the challenges of emerging and reemerging infectious diseases in the world [7], it has become very difficult for the existing IDSR system to detect and analyze small to medium size outbreaks of infectious diseases. As a result, these outbreaks remained hidden and distributed unnoticed over a wide local geographic area because of the growing local food processing technologies in the country. The emergence of the novel coronavirus (Covid-19) in 2020 in the world and the continued reemergence of cholera, dengue, and other infectious diseases in Tanzania necessitate that such measures of engaging additional sources of healthcare data in surveillance, and the use of big data analytics processes to identify patterns, trends, and other valuable new directions on infectious disease surveillance systems are implemented. Though mining the web-based healthcare data through websites, social networks, short text messages and other online news archives using big data analytics technology provide a valuable new direction on infectious disease surveillance system [8][9], but the new technology is still unknown what are the requirements and best frameworks to be adopted for Tanzanian context. Despite the efforts to include non-clinical data in surveillance, there is no framework available to guide public health officials on which sources of data should be included, what analytics processes should be used and how the generated insights can be validated and used for decision making.

Using big data analytics technology in healthcare such as text analytics [10], data streams analytics [11], social network analytics [12] [13], machine learning techniques [14], natural language processing, data mining, and predictive analytics in healthcare [15] can effectively help to analyze diverse-mix of healthcare data from laboratory, diagnoses and medications, drug-resistance patterns, drug interactions and dosing patterns, fraud detection [16] and early warning of disease outbreaks which can boost healthcare system in Tanzania.

The purpose of this study was to identify how healthcare big data can be collected, integrated, and transformed into useful information for preoccupation healthcare planning and implementation in Tanzania, particularly for childhood infectious diseases. The requirements gathering for the big data analytics framework for childhood infectious disease surveillance and response system and validation were conducted via surveys and use-case scenarios respectively, with health facilities' IT personnel, pediatricians, and nurses from four referral hospitals in the country. We modified and existing big data for infectious disease surveillance framework to fit the identified needs and during validation, we found that the framework supported wide coverage of healthcare data collection from online data sets. It allowed transmission and processing of large-scale structured and unstructured healthcare data sets with minimal processing time, provided flexibility to write user-defined programs or run queries on top of the native program to produce the expected results and was overall well accepted by the users.

The rest of this paper is organized into the following sections: Section 2 background, Section 3 related work, Section 4 research methodology, Section 5 results, Section 6 proposed framework, Section 7 use-case diagram, Section 8 proposed system design architecture, Section 9 data flow diagram, Section 10 framework validation experiment, Section 11 validation results, Section 12 discussion and finally Section 13 conclusion.

II. Background

A. Health Information and Healthcare-seeking behavior

Despite advancements in medical and increased vaccine availability to the children, emerging and re-emerging of infectious diseases continues to pose threats to parents, children, and the community at large, based on reported cases of pneumonia, malnutrition, hepatitis, malaria, and other infectious diseases. In 2019 an estimated 5.2 million children under 5 years died mostly from preventable and treatable causes in the world. Children aged 1 to 11 months accounted for 1.5 million of these deaths while children aged 1 to 4 years accounted for 1.3 million deaths of which the leading causes include pneumonia, diarrhea, and malaria. Sub-Saharan Africa remains the region with the highest under-5 mortality rate in the world with more than 80% of the 5.2 million under-five deaths in 2019 (WHO, 2019). This is the average of 1 child in 13 dying before his or her fifth birthday. In Tanzania, according to Tanzania Demographic and Health Survey (TDHS, 2010) reported that 1 out of 20 children die before their first birthday, and 1 out of 12 die before their fifth birthday. These challenges have led to the need for new approaches and technologies for infectious disease alerts, detection, and immediate response.

Also, the use of computer and internet access has been growing at a very high speed in Tanzanian with estimates of 25.7 million internet users in a population of more than 58 million till 2019 (TCRA quarterly communications statistics report, December 2019). 62% have access to mobile phones, and they use them to share information (CIA world Factbook 2019). The 2010 Tanzania Communications Regulatory Authority (TCRA) report [17] has also proved that the number of internet users has been growing at an average rate of 24% per annum from 2005 to 2010. This means that there is also an increased frequency of using internet-based technologies to acquire health and disease information among parents and members of the community. Members of the community have experienced plenty of useful healthcare information available from the mass media. The sources of information involve radio, television, mobile phones, social media, and health websites in which health free-text data and audio of disease causes, diagnosis, prevention, and control are needed.

Socio-demographic characters, possession of health insurance, exposure to mass media (internet, radio, television) have changed the traditional healthcare-seeking behavior characters of relying on health facilities in Tanzania [18]. Exposure to mass media was found to be statistically and significantly associated with appropriate healthcare-seeking behavior change. The influence of seeking appropriate healthcare information such as disease causes, diagnosis, treatment, prevention, and control has changed the healthcare-seeking behavior among parents and members of the community to seek appropriate health information anywhere from the mass media. Some research studies have shown that using mass media (including radio, television, mobile phones, social media, and health websites) has a positive impact on health facility deliveries [19].

B. Big Data Analytics in Healthcare

Big data in the healthcare system is the collection and processing of a multi-diverse mix of healthcare data sets such as structured, semi-structured, and unstructured data which make complicated data mining processes in the traditional system [20]. In big data analytics phenomena in healthcare, data comes from various data sources such as local information news, structured and unstructured data, laboratory test information, radiology images, healthcare website reports, click streaming, Twitter feeds, e-mails, call detail reports, video camera, social network data, weblog files, smartphones mobile apps, audio, healthcare equipment sensors, and others. The data if properly analyzed can greatly aid in evidence-based decision-making on childhood infectious disease management and decision-making. These types of data cannot be processed and stored in a traditional database as they belong to different formats of data sets. The data typically cannot be analyzed with traditional Structured Query Language (SQL) tools such as HIMS. Instead, new non-relational database technology such as Hadoop, MapReduce algorithm, and NoSQL database tools are needed [21]. The high-speed performance, multiprocessing, concurrency, per server throughput, and parallelism processing clusters technologies are the essential requirements on highly scalable healthcare big data analytics [22].

The advantage of using big data analytics technology on infectious disease detection and control is to use e-mail and online free-text health data from the internet to disseminate information of infectious disease outbreaks by e-mailing and posting infectious disease case reports. It can help to conduct disease mapping surveillance by continuously gather and display public health data about new infectious disease outbreak using internet-based data sources such as online news, websites, RSS feeds, expert opinion, and official alerts based on geographical location, time, and disease agent which cannot be supported with the traditional system.

III. Related Work

Many research studies to supplement existing traditional systems and design new models to detect infectious diseases using big data analytics such as social network and internet search queries to gather and process data at a speed that is close to real-time have been conducted in many countries. The following are some related works extracted from the literature review studied on this research:

A. Big Data Analytics using Online Information Aggregates Search Engines

Google Flu Trends Healthcare Big Data Analytics; this service was conducted by Google to predict and locates flu infectious disease outbreaks by making use of online information aggregates search queries.

The San Francisco-Based Global Viral Forecasting Initiative (GVFI) has been used advanced big data analytics on information mined from the internet to identify locations, sources, and drivers of local infectious disease outbreaks before they become global epidemic [23].

Ginsburg et al. [24], also developed a method to collect and analyze healthcare big data through search queries from Google (http://www.google.com/) to track Influenza-Like Illness (ILI) within a given population. They conducted their research on Google search queries taken from historical logs during 5 years (2003 up to 2008) using 50million of the most popular searches.

