2021-11

Big data analytics framework for childhood infectious disease surveillance system using modified mapreduce algorithm: a case study of Tanzania

Mwamnyange, Mdoe

NM-AIST

http://doi.org/http://doi.org/10.58694/20.500.12479/1634

BIG DATA ANALYTICS FRAMEWORK FOR CHILDHOOD INFECTIOUS DISEASE SURVEILLANCE SYSTEM USING MODIFIED MAPREDUCE ALGORITHM: A CASE STUDY OF TANZANIA

Mdoe Aden Mwamnyange

A Dissertation Submitted in Partial Fulfilment of the Requirements for the Degree of Master's in Information and Communication Science and Engineering of the Nelson Mandela African Institution of Science and Technology

Arusha, Tanzania

November, 2021

ABSTRACT

Tanzania has been affected with a potential emerging and re-emerging of infectious diseases such as diarrhea, acute respiratory infections, pneumonia, hepatitis, and measles. There is an increasing trend for the occurrences of new emerging pandemic diseases such as the coronavirus (Covid-19) in 2020 as well as re-occurrence of old infectious diseases such as cholera epidemic in 2015-2017, chikungunya and dengue fever outbreak in 2010, 2012, 2014, 2018, and 2019 which affected different regions in Tanzania. These diseases by far are the main causes of the high mortality rate for women and children of 0-5 years of age. The traditional disease surveillance system as the foundation of the public healthcare practices has been facing challenges in data collection and analysis using health big data sources to prevent and control infectious diseases. Health big data sources on infectious diseases have been recognized as the potential supplement for the provision of evidence-based decision-making worldwide. Tanzania as one of the resource-limited setting countries has lagged because of the challenges in information technology infrastructure and public healthcare resources. The traditional disease surveillance system is still paper-based, semi-automated, and limited in scope which relies on clinical-oriented patient data sources and leaving out nontraditional and pre-diagnostic unstructured big data sources. This research study aimed to improve the traditional infectious disease surveillance system to employ big data analytics technology in healthcare data collection and analysis to improve decision-making. Big data analytics framework for the childhood infectious disease surveillance system was developed which guides healthcare professionals to streamline the collection and analysis of health big data for infectious disease surveillance. The framework was then fairly compared with the existing framework in its performance using infrastructures, data size and transformation, and running-time execution of the systems. The experimental results indicate the efficiency of the framework system performance with the highest running time execution of about 56% quicker over the traditional system. Also, it has the best performance in processing multiple data structures using additional processing units. In particular, the proposed framework can be adopted to improve the prenatal and postnatal healthcare system in Tanzania.

DECLARATION

I, Mdoe Aden Mwamnyange do hereby declare to the Senate of The Nelson Mandela African Institution of Science and Technology (NM-AIST) that this dissertation titled "*Big Data Analytics Framework for Childhood Infectious Disease Surveillance System using Modified MapReduce Algorithm: A Case Study of Tanzania*" is my original work and that it has neither been submitted nor being concurrently submitted for degree award in any other institution.

Mdoe Aden Mwamnyange

Name and Signature of the Candidate Date

The above declaration is confirmed by:

Dr. Edith Luhanga

Name and Signature of Supervisor Date

COPYRIGHT

CERTIFICATION

The undersigned certifies that has read and hereby recommend for acceptance by the Nelson Mandela African Institution of Science and Technology a dissertation titled "*Big Data Analytics Framework for Childhood Infectious Disease Surveillance System using Modified MapReduce Algorithm: A Case Study of Tanzania*" in partial fulfilment of the requirements for the degree of masters in Information and Communication Science and Engineering at Nelson Mandela African Institution of Science and Technology, Arusha Tanzania.

Dr. Edith Luhanga

___________________________	____________________
Name and Signature of Supervisor	Date

ACKNOWLEDGEMENTS

Firstly, I would like to thank God Almighty for His endless blessings, good health, and opportunities He has provided me with that enabled me to undertake a Master's program at The Nelson Mandela African Institution of Science and Technology (NM-AIST) which incorporated this study.

My sincere thanks go to my supervisor; Dr. Edith Luhanga, and Mr. Sanketh Thodge for their support, guidance, and valuable time they dedicated to ensuring this research work is a success.

I would also like to express my gratitude to the African Development Bank (AfDB) for their funding of my studies at NM-AIST and support in this research work.

Special thanks also go to my wife, my kids, and all my family members for their support, prayers, and encouragement during the whole period of undertaking my master's degree studies at NM-AIST.

Sincere appreciation goes to my colleagues at NM-AIST, all staff members, the surrounding community, and my co-workers at Information and Resource Center department (NM-AIST), thank you all.

DEDICATION

I would like to dedicate this work to my family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS AND SYMBOLS

API	Application Programming Interface
ARI	Acute Respiratory Infections
BDAF-CIDSS	Big Data Analytics Framework for Childhood Infectious Disease
BTSM	Brain Tumor Social Media
CDC	Centers for Disease Control and Prevention
CIDSS	Childhood infectious disease surveillance system
CSSE	Center for Systems Science and Engineering
DHIS-2	District Health Information System – 2
EVD	Ebola virus disease
ETL	Extraction, Transformation, and Load
EWARS	Early Warning Alerts and Response System
FIDA	Framework for Infectious Disease Analysis
GPHIN	Global Public Health Intelligence Network
HDFS	Hadoop Distributed File System
HMIS	Health Management Information System
HSSP	Health Sector Strategic Plan
IDWER	Infectious Disease Week Ending Report
IDSS	Infectious Disease Surveillance System
IDSR	Integrated Disease Surveillance and Response
IHR	International Health Regulations, 2005
IoT	Internet of Things
IVP	Immunization and Vaccine Development Program
MRQL	Map Reduce-Based Query Language
NHIF	National Health Insurance Fund
NLP	Natural Language Processing
NMCP	National Malaria Control Program
ODK	Open Data Kit
ProMED	Program for Monitoring Emerging Diseases
RDBMS	Relational Database Management System
R-SQL	Relational Structured Query Language
RSS	Really Simple Syndication
RFID	Radio Frequency Identification
R & D	Research and Development

SMS	Short Message Service
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SQL	Structured Query Language
TDV 2025	Tanzania Development Vision, 2025
TNHP 2017	Tanzania National Health Policy, 2017
USA	United States of America
WHO	World Health Organization

CHAPTER ONE

INTRODUCTION

1.1 Background of the Problem

Tanzania has been affected with a developing emerging and re-emerging of infectious diseases including diarrhea, acute respiratory infections (ARI), pneumonia, hepatitis, and measles. The occurrences of pandemic diseases have been increased recently such as coronavirus of 2019 (Covid-19) in 2020 (Jeon *et al.,* 2020), also frequently occurring of old infectious diseases such as cholera infectious disease in 2015-2017 in which 25 600 people were affected and 401 died. The re-occurrence of chikungunya and dengue fever outbreak in 2010, 2012, 2014, 2018, and 2019 have been affected different regions including Dar es Salaam, Kilimanjaro, Arusha, Morogoro, Tanga, Zanzibar, and other regions (Vairo *et al.,* 2014; Chipwaza *et al.*, 2021; Kajeguka *et al.*, 2016). For instance, in 2019 the last dengue fever was reported in Dar es Salaam, Tanga, Pwani, Morogoro, Singida, and Kilimanjaro in which 3500 cases were reported and 13 deaths (Salvatory Kalabamu & Maliki, 2021). These diseases by far are the main causes of high mortality rates for women and children of 0-5 years of age (Mangu *et al.,* 2016). Continuous attention is needed to counter the weaknesses of the existing infectious disease surveillance system to improve performance which in turn reduces the rate of child mortality and morbidity in the country. Tanzania Development Vision 2025 (Tanzania Development Vision [TDV], 2025) provides guiding principles and direction for long periods of development. According to TDV 2025, Tanzania needs to achieve high quality of livelihood for its citizens to attain sustainable growth by 2025. It identifies the health sector as the number one priority geared towards high quality of life standards for all Tanzanians (TDV, 2025). The National Five Year Development Plan 2016/2017 - 2020/2021 health objectives emphasize the provision of high-quality healthcare services to ensure people participate in social-economic development. The Tanzania Health Sector Strategic Plan 2015 – 2020 (HSSP IV) document emphasizes reaching a high population located in the remotes areas to provide healthcare services delivery (Care, 2020). Also, the Tanzania Health Sector Strategic Plan 2021-2026 (HSSP V) emphasizes the same with the addition of leaving no one behind in healthcare services delivery (Tanzania Health Sector Strategic Plan [HSSP], 2021). These articles emphasize reaching the expectations of the high population adhering to quality standards and evidence-based interventions in healthcare services delivery. Their specific objectives involve improvement of health care services, provide equal access to healthcare services to all people based on geographic areas with high disease burdens and involve community partnership interactions to improve healthcare services delivery. To address these plans, strengthening the health services delivery system using big data sources to improve the health of mothers and children is necessary.