B. Big Data Analytics using Social Networks

A. Signorini, A. M. Segre, and P. M. Polgreen [25] employed big data analytics technology on social media using Twitter post data across the United States by searching through particular areas and analyzing the data to predicate weekly Influenza-Like Illness (ILI) levels. The focus of their efforts was on the period when the H1N1 epidemic was happening in the United States. The overall aim was to examine the use of information about news and geopolitical events embedded in Twitter to track rapidly-evolving public sentiment concerning H1N1 and measure actual disease activity to monitor the seasonal influenza-related traffic within the United States.

They gathered data set consisted of 4,199,166 tweets selected from the roughly 8 million influenza-related tweets (i.e., keywords h1n1, swine, flu, or influenza) observed between October 1, 2009, up to May 20, 2010, using Twitter’s streaming application programmer’s interface (API). The tweets were sifted through looking for posts containing a preset of keywords correlated to H1N1 (h1n1, flu, swine, influenza). They trained 32 times on each 31-week subset of the training data. The estimates of the prediction model for national ILI values produced by the system were fairly accurate, with an average error of 0.28% and a standard deviation of 0.23%.

H. Achrekar, A. Gandhe, R. Lazarus, S. Yu, and B. Liu [26]. Used big data analytics technology by developed a system deemed Social Network Enabled Flu Trends (SNEFT) which continuously monitored tweets to detect and track the spread of ILI epidemics. The study used a data set of tweets and profile details of the Twitter users who commented on flu keywords started on October 18, 2009. They used an OSN crawler that searched online social networks they developed to retrieve tweets from the internet using keywords flu, H1N1, and swine flu.

Yuan et al. [27] also developed a system to collect and analyze the big data of healthcare using search query data. The study used search queries gathered from Baidu (baidu.com) to track ILI epidemics across China. The author gathered their data from Baidu’s database (http://index.baidu.com/) which stored the online search query since June 2006. For this study, they only gathered data from March 2009 to August 2012, which was during the Influenza virus (H1N1) epidemic, and compare their results to that of China’s Ministry of Health (MOH). Yuan et al.’s system was split into four main parts: (a) choosing keywords, (b) filtering these keywords, (c) defining weights and composite search index, and (d) fitting the regression model with the keyword index to that of the influenza case data.

Ashish Naveen, B. Antarip, D. Sumit, N. Saurav, and P. Rajiv[28], created an online big data analytics platform called Abzooba Smart Health Informatics Program (SHIP). Their purpose was to help patients connect to the medical experiences of other patients posted throughout the internet via online discussion message boards. They used a pool of 50,000 discussion messages including posts extracted from websites such as inspire.com, medhelp.com, and others to extract and execute big data text processing to extract information of each entry including posts and replies which have medical significance related to health such as treatments, side effects, medicines, etc.

C. Big Data Analytics for Healthcare in Africa

Although the application of big data analytics in healthcare is still in its infancy stages in Africa compared to the developed nations, some evidence proved that big data analytics is emerging in Africa particularly in Sub-Saharan Africa, and has shown the potential to improve the public health system. The emergency use of the internet, web-based systems, social networks (Baidu, Instagram, WhatsApp, Twitter, and Facebook, etc.), and other mobile devices in Africa is making a foundation source of big data which can help to improve infectious disease surveillance. Through the preliminary evidence of an emerging technology few research studies have been made by researchers to practically demonstrate the usefulness of using big data analytics in public health for the African continent using mobile phones and social networks.

The first example of the use of big data analytics for infectious disease management in Africa was the use of mobile phones in connection with the HealthMap online system in 2014 for detecting the Ebola virus epidemic in Guinea, Liberia, Nigeria, and Sierra Leone in western Africa. The system used emails, RSS feeds, text, and online free-text data on its surveillance.

Wesolowski et al., 2012 [29] used mobile phones to monitor the movement of malaria parasites by analyzing call and text data of mobile phone subscribers of about 15 million people of Kenya. They estimated 14,816,521 Kenyan mobile phone subscribers between June 2008 and June 2009, through mapping every call and text made by each individual to one of 11,920 cell towers located within the boundaries of 692 settlements. The aim was to identify the dynamics of human carriers that drive parasite importation between regions and mapping the routes of parasite dispersal by human carriers in Kenya. The result of this analysis was compared with the hospital records to detect malaria transmission in the local geographical areas in Kenya. This research study assisted the Kenyan government to develop an effective malaria control program. The strength of this study was the ability of the system to pool huge amounts of data from the mobile phone subscribers to track the movement of people. However, the limitation of this system was also the inability to combine structured and unstructured data for sophisticated healthcare data analysis.

The use of internet-based and mobile phone technology systems for disease surveillance in the world has quickly become controlling sources of information on emerging infectious disease surveillance, however, their impacts on public health dynamics remain undetermined. Lack of authenticity, false reports, and information overload restrict the cognizance of their potential for public health practices. In Tanzania, the appropriate usage of big data analytics technologies for infectious disease surveillance is still unknown. The issues on how nontraditional healthcare data can be incorporated with the traditional health data, what are the sources and limitations of the available unstructured infectious disease report cases data, how the online healthcare data can be extracted and transformed into useful information, unaffordable high-performance computing infrastructures for big data analytics are still the challenges. This work aims to address this gap, with a focus on the development of a childhood infectious disease surveillance system.

IV. Research Methodology

A. Study Area

This study was conducted in four regions in Tanzania, Dar es Salaam, Arusha, Kilimanjaro and Mbeya. Six regional referral hospitals were included, namely Amana, Temeke, and Mwananyamala referral hospitals in Dar es Salaam, Mount Meru hospital in Arusha, Mawenzi hospital in Kilimanjaro, and Mbeya referral hospital in Mbeya. The hospitals were purposively selected since they are responsible for coordinating surveillance activities and mobilizing resources and providing technical support for the surveillance activities conducted at lower levels.

B. Participants

A total of 110 participants took part. Forty-nine (49.09%) were pediatricians, thirty-two (32.73 %) were medical records officers and eighteen (18.18%) were IT healthcare professionals in the hospitals. The heads of departments from the hospitals were asked to propose people who met the following criteria: (i) able to read and write in English (since the questionnaire was in English) (ii) have 3+ years' experience particularly in infectious disease data collection and analysis. The heads were also asked to ensure gender balancing in their list of proposed participants.

C. Data Collection

Data collection was conducted between February and May 2019. Surveys, interviews, and observation were used to:

1) Identify challenges facing healthcare professionals from infectious diseases prevention and control perspective.

2) Identify healthcare information gaining mechanisms and decision-making information gaps and.

3) Identify healthcare system opportunities for future infectious disease prevention and detection.

The questionnaires developed asked questions on: major challenges on infectious diseases prevention and control; Infectious disease data collection and analysis; Involvement of citizens/ public to collect and analyze infectious diseases; Organization experience in healthcare big data and data-driven innovation; and the use of healthcare big data technology in the healthcare system, among others. There were a total of 31 questions, with 28 requiring responses on a 5-point Agreement

Likert scale (1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree and 5 = strongly agree).

The interview was used to gain further insights into the infectious disease surveillance process and its successes, challenges, and the solutions they used to overcome the challenges particularly on collecting and analyzing unstructured data. The data were also recorded and analyzed using the descriptive statistics method.

D. *Data Analysis*

A total of 108 questionnaire sheets that qualified for data analysis were returned and analyzed. Descriptive statistics were used to analyze the survey responses. Free-text responses were inductively coded and the frequencies of each theme were calculated.

The analysis of the questionnaire was performed based on the indicated themes. The major challenges that the healthcare professionals experience in their day to day performance were measured by seven questions (inadequate of the infrastructure /facilities; access and quality of the healthcare centers' services; difficulties to reach remote areas; inability to collect infectious disease data from the patient's environments; poor quality of food, water, and housing services; people's culture; and shortage of the number of healthcare staff).

The theme of data collection and involvement of citizens was also measured to assess if the traditional system can integrate healthcare data from other healthcare-related systems such as healthcare websites, social media, mobile phones, and public pharmacies to improve cross-functional communication and collaboration among healthcare systems.

V. Results

The results in this study show that in the traditional system, three of the six hospitals (Mwananyamala, Mawenzi, and Mbeya) only clinical data were used for surveillance. The other three hospitals (Ilala, Temeke, and Mount Meru) relied on clinical data and other healthcare data sources including community case findings as well.

The results of the questions v/s responses in histogram chart:

However, the following results were obtained based on the analysis of the individual cases:

A. *Sources of Non-clinical Data for Surveillance*

The results in questions 1-7 confirmed that the healthcare professionals have greater challenges collecting data of infectious disease report cases from the patients' environments as indicated in question 5. Over 70% (n=78) of the responses strongly agreed that inability to collect data from the patient's environment and involvement of citizens hampered surveillance and response as indicated on the histogram chart in Fig. 1. Data collection from other sources including free-text from mobile phones, social networks like WhatsApp, Facebook, Twitter, and email system would improve infectious disease surveillance systems.

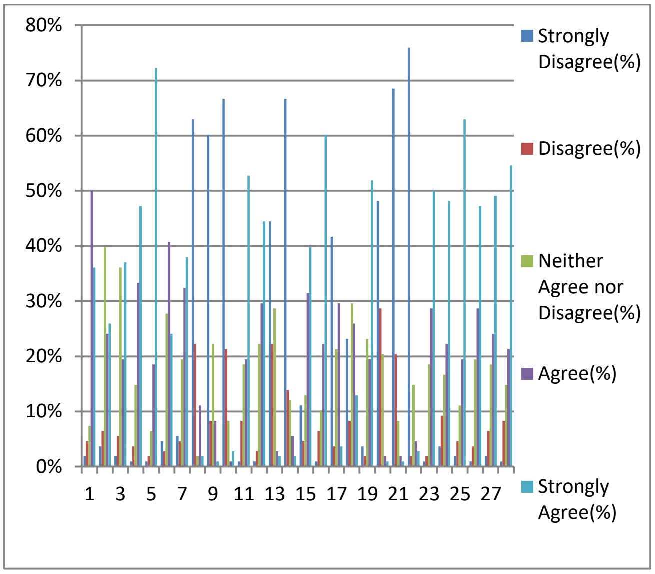


Fig. 1. Histogram Chart of the Questions v/s Response Percentages from the Respondents.

B. *Experience in Healthcare Big Data and Data-driven Innovation*

The results confirmed that 70(65%), out of 108 respondents strongly agreed that they do not have experience in healthcare big data and data-driven innovation. Also, they agreed that they do not have a healthcare big data framework for collecting, analyzing, and transforming very large infectious disease report cases data sets as indicated in questions 17-21 in Fig. 1. In this group of respondents 33(30%) were pediatricians, 21(20%) medical records and 16(15%) were IT professionals.

C. *Current and Proposed Health Data Sources*

Over 75% (n=81) of the 108 respondents strongly agreed that the collection and analysis of infectious disease report cases from other sources using multiple channels such as mobile phones, websites, e-mails, social media, and content management systems would improve the traditional surveillance system as indicated in question 22 in Fig. 1. 86(80%) of 108 respondents chosen mobile phone short-text messages as a good source of information. 54(50%) of 108 respondents selected web-based free-text information as the source of health data and 97(90%) of 108 respondents opted for social media free-text application as the good source of health data.

VI. Proposed Big Data Analytics Framework

The big data analytics framework for the healthcare system developed by [30] was considered as a reference for the development of this framework. This dynamic framework has been considered as a healthcare big data framework base for the general healthcare system which can support tracking and monitoring infectious disease. It has incorporated the general known healthcare big data analytics approach based on the fundamental variables. This is inefficient to be adopted for the process of collection and analyzing the healthcare big data in Tanzania environment.

In this study, the author developed a framework that serves as a reference model to make healthcare professionals in Tanzania ready to explore and implement big data analytics technology in the healthcare system. The data collected for this study could only relate to the short time validation of this framework. The findings are limited based on the local scale of the use-cases scenarios phenomenon. Further research and update are needed throughout the framework studies. The resulting big data analytics framework which was proposed to be suitable for the Tanzania context was shown in Fig. 2.

The proposed framework was divided into the following layers: (a) data capture layer (b) data acquisition layer (c) data analytics layer and (d) information exploration.

A. Data Capture Layer

The data capture layer involves all traditional and non-traditional data sources necessary to provide insights on early infectious disease prevention and control. It involves structured data from HIMS/DHIS-2, Health Insurance, medical healthcare sensors system, online structured healthcare information archives, home patient monitoring sensors, and public pharmacy. These clinical data can be collected from various sources through tables, csv files, json data files, and text file format and stored in the relational databases depending on the content format such as MySql, Oracle, PostgreSQL, and others. Since unstructured data cannot be processed using structured databases [31], then data will be stored in the nontraditional databases which can handle unstructured data such as MongoDB Databases.

B. Data Acquisition Layer

The data acquisition layer is responsible for handling data that comes from various healthcare data sources. In this layer, healthcare data stored in the various structured databases such as tables and csv files and unstructured databases such as free-text, json data, video, and audio format can be transformed into Hadoop Distributed File System (HDFS) format ready to be processed in big data analytics tools. Since the incoming data comes from various data sources, their characteristics are varying in terms of a communication channel, frequency, size, volume, and file format. Therefore, the transformation engine must be able to extract, merge and transform data into key-value pairs.

In this layer, the transformation engine must be able to support functions such as data transfer, cleaning, splitting, sorting, merging, and validating data. For instance, structured healthcare data sets records such as (patient name, age, address, location, and disease descriptions or medical history) can be extracted and transformed into key-value pairs of the HDFS format. This process can also be done in unstructured data whereby data in the format of e-mail, weblogs, or text can be extracted and transformed into key-value pairs as well.

C. Data Analytics Layer

In this layer, data can be processed and analyzed in three ways: Hadoop MapReduce data processing, data streaming, and in-database analytics.

MapReduce data processing works by breaking data processing into two phases: Map phase and Reduce phase. Each phase contains key-value pairs as input and output. The input to our Map phase is the raw or unstructured data, which is processed by split up into key-value pairs. And the output from the *Map function* is processed by the MapReduce framework before being sent to the *Reduce function*. MapReduce processing in big data analytics provides the ability to process a large volume of structured and unstructured healthcare data in batch processing and massively parallel processing [32].

Data streaming processing can help to process and analyze real-time and near real-time stream data processing. In real-time stream data processing, healthcare professionals can track healthcare data-in-motion such as rates of infectious disease spread, prediction of infectious disease outbreaks, respond to unexpected infectious disease outbreak and quickly make an evidence-based decision for early infectious disease notification, prevention, and control.

In this layer, the healthcare data analytics such as healthcare revenue cycle, healthcare supply chain, disease management, disease case-specific management, healthcare cost and quality management, and operational efficiency can be analyzed. The proposed framework can help to conduct these analyses using various big data analytics techniques. For example: in the healthcare revenue cycle, we can use the framework to conduct analytics through collect, compile and analyze health data from various healthcare centers reported on a monthly, quarterly and annual basis to enable healthcare professionals and the community to gain insight into the national average and top quartile of the revenue cycle. This can be done in a framework through developing a modified MapReduce algorithm that can be able to map and reduce health data accordingly based on the modified principles of the MapReduce algorithm.