According to Tanzania National Health Policy of 2017 (TNHP, 2017) Tanzania has been reduced the morbidity and mortality rate to the children through providing vaccines including Rota vaccine, Poliomyelitis (OPV), Tuberculosis (BCG), Pneumococcal Vaccine (PCV13), Measles/Rubella (MR), Hepatitis B, Diphtheria, among others. Tanzania has also supported global initiatives including Poliomyelitis Eradication, Measles, and Neonatal Tetanus elimination for child survival. The survival of children has been improved over 10 years. Children less than 5 years mortality rate has been reduced from 81/1000 in 2010 to 67/1000 in 2015/16. Tanzania provides three health services for children; child growth monitoring, outpatient curative care, and immunization services. Maternal health care including malaria in pregnancy, nutrition, and postnatal care has also been improved (TNHP, 2017). However, diarrhea, ARI, pneumonia, hepatitis, and measles persist to be among the major public health threats in Tanzania hence close attention is needed to control the infectious diseases to reduce child mortality and morbidity.

Tanzania has been investing in high-quality healthcare technologies services including diagnosis using modern electronic technology to improve the delivery of healthcare services which increases the availability of digital health records to achieve the goal of HSSP V and TDV 2025. The technologies include modern laboratory technology, digitization of radiology and imaging equipment, electronic Health Management Information System (HMIS) and District Health Information System (DHIS-2), health insurance systems, availability of CT-Scan, Digital X-Ray and MRI services, laboratory information systems, integration of laboratory services in infectious disease surveillance and founding a National Public Health Laboratory Services for investigation of pandemic diseases. However, the implementation of the new technology has been facing challenges including the high cost of the equipment and consumables, changes in technology, lack of technical skills among physicians to provide services and maintenance of health technology equipment and consumables. These electronic technologies prove the signs of emerging a high volume of health big data sources to facilitate healthcare services in Tanzania. Today, health big data sources involve health data from the Internet, health database systems, social media, health websites, mobile applications, online magazines, online healthcare information aggregates, Web search queries, online discussion messages posts, Really Simple Syndication (RSS) feeds, patient past laboratory information, free-text document reports, expert opinion on diagnoses and medications provide wide opportunities for information exchange through the Web-based systems. These data if properly extracted and analyzed using big data analytics tools can greatly aid in healthcare decision-making.

Despite the efforts and achievements, Tanzania has been facing challenges in the implementation of these modern technologies in infectious disease surveillance for children under 5 years. The challenges are including shortage of human resources for healthcare, increase in emerging and re-emerging of infectious diseases; shortage of healthcare staff to conduct infectious disease surveillance, lack of collaboration among healthcare stakeholders in addressing the re-occurrences of the infectious diseases, and increase in maternal-child/infant mortality and morbidity. Community-based infectious disease surveillance has not been integrated into the national system for disease surveillance and outbreak investigation.

In Tanzania, there are three common types of infectious disease surveillance systems; HMIS/DHIS-2 which in Kiswahili acronym as MTUHA, Infectious Disease Week Ending Report (IDWER), and Integrated Disease Surveillance and Response (IDSR). The MTUHA involves the following infectious diseases surveillance; measles, acute respiratory infections, diarrhea, pneumonia, eye, ear, and skin diseases. The IDWER involves Acute Flaccid Paralysis (AFP), meningococcal meningitis, yellow fever, cholera, malaria, measles, diarrhea, plague, dysentery, typhoid, and others. The IDSR system which was adopted by the government in 2011 supports disease surveillance for 34 diseases including AFP, malaria, cholera, measles, rabies, meningitis, diarrhea under 5 years, plague, polio, typhoid, pneumonia under 5 years, bacillary dysentery, among others (Nkowane, 2019).

Despite the efforts of the existing disease surveillance systems, Tanzania still suffers from the emerging and re-emerging of infectious diseases which cause a high mortality rate for women and children. The weakness of these systems to accommodate active and passive disease surveillance activities is among the reasons. In the existing system, passive disease surveillance reporting activity occurs on a weekly and monthly basis through the HMIS /DHIS-2. Active disease surveillance activities which are disease-specific based on geographical location and time are not well conducted. This is due to a lack of reliable health data sources to support staff for daily and weekly disease surveillance. The traditional disease surveillance activities depend on the use of health data from the IDSR and HMIS system. The disease surveillance activities are not considered a high priority at the local geographical level relative to curative services. The systems do not support a weekly transmission of all infectious disease data including diseases for immediate notification and discrepancies in its reporting mechanism in disease specific-program health data reporting. The system relies on the existing hospital and clinical databases which involve structural raw data collected by medical practitioners from the clinical-patient data sources. Health data collection depends on the access of the patient to healthcare facilities such as hospitals, healthcare centers, and dispensaries and leaving out those who do not access health facilities due to various social and economic reasons. The available nontraditional and pre-diagnostics healthcare big data sources from the modern health technology equipment, community environments, social media, Internet-based, and other online news archives which can be collected and analyzed for infectious disease surveillance and outbreak investigation are not well-utilized meaning that important patterns and trends for evidence-based decision making are missed (Bansal *et al.*, 2016).

Several research papers and articles have been published to address the implementation of health big data sources in disease surveillance (Salathé, 2016). From these work, six motivation factors have been motivating us to increase interest in using big data analytics to enhance childhood infectious disease surveillance systems (CIDSS) to improve decision-making in Tanzania. The first motivation factor was that healthcare big data with its 3Vs (volume, velocity and, variety) characteristics provides many opportunities to enhance infectious diseases surveillance.

Volume refers to the size or amount of infectious disease data to be examined. Healthcare data can be considered 'big' when the size of the data hinders data processing using a traditional software system. Today, in Tanzania, forty-six referral, regional and district hospitals have started to use electronic hospital administration and laboratory data management systems. The electronic systems such as Jeeva, WebERP, Care2X, Harmony, Bumi expert, Daisa among others have been deployed in the hospitals. The laboratory data systems such as Labnet, and BLIS Basic Laboratory Information System have also been deployed to manage laboratory test results in the hospitals. This implies that the modern electronic database systems increases the greater size of data measured from gigabytes (1024 Megabytes) to terabytes (1024 Gigabytes) up to zettabytes (1 trillion bytes) which exceed storage of available hard drives and processing capacity of the traditional computer systems of the hospitals (MoHCDGEC, 2017).

Velocity refers speed with which infectious disease data are collected and transmitted. The traditional CIDSS is hampered by a time delay between infectious disease event reporting, investigation, and intervention. In big data analytics, velocity involves a collection of data in near-real-time to enhance timeliness. Time of infectious disease detection can be an important factor to determine infectious disease outbreak events. The total duration of an infectious disease outbreak event may be linked with the number of days when the first reported case of the patient with symptoms was investigated. In the event when a public infectious disease emerging such as COVID-19 is about to erupt or have already erupted, the analysis results based on the system that frequently generate huge amounts of data from a diverse-mixes of data such as tracking peoples' movements, social interactions, identify and isolate instant infected people for early warning and tracking infectious disease dynamics requires high speed with near-real-time data collection surveillance system.

Veracity refers to the range of healthcare big data sources, file types, and data structures. It involves infectious disease data sources, types, and structures including numerical data structures, free-text data, unstructured text messages, csv files, PDF file format, emails, images, video format, voice sound file format can be handled, processed, and stored all together. Preliminary evaluation proves that these functions cannot be accommodated in the traditional infectious disease surveillance in Tanzania.

The second motivation factor was the need for using additional health data sources to support traditional infectious disease surveillance systems. Big data on infectious diseases are available from several different big data sources and technologies. For instance; Smartphones technology enhances understanding of infectious disease transmission through call-data records and Short Message Service (SMS) data sources collection and analysis. It involves a collection of data on the time communication was made, geographical location, call duration, and data size transmitted. By connecting the telecommunication tower location to a map, it becomes possible to track phone call movement for tracking infectious disease dynamics. This tool has been used in the Ebola Track system to monitor persons who visited Ebola outbreak areas in Western Africa in 2014 (Chowell *et al.*, 2016). Social media network technologies including Baidu, Instagram, Twitter, WhatsApp, and others have been used in big data analytics to strengthen infectious diseases surveillance to prevent pandemic diseases (Ahmed *et al.,* 2020). Free-text data, short-text messages, short stories, likes, hashtags, and location data from the social media network can be collected through the Application Programming Interface (API) in real-time to understand human mobility patterns, human contact rates during infectious disease outbreaks using spatial clusters, social network analysis, and communication pattern analysis (Kwok *et al.,* 2021). In big data analytics, an internet-based system such as a Web-based system, online systems, forums, online news aggregates, email systems, and magazines can be used to monitor, summarize, filter, and map infectious diseases data to improve disease surveillance. This can be done through the collection and analysis of free-text messages, emails, disease report cases, expert opinion reports, stories associated with health news, and others that cannot be done in the traditional surveillance system in Tanzania (Heisey-Grove *et al.*, 2020).

The third motivation factor was that the traditional surveillance system has been challenged by the re-occurrences of infectious diseases including cholera, dengue fever, diarrhea, malaria, and pneumonia in Tanzania (Mwanyika *et al.,* 2021). To address the challenges, infectious diseases surveillance systems and significant syndrome clinical-focused surveillance systems that function independently to monitor and control childhood infectious diseases at the local geographic location have been invested in Tanzania. These systems may be improved to consider using opportunities of accessing additional crowd-sources of information on symptoms from the Internet-based systems in near-real-time before clinical diagnosis to understand infectious disease transmission and dynamics from specific geographical locations to improve decision-making.

The fourth motivation factor was the interest in the integration of local pharmacies and retail business information systems with the traditional disease surveillance system to improve decision-making. Pharmacies and local medicine shops-based surveillance of drug prescriptions and non-prescription retail sales have been successfully conducted in many countries in the world including Africa, Europe, and America. Data from shopper loyalty cards, retail scanning data, and medicine-shopping database search have been used to collect utility of over-the-counter sales information to identify infectious disease surveillance and disease outbreaks investigations. Retail data of over-the-counter food sales information has been used to conduct food-borne infectious disease surveillance. For instance, in 2012 credit cards having retail shopping information histories of 3-months from the supermarkets were used to determine frozen fruit blend as the source of hepatitis A outbreak in British Columbia (Swinkels *et al.*, 2014). The increase in drug retail sales information for diarrhea and nausea from local pharmacies has also been associated with the water-borne disease outbreaks of Cryptosporidium, *E. coli* in Canada (Pivette *et al.*, 2014).

The fifth motivation factor was the interest in mining a high volume of electronic health records databases from the existing modern health technology systems using an efficient high-performance computer system to investigate infectious disease outbreak cases (Bennett & Doub, 2011). Today, there is an increasing availability of electronic medical records of the children and other patients from the hospitals, healthcare centers, dispensaries, and traditional midwifery which are available from the HIMS / DHIS-2 systems in Tanzania. Example of these electronic health records involves records from the medical doctors and physicians, laboratory test results, photo scanning images, healthcare insurances, out-patient and in-patient hospital records, hospital discharge records, death certificates records among others. These data if properly extracted and analyzed using big data analytics tools can help to improve infectious disease surveillance and disease outbreaks investigations. Electronic health records data have been used to conduct real-time and passive public health surveillance to support local geographic location disease surveillance in many countries in the world. For instance, detailed electronic insurance claim forms from the medical doctors and physicians have been used to conduct an investigation on infectious disease of the influenza-like illness (ILI) at the local geographical location in the USA (Ginsberg *et al.,* 2009). While the high volume of electronic health records databases appears to be a meaningful infectious disease big data source, several challenges may limit the application functions of these big data sources to enhance infectious disease surveillance and disease outbreaks investigation in Tanzania. Most electronic health records systems such as HIMS and DHIS-2 have been developed to serve clinical health hospital systems and are owned by the government and not for public or individual consumption without permission. There are strict policy and privacy questions on which health data should be used for the public. Confidentiality is also a great issue to access the electronic health records databases in the country. A patient might have multiple healthcare insurance providers with different insurance coverage such as National Health Insurance Fund (NHIF), Assemble insurance, Platinum health insurance, Jubilee life insurance, and others for medicine prescriptions that require an active networking database system that connects all individual systems.

The sixth motivation factor was the self-participatory of the remote areas population data sources in the public health applications. While the traditional CIDSS relies on the hospital and clinical patient data for disease surveillance, the online self-participatory disease surveillance system helps to communicate with a great number of people to report symptoms and health status. Internet-based and mobile applications have been effectively used to facilitate infectious diseases data collection from the local geographical areas and disadvantaged-population of the remote areas such as influenza, zika, Ebola, COVID-19, and others. For instance, in Europe, the Influenza.net online surveillance system has been successfully used to collect infectious disease data of influenza-like illness in all European countries (Paolotti *et al.,* 2014). Flue Near You online infectious disease surveillance system has also been successful in collecting crowd-sourced health data of influenza-like illness online across the USA (Ginsberg *et al.,* 2009). In 2009, the vaccination sentiments measurements of an influenza A(H1N1) pandemic infectious disease measured in the Twitter social network were correlated positively with the vaccination measurement report conducted across the USA (Salathé & Khandelwal, 2011). These studies indicate that the online health big data streams can help the traditional disease surveillance systems to enhance decision-making if the big data framework and limits of its applications are well-identified in Tanzania.

However, the high volume of healthcare big data in modern electronic systems, social media, and Internet-based systems such as online free-text messages and health news information may limit the reliability and effective use of this information to precisely detect the indicative signals of infectious disease surveillance and disease outbreaks investigation. The reliability of online healthcare big data has been disputable due to a lack of validity, consistency, and accuracy problems. Huge amounts of infectious disease data are continuously generated through the Internet-based systems, but the values are covered behind the noise of the data. Therefore, data cleaning, mining, merging, linking, transforming, and reduction of noisy data is inevitable. To apply big data sources in disease surveillance and improve data reliability, hospitals need to identify sections where big data can contribute values for infectious diseases surveillance by evaluating the impact of noisy data, confirm the reliability of existing data, and confirm how the noisy information can affect the results. In order to improve health data validity and reliability from the internet-based big data sources, hospitals may need to develop and maintain their own independent online healthcare data platform repository such as secure online websites, set disease-specific social media hashtag user account, or using trusted public healthcare data repository such as Github to become a standard healthcare data sharing source whereby individuals interested to analyze the infectious diseases trends and their impacts to the society can perform the function. For instance, The Brain Tumor Social Media (#BTSM) hashtag on Twitter which was established in February 2012 helps to display all tweets information aggregated by the same brain tumor disease-specific topic from the subscribed big data internet sources (Feliciano *et al.,* 2020). HealthMap FoodBorne system established in Boston children's hospital is also used by the hospital as a specific health data source repository to identify tweets regarding foodborne illness in the USA.

The main purpose of the study was to design and develop Big Data Analytics Framework for Childhood Infectious Disease Surveillance System (BDAF-CIDSS) to streamline health big data collection and analysis to enhance early childhood infectious disease surveillance in Tanzania. The study focused on determining how the Internet-based health big data are collected, analyzed, and transformed into useful information for preoccupation with healthcare planning, implementation, and decision making. A BDAF-CIDSS was then designed and developed to guide healthcare physicians to collect and analyze the health big data sources of the infectious diseases affecting children in Tanzania. Then, the framework was validated using various use-cases scenarios to facilitate actual implementation.