D. Information Exploration Layer

The information exploration layer provides output results based on visualization reports, real-time information monitoring, and other useful healthcare business insights reports. Some information and visualization reports which were felt important were healthcare revenue cycle, healthcare supply chain, healthcare information technology, healthcare cost, and quality and operational efficiency. Because it will help to provide healthcare executives and leadership teams with objective, actionable best practice research and implementation resources. It will help healthcare professionals to make early evidence decisions such as infectious disease alerts, warnings, and notifications to the citizens before an infectious disease outbreak.

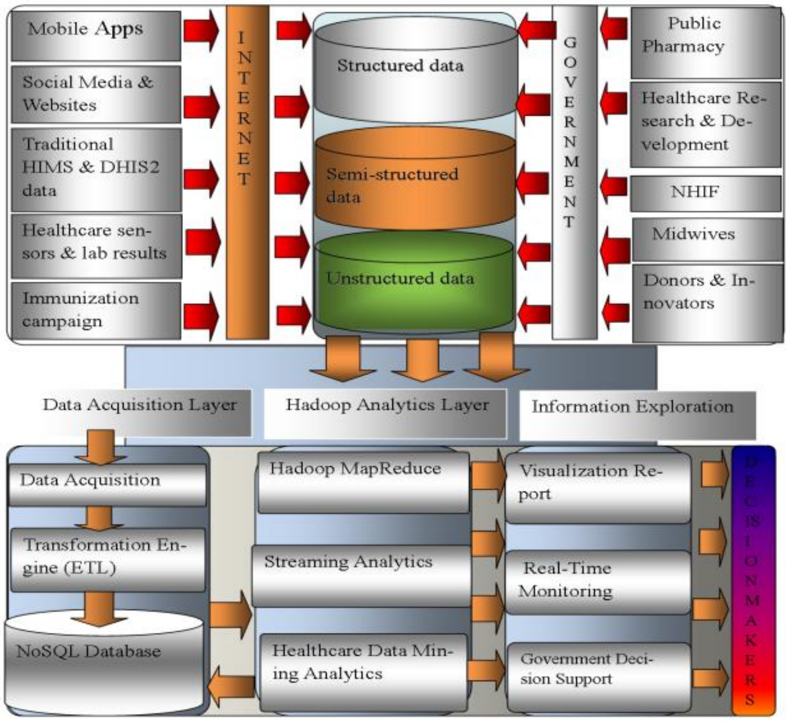


Fig. 2. Proposed Big Data Analytics Framework.

VII. Framework Model use Case Diagram

The framework model use case diagram in Fig. 3 is a simple representation of user interaction which provides a simplified real picture to the stakeholders of how the framework can be implemented in the real world as presented in the proposed system design architecture in Fig. 4. Based on the use case diagram, we can divide the diagram into four areas: (a) Healthcare big data sources (b) Data ingestion zone (c) Big data analytics zone, and (d) Big Data Application zone.

A. Healthcare Big Data Sources

Healthcare big data sources involve roles of data collection from the various healthcare data sources. It involves the collection of infectious disease data from patients using various tools including mobile-apps, web-based systems, social media, content management systems, clickstreams, weblogs, and online archives. The traditional system already has dynamic categories of healthcare provider users who collect infectious disease data using HIMS/DHIS-2, Infectious Disease Week Ending System (IDWE), and the community healthcare activists who collect and submit data to the hospitals. These categories of users will be improved by assigned activities of collecting infectious disease data using digitized data through mobile applications and web-based systems instead of the existing manual paper-based system. Initially, doctors, IT personnel, laboratory scientists and medical records personnel can help to collect other data of infectious disease from pharmacies, social media, clickstreams, weblogs, and online archives through healthcare websites, mobile applications, and online healthcare systems as indicated on the data flow diagram in Fig. 5.

B. Data Ingestion Zone

Data ingestion zone involves data integration and streaming processes. IT personnel, doctors, medical records personnel, and laboratory scientists can help to conduct this process. It involves running queries and commands to extract, transform and load infectious disease data from the data sources into the Hadoop platform. Structured databases such as tables, csv files, and json data from local pharmacies, healthcare insurances, and others can be collected and integrated at this stage. The unstructured data will also be extracted and transformed before transmitted into the Hadoop Big Data analytics engine as indicated in Fig. 5.

C. Big Data Analytics Zone

The big data analytics zone involves running healthcare data analytics. It involves executing healthcare data jobs using Hadoop MapReduce data processing, Data streaming, and In-Database Analytics. These activities can be done by the doctors, healthcare executive officers, medical officers, and medical records personnel in collaboration with the IT personnel. Structured and unstructured health data will be executed as healthcare data jobs in Hadoop Cluster using the MapReduce algorithm as indicated in Fig. 4. The real-time monitoring streaming will be monitored by the Healthcare specialist and the healthcare big data analytics visualized reports will be submitted to the decision-makers as indicated in Fig. 5.

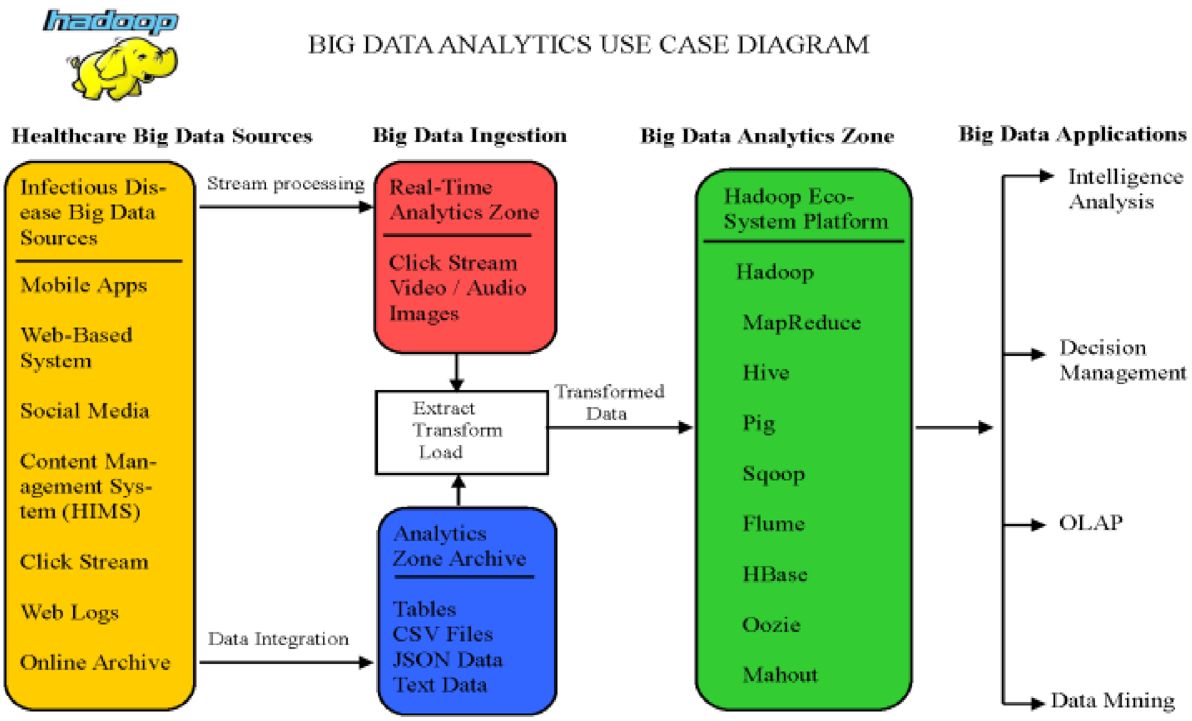


Fig. 3. Use Case Diagram for the Implementation of the Proposed Framework.