1.2 Statement of the Problem

Data collection analysis and interpretation are crucial in any infectious disease surveillance system to facilitate disease detection, identification, and early warning. But, the traditional infectious disease surveillance system faces challenges of health data collection and analysis of multiple big data sources to facilitate informed evidence-based decision-making. Big data capture, searching, storage, sharing, integration, analysis, and visualization are the key challenges of the traditional system. The big data analytics framework makes it possible for the traditional system to explore and manage the health big data sources to enhance informed evidence-based decision-making to fight against the re-occurrences of infectious diseases.

1.3 Rationale of the Study

Tracking diseases, pathogens, and clinical outcomes through data collection and analysis are the main focus of any public health disease surveillance system to control the emerging and re-emerging of infectious diseases. Regardless of the emerging of various tools to enhance infectious disease surveillance such as the Internet, smartphones, online health database system, websites, social media, and others that could enhance infectious disease surveillance hospitals and healthcare centers have yet managed to use the technologies to reduce the trend of the occurrences of the infectious diseases for proper evidence-based decision-making. Data collection and analysis do not include other diverse mixes of Internet-based healthcare-related data from other multiple channels including modern health technology systems, community case findings, health insurances, public pharmacies, Research and Development (R&D) institutions, social media, and online information archives which used to facilitate passive and active infectious disease surveillance.

In Tanzania, there are three common types of infectious disease surveillance systems HMIS/DHIS-2, IDWER, and the IDSR system. The infectious disease surveillance activities are done through the following; National level disease-specific surveillance and the regional, council, and health facility levels as presented in the framework in Fig. 1 as follows:

1.3.1 National Level

At the national level, 15 disease specific-programs are linked at the IDSR system whereby each program uses data from the IDSR system for analysis. The operation of these programs often depends on the availability of supporting funds from donors or other sources to support their surveillance activities. For instance, National Malaria Control Program (NMCP) tracking malaria-reported cases, inpatients, and deaths cases at the health centers. It uses disease surveillance health data from the IDSR and HMIS system for analysis. Another example is the Immunization and Vaccine Development (IVP) program. The program actively monitored cases of acute flaccid paralysis and measles through case-based disease surveillance. It was also funded by the resources used for polio eradication which has been phased out.

1.3.2 Regional Levels

At the regional level, disease surveillance is done through the regional medical office. The IDSR staff from the regional level coordinates disease surveillance activities, mobilizes resources, and provides technical assistance at the lower level. Regional disease surveillance staffs are the employees of the regional administration. Disease surveillance activities depend on the availability of information in the IDSR and HMIS system for analysis.

1.3.3 Council/District Level

Infectious disease surveillance at the council level is done through the council health office. Surveillance staffs at the district/council levels are the employees of the government who work under the council medical office. The operational support of the disease surveillance activities at the council depends on the availability of funds for fuel, transport, and stipends for staff to travel to the regional level and health center levels. Usually, the council administration offers insufficient funds to support disease surveillance activities because of budget constraints. Therefore, disease surveillance activities are done quarterly instead of on a daily or monthly basis based on the IDSR disease surveillance standard.

1.3.4 Health Center/Dispensary Levels

At the dispensary and health center, there are designated people from the community and the medical or clinical officer from the dispensary. The designated people are trained to conduct infectious disease surveillance activities at the local geographical level using a paper-based system. They report immediately notifiable infectious disease report cases and prepare weekly ending reports using a paper-based system before submitting them to the council level. Disease surveillance activities depend on face-to-face interactions of these designated trained people with the members of the community.

The traditional CIDSS is paper-based, semi-automated, and limited in scope which depends on the availability of supporting funds, trained staff, and limited health data sources. Regardless of the establishment of the modern health technology systems at hospital levels, every hospital has its independent way of keeping track of the healthcare data of its patients. The great challenges appear when there is an issue of information exchange among hospitals and government administration. This can be observed when a patient is referred from one hospital to another. It is very difficult for the patient data of infectious disease collected from one hospital-level accessed by another hospital level. Health data from Internet-based systems and other automated systems are not incorporated in the surveillance at all levels. The systems have severe time lags between the infectious disease event and reporting, also a limitation on the spatial resolution which limits analytics findings and their interpretations for decision making. This situation encourages patients' readmission, unnecessary treatments, lack of patients' medical history, lack of past laboratory test results, limited analysis, and unnecessary costs.

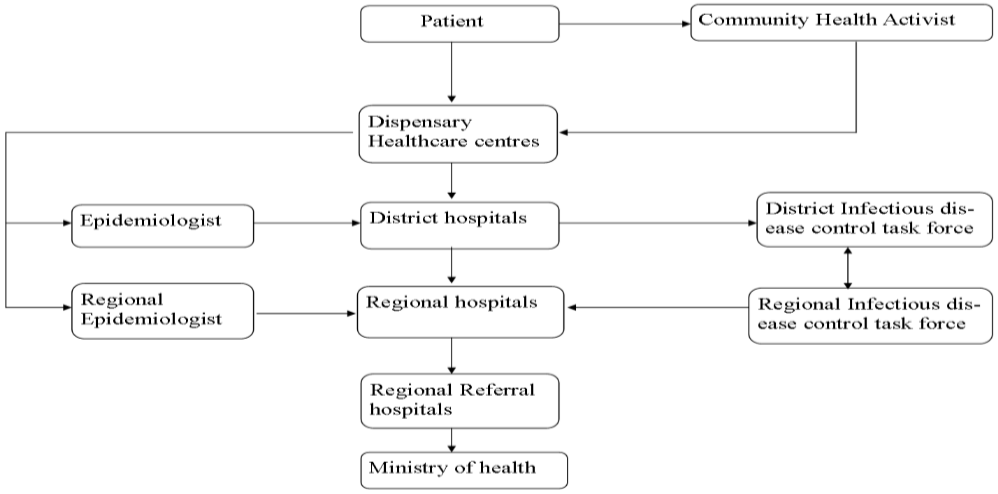

Patient
Community Health Activist
Dispensary Healthcare centres
Epidemiologist
District hospitals
District Infectious disease control task force
Regional Epidemiologist
Regional hospitals
Regional Infectious disease control task force
Regional Referral hospitals
Ministry of health

Figure 1: Traditional Childhood Infectious Disease Surveillance System Framework

Due to the challenges of emerging and re-emerging infectious diseases in Tanzania, it has become very difficult for the existing disease surveillance system to detect and analyze small to medium-sized outbreaks of childhood infectious diseases. As a result, these outbreaks remain hidden and can be distributed unnoticed over a wide local geographic area because of the growing local food processing technologies in the country. The emerging of Covid-19 in 2020, and the re-emerging of the old infectious diseases such as cholera (2015-2017), dengue fever (2010, 2012, 2014, 2018, and 2019), chikungunya, and among others, demonstrate how infectious diseases surveillance system needs not only an emergency management system but also a system that analyze data leading to timely infectious disease detection with dynamic, sophisticated data collection, analysis, visualization, and reporting (Salvatory Kalabamu & Maliki, 2021).

In today's world, modern health technology systems, mobile technology, Web 2 technology, and advancement in computer architecture and software promote rapid access of Internet-based healthcare data sources involving nontraditional and pre-diagnostic healthcare data, clinical information and community cases findings, online healthcare information aggregates, Web search queries, online discussion messages posts, RSS feeds, patient past laboratory information, free-text document reports, expert opinion on diagnoses and medications. Therefore, the proposed BDAF-CIDSS model will streamline and support collection and analysis of healthcare big data of infectious diseases in Tanzania for the following reasons; to improve the traditional disease surveillance system in data collection and analysis, to share health data worldwide using the cloud in accuracy and timely manner; to avoid unnecessary treatments; to avoid patient readmission; treating large amounts of child-patients more quickly; conduct real-time analytics and predictions to improve patients lives while cutting costs. Despite many research studies and articles discussing the use of big data analytics in infectious disease surveillance in the world, still there is a lack of an effective framework that suite as a reference for the Tanzanian context. Therefore, the proposed BDAF-CIDSS is necessary to enhance healthcare physicians and decision-makers to prevent and control infectious diseases using big data analytics technology, which is also applicable in instances of other infectious diseases surveillance to support the achievement of HSSP V and the vision of TDV 2025.

1.4 Research Objectives

1.4.1 General Objective

The general objective of this study was to develop Big Data Analytics Framework for Childhood Infectious Disease Surveillance System for the Tanzanian context.

1.4.2 Specific Objectives

The specific objectives of this study were:

(i) To determine what data, methodologies, and tools can enhance Childhood Infectious Disease Surveillance System in Tanzania.

(ii) To develop a Big Data Analytics Framework for Childhood Infectious Disease Surveillance System in Tanzania.

(iii) To validate the Big Data Analytics Framework for Childhood Infectious Disease Surveillance System using use-case scenarios.

1.5 Research Questions

The following research questions were used to guide the conduction of this research study to attain its objectives:

(i) How the traditional Childhood Infectious Disease Surveillance System is performed in Tanzania?

(ii) What are the challenges with the current Childhood Infectious Disease Surveillance System in Tanzania and how should they be addressed?

(iii) What tools and methodologies are suitable for improving the performance of the traditional Childhood Infectious Disease Surveillance System to capture and analyses nontraditional and pre-diagnostic Internet-based childhood healthcare data?

(iv) What are the requirements for a Big Data Analytics Framework for Childhood Infectious Disease Surveillance System in Tanzania?

(v) How acceptable and useful is a Big Data Analytics Framework for Childhood Infectious Disease Surveillance System?

1.6 Significance of the Study

The significance of this research was to facilitate the development of other healthcare big data analytics frameworks and healthcare system design models for other disease surveillance systems through the use of the proposed BDAF-CIDSS to reduce emerging and re-emerging of the infectious diseases geared towards the achievement of HSSP V and vision of TDV 2025. The study would also be interested in the Ministry of Health as a policy and decision-makers to increase the volume and variety of health data in infectious disease surveillance for effective decision-making support for all disease surveillance levels. The Ministry is the key player in the implementation of HSSP V and TDV 2025. Scholars, Physicians, and researchers will use the derived framework to study big data volume, velocity, and variety to effectively apply to understand other infectious diseases dynamics using various big data analysis techniques and algorithms. Further, the study offers the following advantages:

(i) To enable healthcare organizations in Tanzania to understand the application of big data analytics phenomenon in the public health system, capacity, and its benefits to promote them to prepare more applicable data-driven analytics.

(ii) To improve organization performance by creating a more effective data-driven healthcare business model.

(iii) To enhance healthcare organizations to collect, process, and analyze Internet-based healthcare data from the diverse mixes of healthcare network services to support evidence-based decision making.

(iv) To validate the suitability of the proposing Big Data Analytics Framework for Childhood Infectious Disease Surveillance System for use in the Tanzanian context.

1.7 Delineation of the Study

The study aimed to design and develop Big Data Analytics Framework for Childhood Infectious Disease Surveillance System (BDAF-CIDSS) to streamline health big data collection and analysis to enhance early childhood infectious disease surveillance in Tanzania. The focus was to determine how the Internet-based health big data are collected, analyzed, and transformed into useful information for preoccupation with healthcare planning, implementation, and decision making. A BDAF-CIDSS was then designed and developed to guide healthcare physicians to collect and analyze the health big data sources of the infectious diseases affecting children in Tanzania. Then, the framework was validated using various use-cases scenarios to facilitate actual implementation

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter presents the general concepts of infectious diseases, disease surveillance systems, healthcare framework architectures, and related works for tracking infectious diseases with the intent of building more knowledge, understanding, assessment, and analyzes the challenges and successes of the systems in Tanzania and other countries.

2.2 Infectious Diseases

According to the Journal of Immunological Techniques and Infectious Diseases (2020) Infectious diseases are diseases or disorders which are caused by pathogenic micro-organisms including viruses, parasites, bacteria, or fungi. Many of these micro-organisms live in and on human bodies. Normally, they are harmless but under certain conditions, they may cause diseases. Infectious diseases can be transmitted from person to person, either directly or indirectly by various means such as bites by insects or animals or exchanging body fluids through sexual practices, coughs, sneezes, and others. Most infectious diseases have slight complications, however, other infectious diseases like pneumonia, dengue, cholera, and others may become public health threats. Some infectious diseases have been associated with the development of long-term cancer disease. For instance, Hepatitis B and C viruses have been associated with liver cancer, and Helicobacter pylorus has been associated with stomach cancer.

Infectious diseases may cause rapid disease outbreaks in the form of an epidemic or pandemic which may cause high child mortality and morbidity rates if they are not monitored and controlled at the early stages of transmission. The use of functional CIDSS for monitoring trends and its transmission for early warning and alerts for intervention has been proven to improve childhood infectious disease control and management.

2.3 Infectious Disease Surveillance System

Infectious Disease Surveillance System (IDSS) is an important public health measure to assess and monitor the magnitude and trends of childhood infectious diseases to develop an early warning and alert system for initiating public health action and intervention before the disease outbreaks. A functional IDSS involves healthcare data collection, processing, analysis, and interpretation to provide healthcare information for public action. To improve public health security, strengthening infectious disease surveillance systems in the country is essential to achieve international public health regulations security (Kamradt-Scott, 2019).

The popular example of IDSS in Sub-Saharan Africa is the Early Warning Alerts and Response System (EWARS) developed by World Health Organization (WHO)’s initiatives for resource-limited settings countries. The EWARS system rapidly detects priority infectious diseases and allows rapid response before they develop into disease outbreaks. It has been used in Nigeria, Ethiopia, Fiji, South Sudan during the persistent conflicts, and Mozambique. It is capable of tracking and monitoring infectious diseases of more than 500 000 people. In this system, each healthcare site, dispensaries, and healthcare centers receive a mobile phone loaded with a custom-developed infectious disease reporting application that allows healthcare workers to enter information when they see patient with symptoms of one of the priorities diseases or conditions such as acute diarrhea, cholera, measles, malaria and pneumonia (Pavlin & Cic, 2016).

However, the gap in the EWARS system was the slow analysis and response planning due to the use of standard servers with moderate computational capabilities. This approach is inefficient when conducting an in-depth evaluation that requires the collection, processing, and analysis of a large volume of health data of infectious diseases.

2.4 Framework for Healthcare Delivery

A healthcare framework is a systematic approach to healthcare delivery. It involves interdependencies of the physical structure and dynamics of the healthcare system to healthcare delivery. It provides a systematic division of labor and inter-dependencies among healthcare professionals, patients, and other healthcare stakeholders as shown in Fig. 2 (Reid *et al.,* 2005).

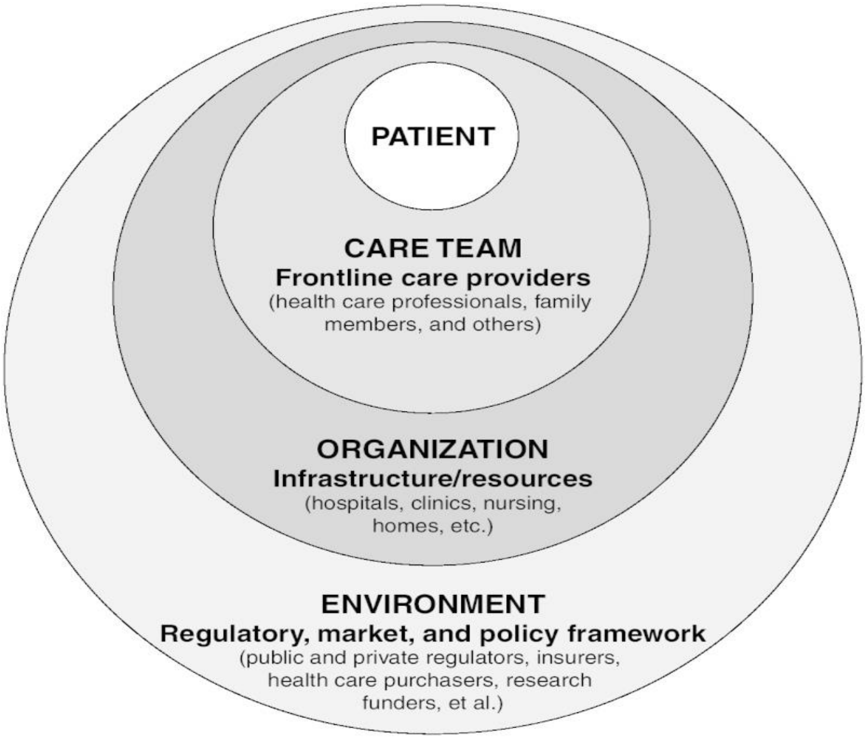

PATIENT
CARE TEAM
Frontline care providers
(health care professionals, family members, and others)
ORGANIZATION
Infrastructure/resources
(hospitals, clinics, nursing, homes, etc.)
ENVIRONMENT
Regulatory, market, and policy framework
(public and private regulators, insurers, health care purchasers, research funders, et al.)