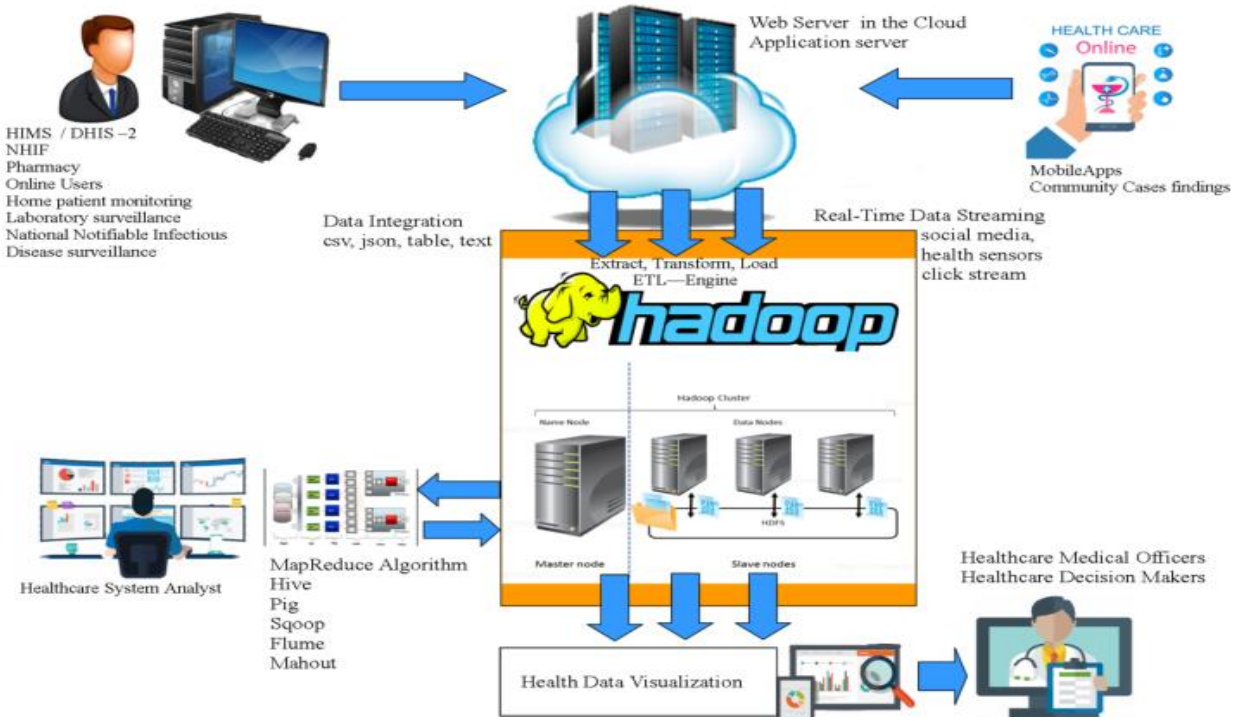


Fig. 4. Proposed Big Data Analytics Framework System Design Architecture.

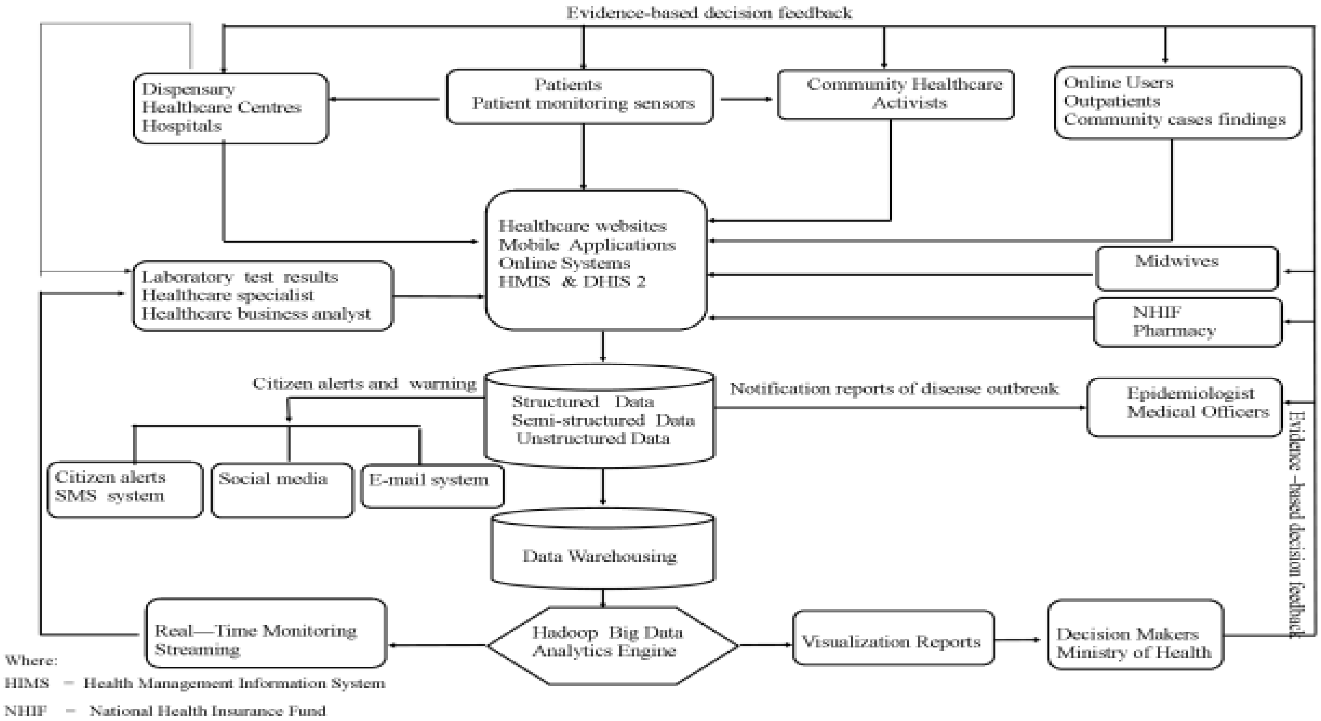


Fig. 5. Data Flow Diagram of the Proposed Big Data Analytics Framework.

D. Big Data Application Zone

The big data application zone involves decision-making processes based on the processed healthcare big data analytics reports. It involves healthcare master data management security and privacy, data standardization, policies, and data incorporation to create immediate, completeness, accurate evidence-based decision-making. It involves executing and managing complex analytics algorithms on data mining and intelligence analysis to manage healthcare business information. These activities can be performed by the higher authority and decision-makers in the healthcare organization

VIII. Framework Validation Experiment

To validate the proposed big data analytics framework for usability and applicability, we conducted the following experiment using the following use-cases scenario:

A. Use Case-Scenario I

Integrate healthcare data sets from various healthcare-related data sources.

During the research study, one of the great challenges facing traditional systems identified was integrating healthcare data sets from other healthcare-related data sources and analyzing them all together for evidence-based decision-making.

Currently, in the traditional system, each organization has its independent system of tracking infectious disease report cases. The hospital conducts disease surveillance using a paper-based system, National Health Insurance Fund (NHIF) also has its way of tracking costs, local pharmacies also have their way of tracking medicines provided to the patients suffering from infectious disease symptoms. These systems cannot communicate with each other even though they dealt with the same common function of prevention and control of infectious diseases.

In this use-case scenario, we modified the MapReduce algorithm programming design model (API) to enable healthcare professionals to integrate healthcare data sets from various healthcare-related data sources. The aim was to use the framework to track patients against the repetition of drug use for the same infectious disease which in turn develop drug resistance adverse effects. This simulation can help healthcare professionals improve drug risk management and cost implication management for evidence-based decision-making.