Figure 2: Four levels of healthcare framework (Ferlie & Shortell, 2001)

According to Ferlie and Shortell (2001) the effective healthcare system model should be divided into four levels:

2.4.1 Patient Level

Patient level involves patients who are ill or injured and needs healthcare services or treatment from medical doctors and physicians. In an infectious disease surveillance system, the need for information exchange between patients themselves and healthcare professionals is essential to improve the disease surveillance system and decision-making. Patients and healthcare professionals should have access to the same information or disease surveillance system to facilitate two-way communication between them to reduce the possibility of infectious disease threats in the country.

2.4.2 Care Team Level

Care Team Level involves healthcare professionals, family members, community members, and all others, who supporting healthcare service delivery to the patients. They should be approachable for the needs and preferences of the patients and their families.

2.4.3 Organization Level

Organization level involves hospitals, healthcare centers, dispensaries, and clinics that provide healthcare services to support the work performed by the care teams. It involves the decision-making process, information systems infrastructures, process management, and data storage and management. For an effective disease surveillance system, healthcare organizations need to find ways to invest in information technology infrastructures to support continuous communication and information exchange between patients, healthcare professionals, and organizations to improve evidence-based decision-making (Heisey-Grove *et al.*, 2020).

2.4.4 Environments Level

Environments level involves political, economical, policymakers, public and private regulators, health insurers, and healthcare research funders or donors which influence the structures and performance of the healthcare system. These actors should support the goals and objectives of the effective disease surveillance system to improve the performance of the infectious disease surveillance to enhance evidence-based decision-making.

2.5 Big Data Analytics in Public Health Delivery

Big data analytics technology in public health information dissemination can improve efficiency, productivity, and sustainability of the healthcare system by providing efficient data collection, mining, and real-time data analytics. In big data analytics phenomena in public health practices, data retrieved 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, social network data, web log files, mobile apps data, audio, health wearable device sensors and other (Priyanka & Kulennavar, 2014). These data are very crucial in the IDSS for proper decision-making. For instance, social media networks or in big data called sentiment analysis (Alaoui *et al.*, 2018), help to understand what people are thinking or how they feel about something in healthcare system delivery or disease spread information (Vatrapu *et al.*, 2016). It can be extracted from the social media sites such as Instagram, blog entries, Facebook, Twitter, e-mail exchanges, text messages, search engine indexes, online archives, clickstreams, multimedia files, and photographs (Oussous *et al.*, 2018). Studying child-health phone call data records, online news archives, social media network information, online child-health support inquiries, social media commentary, blog feedback from patients, families, and community members helps to reflect what is in their minds regarding infectious diseases in the community. The approaches can also be used to study other types of healthcare data and in other situations.

The challenges of these types of data cannot be stored in a normal traditional database system because they belong to different formats of datasets. The data typically cannot be analyzed with traditional Structured Query Language (SQL) - tool such as existing HMIS/DHIS-2, rather using non-relational database systems such as NoSQL database tool (Hu *et al.*, 2014). Using big data analytics technology in healthcare such as Text Analytics (Nagwani, 2015), Data Streams Analytics (Chandak, 2016), Social Network Analytics (Alaoui *et al.*, 2018; Mavragani & Ochoa, 2018a; Magumba *et al.*, 2018; Mavragani & Ochoa, 2018b), Machine Learning techniques (Pal *et al.*, 2019; Ed & Maalmi, 2019), Natural Language Processing (NLP) (Osadchiy *et al.*, 2020), Data Mining and Predictive Analytics (Baechle & Agarwal, 2017) effectively help to analyze data from the laboratory, diagnose and medications, drug-resistance patterns, drug interactions and dosing patterns, early warning of disease outbreaks, and fraud detection (Herland *et al.*, 2018). 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 (Torabzadehkashi *et al.*, 2019).

2.6 Big Data Analytics in Healthcare Delivery in Developed Nations

The use of big data sources for disease surveillance in the world has quickly become a controlling source of information on emerging infectious diseases 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. The following are some of the Internet-based infectious disease surveillance systems using big data analytics in developed nations:

2.6.1 Global Public Health Intelligence Network

The Global Public Health Intelligence Network (GPHIN) (Canada Public Health Agency, 1994). Is a system developed in 1997, in partnership with the government of Canada and WHO (Keller *et al.*, 2009). It is an early-warning and situational awareness system for chemical and biological public health threats that adopted big data analytics technology. It uses open-source online news media to continuously scan and extract more than 30 000 news reports to collect suspected information of possible infectious disease outbreaks worldwide. Every day, 20 000 news reports are evaluated for evidence-based decision-making. The information sources include healthcare websites, google alerts, internet searches, RSS feeds, news wires, social networks, expert opinion, validated official alerts, local, national, and international newspapers retrieved in multiple languages (such as English, French, Arabic, Farsi, Russian, Chinese, Spanish, Portugal, etc). The system retrieves relevant articles and online news from news aggregators such as Albawaba - The Middle East Online News Platform (Al-Bawaba Middle East Limited, 2000) and Factiva – The Global News Monitoring and Search Engine (Dow-Jones & Company, 1999) in every 15 minutes, using extensive search queries. It disseminates time-sensitive information to public health professionals in Canada and worldwide for healthcare risk management, prevention, and control measures. In November 2002, the GPHIN helped to inform public health officials on the severe acute respiratory syndrome (SARS) disease outbreak which started in Foshan Municipality, in Guangdong, China, by interpreting news reports on the web and blog sites discussions (Engla & Journal, 2009).

2.6.2 Google Flu Trends Healthcare Big Data analytics

Google Flu Trends Healthcare Big data analytics; this service was conducted by Google to predict and locates flu infectious disease outbreaks by using online information aggregates search engine queries. It was developed essentially to conduct an analytical estimate of the level of weekly influenza activities based on search engine queries received by Google. Big data analytical estimates were derived by Google using a model to compare with the data from the Centers for Disease Control and Prevention (CDC) in the USA. The model produced the best results of a mean correlation of 0.9 with the CDC (Butler, 2013).

2.6.3 Abzooba Smart Health Informatics Program

Naveen *et al.* (2012) created an online big data analytics platform called Abzooba Smart Health Informatics Program (SHIP). Their purpose was to connect a patient with other patients who posted their medical experiences to the internet via online discussion message boards. They used 50 000 message posts from the discussion messages boards extracted from the websites including inspire.com, medhelp.com, and others to extract and execute big data text processing to extract information of each entry including posts and replies related to healthcare involving illness, medicines, treatments, etc.

2.6.4 Search Engine Query Data Retrieved from Baidu (baidu.com)

Yuan *et al.* (2013) also developed a system to collect and analyze health big data using search engine query data retrieved from Baidu (baidu.com) to track ILI epidemics across China. They collected their data from Baidu’s database which stores the online search query since June 2006.

They retrieved data from March 2009 to August 2012, which was during the H1N1 epidemic, and compare their results to that of China's Ministry of Health (MOH). The system involves choosing keywords, filtering, defining weights, and composite search index, and fitting the regression model with the keyword index to the influenza cases data.