Because of the confidentiality and integrity of health data sets, we generated healthcare data sets dummy data in json format using the online tool at www.mockaroo.com for the experiment:

NHIF data sets: patient_id, patient_name, drug_description, amount, disease_case_description, and

Hospital data sets: patient_id, first_name, last_name, address, gender, age, and disease_case_diagnosed.

The following modified MapReduce algorithm design model was developed:

I. Modified MapReduce Algorithm Design Model:

Reduce-side joins MapReduce algorithm design model:

```
Procedure: Reduce-Side Joins Multiple Datasets from different files
Input:  Hospital and NHIF datasets
Output: Combination of Hospital and NHIF datasets

Begin:

// Mapper:
// Task I: Read two input files one tuple at a time
: Tokenize each word in a tuple and fetch Patient_ID, Name, Infectious_disease, and Amount

//Task II: Add tags "hosp" to indicate Hospital tuple and "nhif" for NHIF input data to produce Key-Value pairs for Mapper as:
      Key –Value pair [Patient_ID, hosp name]
       Key –Value pair [Patient_ID, nhif name]

//Sorting and Shuffle:

//Task: Aggregate the value to each Key to produce key list as {Patient_ID1 – [(hosp name1), (nhif amount1), (nhif amount2),   (nhif amount3)….]}

//Reducer:
//Task I: Process sorting output to have Patient_ID key and list of Amount from NHIF and Hospital details.

// Task II:  Loop the values to check if they belong to Hospital or NHIF details

//If the value belongs to NHIF;
      1. Show infectious disease trend
      2. Increase counter by 1
      3. Accumulate amount spent, then
      4. Get Total Amount.

// Else,
              Store variable for future assignment;

End Task:
```

B. Use Case Scenario II

Identify the number of notifiable infectious disease report cases at the local geographic areas from 2010 to 2019:

Another great challenge of the traditional system was to identify some historical infectious disease report cases at the local geographical areas (disease report cases counts in weekly, monthly, or yearly for many previous years) to identify trends of the disease before they developed into large massive disease outbreaks for early warning, alerts, quick response, and government intervention.

Currently, this function is done in the traditional system using a paper-based system through counting the number of infectious disease report cases from old documents (mtuha) which is very difficult when it comes to the issue of counting the number of historical infectious disease report cases involves many previous years example cases from 2010 to 2019.

In this validation, we tested the applicability of the framework using a modified MapReduce algorithm to count the number of infectious disease report cases based on local geographical areas for many previous. In this experiment, infectious disease report cases data files (2010-2019) were generated and the following modified MapReduce algorithm design model was developed:

II. Modified MapReduce Algorithm Design Model;

InfectiousDiseaseReportCases WordCount algorithm design model:

```
Procedure: Count Number Of InfectiousDiseaseReportCases WordCount

Input:  InfectiousDiseaseReportCases 2010 – 2019 datasets

Output: Number Of Counts of InfectiousDiseaseReportCases for Each Local Geographical Area

Begin:

// Mapper:

// Task I: Read ten input files one tuple at a time

      : Tokenize each word in a tuple and fetch Area_Name and Disease_Name words that matching

//Task II:  Splits the line into tokens separated by whitespaces and emits Key-Value pair as

Key – Value pair [Area_ID, Area _name

Key – Value pair [Disease_name, DiseaseReportCase]

//Sorting and Shuffle:
//Task: Aggregate the value to each Key to produce key list as {Area_ID1 – [(Area_name1, Disease_name1),DiseaseReportCase 1), (Area_name2, Disease_name2), DiseaseReportCase 2), (Area_name3, Disease_name3), DiseaseReportCase 3),………..]}

//Reducer:
//Task I: Process sorting output to have Area_ID, Area_Name, Disease_name key, and list of DiseaseReportCases from each Area_Name

//Task II:  Loop the values to check the Frequency for each Area_ID, Area_Name, Disease_Name, key to sum up the DiseaseReportCases count

//If there is more value of DiseaseCasesReport in one Area;

1. Count number of DiseaseReportCases

2. Increase counter 1

 3. Accumulate the  number of DiseaseReportCases, then

4. Display Area_Name, Disease_Name, and Number of DiseaseReportCases.

// Else,

              Store variable for future assignment;

End Task:
```

C. Use Case Scenario III

Collecting and analyzing Infectious disease data from the online news archives.

Another great challenge of the traditional system was to collect and analyze online infectious disease data sets from online websites, social media, and online healthcare news archives. This function is currently not done in the traditional system which hinders data coverage and completeness on the evidence-based decision-making.

In this validation experiment, we tested the applicability of the framework to collect and analyze online healthcare news archives data sets. 12 text-free document files from the healthcare news archives from the internet were collected from Google scholar and Tanzania Online Daily News using WebCrawler spider developed using Python and Java programming languages. Our goal was to use the framework to conduct healthcare information mined from the internet to identify news articles that contain medical-significance-related information on the key infectious diseases. Our keywords for distributed cache were pneumonia, hepatitis, measles, malnutrition, diarrhea, and acute respiratory infection. In our experimental studies the following modified MapReduce algorithm was set:

III.Modified MapReduce Algorithm Design Model:

```
Distributed Cache MapReduce Algorithm:
Procedure: Distributed Cache MapReduce AlgorithmDesign

Input: 12 – Input files of Text documents with 1 – Keyword file datasets

Output: Number of Keywords matching in each text document

Begin:

// Mapper:

// Task I: Check the existence of both input and output parameters

: Read and write twelve input files one line at a time

: Tokenize each word in a tuple and fetch words that matching with keywords in the Cache file

//Task II: Set String Keywords in a Hash set

: Call Distributed Cache static helper and pass URI-reference in HDFS Cache file

: Set the output Key as LongWritable for the line numbers and Value as Text

: Tokenize each line by spaces, and a wordlist set used to store each distinct word we are interested in searching.

: Check if the line contains in our Keyword list

: If a match is found;

: Emit the line number it was found on as the key and the token itself as the value as Key-Value pair: [Key: Line number, token]

  //Sorting and Shuffle:

//Task: Pull the complete list of cache file URIs in the distributed cache and check the URI array returned

//Task I: Loop the values to check if the URI Array passes the test.

//If the value belongs to URI Array;

            1. Grab the keywords file located in HDFS

            2. Write the keywords in a temporary working directory

            3. Save the contents in a local file named ./keywords.txt

// Else,

    Store variable for future assignment;

    End Task:
```

IX. Framework Validation Results

A. Use-Case Scenario I

In this experiment, the healthcare professionals observed that the proposed big data analytics framework system can integrate structured and unstructured data for multi-processing in large-scale data operations to produce expected results. As indicated in this experiment, the system can integrate structured data format from multiple sources and process them to interpret the results as they wished to solve the problem of integrating and analyzing all together hospital data, health insurance, and pharmacy data as indicated in Fig. 6.

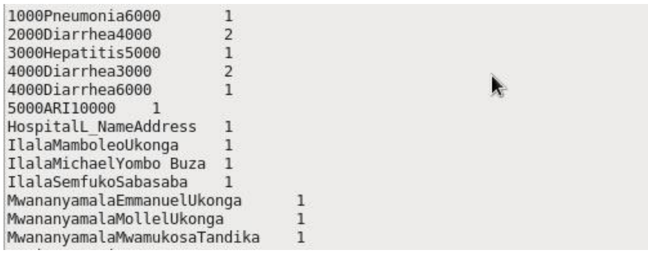
B. Use-Case Scenario II

In this experiment, the healthcare professionals observed that the proposed big data analytics framework system can count the number of infectious disease report cases based on local geographical areas from 2010 to 2019 as indicated in Fig. 7.

This experiment proved that the proposed big data analytics framework system can count the number of infectious disease report cases in a very efficient and fastest method than the traditional system. This has been recommended by a good number of healthcare professional participants from Temeke, Ilala, Mwananyamala, and Mount Meru Referral Hospitals.