2.6.5 Big Data Analytics Technology on Social Media using Twitter Post Data

Signorini *et al.* (2011) employed big data analytics technology on social media using Twitter post data across the United States by searching through specific areas and analyzing the data to predicate weekly ILI levels both across and within the regions. They focus on the period when the H1N1 epidemic was happening in the United States. They collected a large number of Tweets from October 1, 2009 - May 20, 2010, using Twitter's streaming API. The tweets were analyzed by looking for posts containing a preset of keywords correlated to H1N1 (*h1n1*, *flu*, *swine*, *influenza*).

2.6.6 Healthmap

According to Brownstein (2006) Healthmap refers to an early warning alert and disease mapping surveillance system which 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. The system receives 1000-10 000 visits per day (Brownstein *et al.*, 2008). Online health data are automatically collected every hour and classified using free-text analytics data mining to identify kinds of infectious disease and their location. When the health data is processed, the infectious disease outbreak data is visualized on a web-based world map report.

2.6.7 Program for Monitoring Emerging Diseases

According to Morse (1994) Program for Monitoring Emerging Diseases (ProMED) is a Web-based disease surveillance system found in 1994, which continuously gathers and display infectious disease data outbreak worldwide. The system gathers healthcare information from websites, government official health alerts, internet searches, online newspapers, social networks, government reports, and others. The system uses an e-mail system on the internet to disseminate information on infectious disease outbreaks by e-mailing and posting infectious disease case reports. In 2006, the ProMED-mail helped in informing public health officials in Kenya regards cattle die-off online reports in northeastern Kenya which turned into Rift Valley fever infectious disease outbreak.

2.6.8 Web-based Dashboard System for Corona Virus Disease of 2019

Web-based Dashboard system for COVID-19; is an Internet-based system developed by Johns Hopkins University, Baltimore, MD, U.S.A used to track covid-19 confirmed cases in real-time (Center for Systems Science and Engineering [CSSE], 2019). The system uses a live data stream Internet-based which involves local media and government report data to accumulate COVID-19 report cases in near real-time in China. It collects data from the province level through online news, Twitter, and free-text sent through the dashboard. In December 2019, the system was used to detect a local pneumonia disease outbreak in Wuhan (Hubei, China) which initially was the unknown cause, and was later determined to be caused by a coronavirus,1 namely as SARS-CoV-2 (Dong *et al.,* 2020). As of February 17, 2020, the disease outbreak has been spread to every province of China and other 27 countries, with more than 70 000 confirmed cases.

2.7 Big Data Analytics in Healthcare Delivery 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 emergence of the use of the modern health technology systems, 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 healthcare big data which can help to improve CIDSS. Through the preliminary evidence of 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. For instance, mobile phones data was used to detect the Ebola virus disease (EVD) outbreak in Western Africa in 2014 (Chowell *et al.,* 2016). When the EVD was detected through the initiatives of HealthMap, there was no efficient IDSS that detects infectious disease outbreaks at the local geographical areas in Western Africa. The data collection tool which was helpful during the Ebola crisis was mobile phones. During the crisis, healthcare data scientists were collected health data from the mobile phone companies to monitor the movements of the patients suspected of EVD viruses to forecast the spread of the disease. The strength of this study was the ability of the system to collect call-data records and SMS data sources on the time communication was made, size of data transmitted, geographical location, and call duration to track movement between calls for tracking infectious disease dynamics. However, the limitation of this system was the inability to combine structured and unstructured data for sophisticated healthcare data analysis.

Wesolowski *et al.* (2012) used mobile phones data to monitor the movement of malaria parasites by analyzing healthcare data of mobile phones of about 15 million people 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 mobile phones 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.

2.8 Related Works

Many research studies to supplement existing traditional CIDSS 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 in the world ranging from data collection through the web-based system (Cheng *et al.*, 2011; Dong *et al.*, 2020), search engine queries (Chan *et al.*, 2011; Salathe *et al.*, 2012), social network analysis (Hung *et al.*, 2020; Sobowale *et al.*, 2020), online news (Kwok *et al.*, 2021), online forums (Sanders *et al.*, 2020), mobile phones (Li *et al.*, 2020), up to patients movement records (Boulos *et al.*, 2011). In addition to this type of related work, various big data analytics framework studies to facilitate infectious disease surveillance using big data sources and scientific unstructured data analysis have been developed. The following literature reviews studied were extracted on this research:-

Erraguntla (2019) developed big data analytics Framework for Infectious Disease Analysis (FIDA), which supports the integration of structured and unstructured health datasets. The health datasets involve disease incidents, health status, demographics, environmental conditions, and bio-surveillance. The FIDA system supports the collection and analysis of structured data including weather, disease incidents, health status, and demographics data. Also, it supports the collection and analysis of unstructured health-related data including tweets, publications, news articles, really simple syndication feeds, and health websites. The framework involves data collection from various big data sources, modeling and analyzes infectious diseases in a different population, and evaluation of intervention based on resources. The features of the FIDA framework system involve; infectious disease data collection, integration, and analysis. It provides support to infectious disease stages including emergence and transmission. It supports the application of machine-learning techniques, disease dynamics simulation, and modeling. It provides end-to-end infectious disease support for modeling, prediction, analysis, and intervention. The architectural framework of FIDA helps epidemiologists and physicians to break up disease cases into sub-problems and develops solutions with appropriate modeling approaches.

The framework used various case study scenarios to study disease syndromic surveillance in the USA. For instance, in February 2012, FIDA was used to test and validate the Campylobacter infectious disease outbreak caused by contamination of unprocessed milk in Pennsylvania State in the USA. It used social media network data from Twitter collected from 15 January 2012 to detect the Campylobacter disease outbreak. Signs, symptoms, and disease names described by CDC were used to search and collect news feeds and tweets from the Twitter social network. In this study "count of tweets" techniques were used as a metric to detect unusual tweets that related to Campylobacter infectious disease. After performing signs and symptoms analysis, FIDA correctly detected the Campylobacter infectious disease outbreak in Pennsylvania and other states as presented in Fig. 3.

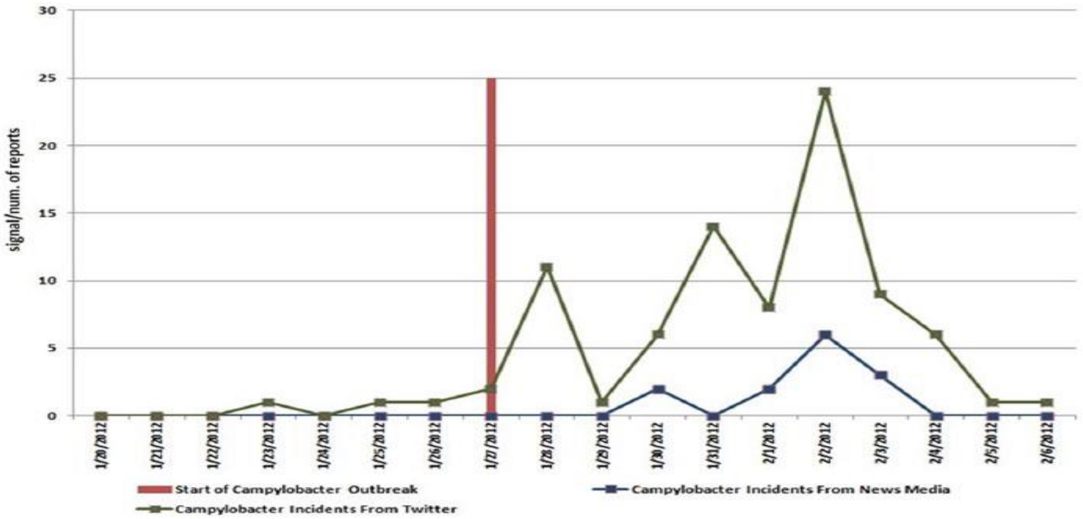


Figure 3: Campylobacter signals from tweets and news feeds detected by framework for infectious disease analysis (Erraguntla, 2019)

In 2014, FIDA was also used to conduct performance measurement using a linear regression prediction model to efficiently capture the transmission nature of ILI illness and predict the seasonal start time of the flu season in the USA. Using Twitter social media, FIDA extracted tweets and news feeds from the HealthTweets online data repository performed by researchers from Johns Hopkins Social Media and Research Group (Dredze *et al.*, 2014). The FIDA used Twitter social media network with linear regression model and other prediction models to conduct infectious disease prediction at the national level.