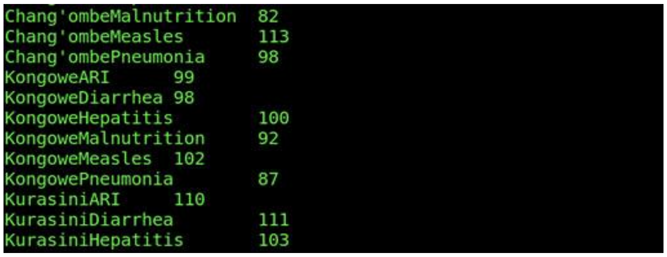
C. Use-Case Scenario III

In this result, out of 12 collected online healthcare news archives, 8 news articles found contain significance related information on various diseases including diarrhea, malnutrition, pneumonia, measles, and hepatitis as indicated in Fig. 8.



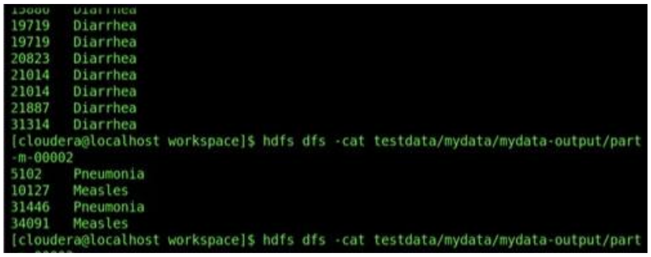
```
1000Pneumonia6000       1
2000Diarrhea4000        2
3000Hepatitis5000       1
4000Diarrhea3000        2
4000Diarrhea6000        1
5000ARI10000    1
HospitalL_NameAddress   1
IlalaMamboleoUkonga     1
IlalaMichaelYombo Buza  1
IlalaSemfukoSabasaba    1
MwananyamalaEmmanuelUkonga      1
MwananyamalaMollelUkonga        1
MwananyamalaMwamukosaTandika    1
```

Fig. 6. Output Results from Part-00000 file in Hadoop Cloudera Express.



```
Chang'ombeMalnutrition  82
Chang'ombeMeasles       113
Chang'ombePneumonia     98
KongoweARI      99
KongoweDiarrhea 98
KongoweHepatitis        100
KongoweMalnutrition     92
KongoweMeasles  102
KongowePneumonia        87
KurasiniARI     110
KurasiniDiarrhea        111
KurasiniHepatitis       103
```

Fig. 7. Analytic Results of the Processed MapReduce Algorithm Program.



```
19719   Diarrhea
19719   Diarrhea
20823   Diarrhea
21014   Diarrhea
21014   Diarrhea
21887   Diarrhea
31314   Diarrhea
[cloudera@localhost workspace]$ hdfs dfs -cat testdata/mydata/mydata-output/part
-m-00002
5102    Pneumonia
10127   Measles
31446   Pneumonia
34091   Measles
[cloudera@localhost workspace]$ hdfs dfs -cat testdata/mydata/mydata-output/part
```

Fig. 8. Number of the Appearance of Keywords of the Online News Archive Files.

This experiment proved that the proposed big data analytics framework system offers more support for healthcare data processing. It offers an opportunity to collect and analyze web-based and internet healthcare data through writing user-defined programs or run queries using other languages such as Python programming language to produce the expected results from the online unstructured healthcare data sets.

X. Discussion

The big data analytics framework model for childhood infectious disease surveillance and response system has been designed for patients, community, healthcare professionals, and decision-makers to meet their specific needs to prevent and control infectious diseases affecting children 0-5 years of age in Tanzania. The framework model has been developed to overcome the following healthcare issues that prevail in the traditional system:

A. Data Collection

The framework model has been developed to accommodate the data collection and analysis process from various healthcare stakeholders including patients through mobile apps and healthcare sensors, community through social media and websites, public pharmacies, laboratory test results, healthcare insurances, and others. The collection of web-based free-text data and mobile phone data will improve the traditional infectious disease surveillance system in Tanzania.

In data collection, the framework model can be widely implemented using a mobile application, short-text messages (sms), online healthcare system, social network, blogging, Internet protocol address, weblogs, and healthcare websites which can be integrated into the same database. This will improve the healthcare data collection process from the citizens including traditional, nontraditional, and pre-diagnostic data from community-level case findings and healthcare centers.

B. Early Detection

Infectious diseases control measures are always done using monitoring tools that help to monitor and limiting infectious disease spread to prevent disease outbreaks by identifying and managing infectious disease report cases through early detection, notification, and warning. Through the proposed framework model, the infectious disease notification alerts including warnings, notification messages, and disease outbreaks notifications can easily be sent to the citizens and healthcare professionals through text messages, e-mail systems, and social network pop-ups for quick action as presented on the data flow diagram in Fig. 5.

C. Healthcare Information Analysis

Having an integrated commodity computers cluster with big data analysis technology makes it easier to perform various types of data analysis in the public health sector. The use of the proposed framework in disease surveillance will help to solve technical and computational challenges that face traditional systems on the ongoing digital data revolution which requires high-performance computation system access to a high volume of stream data and the availability of high-performance computer clusters machine.

D. Evidence-based Decision-making

Infectious disease surveillance and response system conducted in most developing countries are conducted in the condition of resource-limited settings in which often suffers from low reporting coverage, poor data quality, and completeness which in turn provide insufficient data accuracy, poor timely disease outbreak detection and lack of evidence-based decision support. Using the proposed framework model, the evidence-based decision-making process will be more accurate and relevant due to the high quality of healthcare data contents, coverage, and completeness. This will improve collaboration and coordination among healthcare professionals and other stakeholders.

XI. Conclusion

The Big Data Analytics Framework for Childhood Infectious Disease Surveillance and Response System has focused mainly on the performance of the traditional system in Tanzania. Our framework is a simple data-parallel programming model enhanced with sorting, grouping, and reduction capabilities and with the ability to scale to very large volumes of healthcare data. It also works with existing SQL databases and analytics using hive tools. Its distributed implementation requires an underlying distributed file system to access input data, giving preference to local file system access and storing the output. It can be expressed as a data function from input to output framework model.

This approach can be used in similar environments worldwide, but particularly in developing countries, where many of the countries have similar conditions of not paid attention to the infectious disease data quality, coverage and representatives. Whether the infectious disease surveillance endpoint is situational awareness, disease outbreak detection, identifies estimation trends or disease-cost estimation analysis, infectious disease data quality, coverage, and completeness is the key factor during each stage. This approach can play a unique role in developing countries where dispensaries, healthcare centers, hospitals, and primary care settings are performed under limited resource settings while today’s healthcare big data generation and advancement of technology realities demand integrated, relatively low-cost approaches to improve decision-making to comply with the standard of the World Health Organization and International Healthcare regulations.

This study has made the following contributions. First, we managed to propose the big data analytics framework for guidance to build a systematic infectious disease surveillance system that monitoring community case finding, online web-based and mobile phone data for infectious disease surveillance in Tanzanian. With such a framework, we can systematically collect infectious disease data from the Internet and mobile phones through web-based mapping, search engines, social networks, and local infectious disease cases, thus providing accurate and timely information to decision-makers. We believe that such a framework is very important to patients, researchers, epidemiologists, decision-makers, and other public healthcare providers.

Second, the techniques and methods used are based on big data analytics using the MapReduce algorithm which has been reported as the best performing algorithm in big data analytics. It allows distributed and parallel processing of large-scale data sets across commodity computers cluster which can easily be applied in resource-limited setting counties like Tanzania to improve high-performance computation.

The study has the following limitations which can be explored by the researchers for further studies: It is easy to imagine the potential benefits of extracting healthcare information from big data, access to such information is limited, costly, security and legal concerned and even impossible for many research societies. The online healthcare data needs to be evaluated and filtered to increase the signal-to-noise ratio for suitable healthcare data analysis. Another limitation is that most people in rural areas in Tanzania tend to lack or have limited Internet access. Online healthcare data needs web queries and search engines based surveillance. This depends on the availability of sufficient web-internet access to generate signals for data response.

References

[1] P. Sjoquist, “Tanzania,” Institutional Adjust. Econ. Growth Small Scale Ind. Econ. Transit. Asia Africa, pp. 163–199, 2019, doi: 10.4324/9780429441561-7.

[2] M. Acheampong, C. Ejiofor, A. Salinas-Miranda, F. M. Jaward, M. Eduful, and Q. Yu, “Bridging the under-five mortality gap for Africa in the era of sustainable development goals: An ordinary least squares (OLS) analysis,” Ann. Glob. Heal., vol. 84, no. 1, pp. 110–120, 2018, doi: 10.29024/aogh.9.

[3] J. D. Keenan et al., “Azithromycin to reduce childhood mortality in sub-Saharan Africa,” N. Engl. J. Med., vol. 378, no. 17, pp. 1583–1592, 2018, doi: 10.1056/NEJMoa1715474.

[4] M. C. Masanja H, De Savingy D, Smithson P, Schellenberg J, John T, “Child survival gains in Tanzania: analyisis of data from demographic and health surveys,” Lancet, vol. 371, no. table 1, pp. 1276–1283, 2008, doi: http://dx.doi.org/10.1016/S0140-6736(08)60562-0.

[5] B. M. Nkowane, “Streamlining and strengthening the Disease Surveillance System in Tanzania: Disease Surveillance System review, asset mapping, gap analysis, and proposal of strategies for streamlining and strengthening disease surveillance,” 2019.

[6] A. M. Kanté et al., “Childhood Illness Prevalence and Health Seeking Behavior Patterns in Rural Tanzania,” BMC Public Health, vol. 15, no. 1, pp. 1–12, 2015, doi: 10.1186/s12889-015-2264-6.

[7] D. M. Morens and A. S. Fauci, “Emerging Infectious Diseases: Threats to Human Health and Global Stability,” PLoS Pathog., vol. 9, no. 7, pp. 7–9, 2013, doi: 10.1371/journal.ppat.1003467.

[8] J. L. Hurtado, A. Agarwal, and X. Zhu, “Topic discovery and future trend forecasting for texts,” J. Big Data, 2016, doi: 10.1186/s40537-016-0039-2.

[9] K. Priyanka and N. Kulennavar, “A survey on big data analytics in health care,” IJCSIT, Int. J. …, vol. 5, no. 4, pp. 5865–5868, 2014, doi: 5: 5865-5868.

[10] N. K. Nagwani, “Summarizing large text collection using topic modeling and clustering based on MapReduce framework,” J. Big Data, pp. 1–18, 2015, doi: 10.1186/s40537-015-0020-5.

[11] M. B. Chandak, “Role of big - data in classification and novel class detection in data streams,” J. Big Data, 2016, doi: 10.1186/s40537-016-0040-9.

[12] I. El Alaoui, Y. Gahi, R. Messoussi, Y. Chaabi, A. Todoskoff, and A. Kobi, “A novel adaptable approach for sentiment analysis on big social data,” J. Big Data, 2018, doi: 10.1186/s40537-018-0120-0.

[13] A. Mavragani and G. Ochoa, “Forecasting AIDS prevalence in the United States using online search traffic data,” J. Big Data, 2018, doi: 10.1186/s40537-018-0126-7.

[14] A. Ed and K. Maalmi, “A new Internet of Things architecture for real - time prediction of various diseases using machine learning on big data environment,” J. Big Data, 2019, doi: 10.1186/s40537-019-0271-7.

[15] C. Baechle and A. Agarwal, “A framework for the estimation and reduction of hospital readmission penalties using predictive analytics,” J. Big Data, pp. 1–15, 2017, doi: 10.1186/s40537-017-0098-z.

[16] M. Herland, T. M. Khoshgoftaar, and R. A. Bauder, “Big Data fraud detection using multiple medicare data sources,” J. Big Data, pp. 1–21, 2018, doi: 10.1186/s40537-018-0138-3.

[17] R. Authority, “For the year ended 30th June, 2010,” 2010.

[18] A. Nillson, “Using mass media as channel for healthcare information,” pp. 1-, 2014.

[19] J. Adinan, D. J. Damian, N. R. Mosha, I. B. Mboya, R. Mamseri, and S. E. Msuya, “Individual and contextual factors associated with appropriate healthcare seeking behavior among febrile children in Tanzania,” PLoS One, vol. 12, no. 4, pp. 1–15, 2017, doi: 10.1371/journal.pone.0175446.

[20] W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: promise and potential,” Heal. Inf. Sci. Syst., vol. 2, no. 1, p. 3, 2014, doi: 10.1186/2047-2501-2-3.

[21] H. A. N. Hu, Y. Wen, S. Member, and T. Chua, “Toward Scalable Systems for Big Data Analytics : A Technology Tutorial,” IEEE Access, vol. 2, pp. 652–687, 2014, doi: 10.1109/ACCESS.2014.2332453.

[22] M. Torabzadehkashi, S. Rezaei, A. Heydarigorji, H. Bobarshad, and V. Alves, “Computational storage : an efficient and scalable platform for big data and HPC applications,” J. Big Data, 2019, doi: 10.1186/s40537-019-0265-5.

[23] A. K. Roy, “Impact of Big Data Analytics on Healthcare and Society Journal of Biometrics & Biostatistics Impact of Big Data Analytics on Healthcare and Society,” no. January, 2016, doi: 10.4172/2155-6180.1000300

[24] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, “query data,” Nature, vol. 457, no. 7232, pp. 1012–1014, 2009, doi: 10.1038/nature07634.

[25] A. Signorini, A. M. Segre, and P. M. Polgreen, “The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic,” PLoS One, vol. 6, no. 5, 2011, doi: 10.1371/journal.pone.0019467

[26] H. Achrekar, A. Gandhe, R. Lazarus, S. Yu, and B. Liu, “Twitter improves seasonal influenza prediction,” 2003.

[27] Q. Yuan, E. O. Nsoesie, B. Lv, G. Peng, R. Chunara, and J. S. Brownstein, “Monitoring Influenza Epidemics in China with Search Query from Baidu,” vol. 8, no. 5, 2013, doi: 10.1371/journal.pone.0064323.

[28] A. Naveen, B. Antarip, D. Sumit, N. Saurav, and P. Rajiv, “The Abzooba Smart Health Informatics Platform (SHIP) – From Patient Experiences to Big Data to Insights,” arXiv Prepr. arXiv1203.3764, p. 3, 2012.

[29] A. Wesolowski et al., “Quantifying the impact of human mobility on malaria,” Science (80-.)., vol. 338, no. 6104, pp. 267–270, 2012, doi: 10.1126/science.1223467.

[30] Y. Wang, L. A. Kung, and T. A. Byrd, “Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations,” Technol. Forecast. Soc. Change, vol. 126, pp. 3–13, 2018, doi: 10.1016/j.techfore.2015.12.019.

[31] H. M. Al-barhamtoshy and F. Eassa, “A Data Analytic Framework for Unstructured Text A Data Analytic Framework for Unstructured Text,” no. June, 2014, doi: 10.13140/2.1.4330.0485.

[32] M. Barkhordari and M. Niamanesh, “Chabok : a Map - Reduce based method to solve data warehouse problems,” J. Big Data, 2018, doi: 10.1186/s40537-018-0144-5.