Jia *et al.* (2020) developed big data analytics framework for infectious disease surveillance to fight against pandemic disease incidents including COVID-19 in China. In their work, they developed big data analytics framework which collects and analyze health data from various big data sources using Internet of Things (IoT), mobile phones, social network, search engines, and genetic data. The framework allows the connection of internet of things technology tools using Radio Frequency Identification (RFID), lesser scanner, and infrared sensors system to collect and analyze health big data in near real-time. Health data involves personal data, user feedback, and status data. Through the use of flights booking information, Internet Protocol Address (IP), mobile phone data, and hotel information, the framework was used to associate bookings and confirmed cases of COVID-19 patients. The framework uses a big data analytics graph to clarify the Covid-19 cases and path of diseases transmission. The graph was used to store health data including personal data, geographical location, and infection time to plot a graph that realizes the propagation path visualization. Also, using NLP techniques, the framework supports performing speech recognition using text classification on reports, free-text files, and news archives.

Wang *et al.* (2018) developed big data analytics framework architecture to describe the capability and potential benefits of big data analytics to healthcare delivery. They explored in detail how a big data analytics framework can help healthcare in data capture, transformation, and visualization. The framework proposed five major layers including data, data aggregation, analytics, information exploration, and data governance. The data layer; divides health data into structured data such as tables, semi-structured data such as logs, and unstructured data such as free-text data. Data aggregation layer; collects health big data and digest them into three stapes; data acquisition, transformation, and storage. Analytics layer; processing and analyzing all kinds of health data using various algorithms such as MapReduce algorithm before submitted to information exploration layer for health data visualization reports. The framework allows the connection of various big data sources including mobile devices, clinical results, network sensors, and social media networks, to collects health big data for infectious disease prevention and control. Also, it offers analytical capability using various big data analytics methods to parallel process large volumes of health data, manipulate real-time data and capture all medical data from the connected systems. The big data analytics framework systems shown above tackle the challenges of traditional IDSS in different ways by underlining different approaches to disease surveillance activities. However, none of them has provided information on Internet-based data acquisition and integration with other streams of a large volume of health data in CIDSS. Also, they do not provide proper ICT infrastructure for resource-limited setting countries for assessing rapidly childhood-infectious disease outbreaks. Most of these framework systems are operated by non-profit organizations that rely on volunteers worldwide who sent data on infectious diseases and processed by volunteers who possessed expertise skills. This approach is inapplicable to the surveillance of CIDSS in a resource-limited setting in countries like Tanzania. The main purpose was to rely on these early big data analytics researchers to advance our study in this novel big data analytics technology to develop BDAF-CIDSS for the Tanzania context.

This study is another additional big data analytics framework for infectious disease surveillance based on modern health technology database systems, Internet-based big data sources, and mobile phone data technology to support infectious disease surveillance activities in the Tanzanian context. Using diarrhea, ARI, pneumonia, hepatitis, and measles as an example, we examine and merge the health big data sources of the health database systems, internet-based and mobile phones data related to infectious diseases to determine if, and to what extent, health big data analytics helps to determine the public health dynamics with compared to the traditional CIDSS activities. This study, proposed a general framework for building BDAF-CIDSS based on health database systems, Internet-based and mobile phones data monitoring for CIDSS, to address the challenges of combining structured and unstructured (free-text) data for sophisticated healthcare data analysis using efficient big data analytics technology tools and methods for the Tanzanian context.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Study Design

To determine the data, methods, and tools that would enhance infectious disease surveillance and user requirements for a big data analytics framework, a survey study with follow-up interviews was used. In assessing the current CIDSS, data on associated system infrastructure, data storage, analysis tools and capacity of system integration, geographical coverage, data quality, and completeness for disease outbreak control measures were collected. Data collection took place between February and May 2019.

After data analysis, the framework was developed and then validated with a selection of the participants from the survey study. The data collected was primarily quantitative, focusing on participants' opinions (on a scale) of the usefulness of the framework. There were no personally identifying information collected and no actual patients' records were viewed. Instead, dummy records were used in use-case scenarios to demonstrate the actual performance of the proposed BDAF-CIDSS in the real world.

3.2 Study Area and Participants

This research study was conducted in four regions Dar es Salaam, Arusha, Kilimanjaro, and Mbeya in Tanzania. These areas have been selected based on the frequent occurrences of the emerging and re-emerging of infectious diseases in the country. The areas have a high level of in-and-out movement of people, social network connections, high interaction of family and friends who play a functional role in the spread of infectious diseases. It involved 6 regional referral hospitals in Tanzania involving Ilala, Temeke, Mwananyamala, Mount Meru, Mawenzi, and Mbeya regional referral hospitals.

3.3 Sampling Design

Purposive sampling was used to select regional referral hospitals based on geographical areas since they are the supervisors of the CIDSS before the analytical data reports are submitted to the Ministry of Health in the country which is the higher authority for decision making. In each hospital, pediatricians, pediatric nurses, medical records officers, and IT personnel were selected by their heads of the department based on the following criteria: must be able to read and write in English (since the questionnaire was in English) and must have 3+ years' experience particularly in childhood-infectious diseases data collection and analysis. The heads of departments were also requested to aim for gender balancing. Up to 6 people were randomly selected from each department in each hospital and the total sample size was 110 participants.

3.4 Data Collection Tools

3.4.1 Questionnaire/Survey Design

The aims of the questionnaires were to:

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

(ii) Identify child-healthcare information gaining mechanisms and decision-making information gaps.

(iii) Identify child-healthcare system opportunities for future CIDSS.

The questionnaire had a total of 31 questions; 28 questions assessed opinions of respondents based on a 5-point Agreement Likert scale of (1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree and 5 = strongly agree) and 3 questions were not for analysis but information purposes.

The first part of the questionnaire assessed opinions on the challenges with existing facilities and tools for childhood infectious disease control. Specifically, participants were asked about the quality of tools used for data collection and analysis, child-healthcare facilities' services regarding infectious disease control, and how many factors such as inability to collect data from rural areas, from patients themselves (patient-generated data), online news, mobile phones data, social networks data, health database systems and the number of healthcare staff affected their ability to monitor childhood-infectious diseases.

The second and third parts of the questionnaire assessed which electronic sources including the Internet, mobile phones, and other organizations' data and which citizen/public data are used to inform on the status of childhood-infectious diseases and how acceptable health facility personnel found the idea of using Internet-based, online health database systems and mobile phone data sources for augmenting surveillance.

The fourth and fifth parts assessed expertise in big data analytics, the existence of big data frameworks, and the acceptability of using big data analytics in CIDSS.

3.4.2 Observation

The observation method was also used to observe the nature of childhood infectious disease data collection, processing, and analysis. The aim was to observe the infrastructures and nature of the systematic flow of infectious disease report cases data, and the time taken from the patient, doctors up to the data storage for analysis and interpretation. During observation, pediatricians' week-ending report forms data and HMIS database system were analyzed to determine the types of data collected. The ICT infrastructure including computers, storage devices, and local area network devices was also assessed to identify performance speed and capacity which is necessary for big data analytics.

3.4.3 Interview

Some of the selected representatives in pediatric, medical records data storage, and IT departments were interviewed to gain further insights into the CIDSS process and its successes, challenges, and the solutions they use to overcome the challenges particularly on collecting and analyzing health big data sources through the Internet.

3.5 Data Analysis

Qualitative and Quantitative statistical methods were used to analyze survey responses. In analyzing quantitative data Microsoft Excel spreadsheet was used. The responses from the survey and interview were analyzed to determine themes (inductive coding). Relative frequency distributions and histogram graph chart were then plotted to determine the most common challenges, solutions, and information-gaining mechanisms

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Results

This chapter presents the results and discussion conducted during the validation experiment from the research study area of the proposed BDAF-CIDSS and presents the output of the validation experiment of the developed framework.

4.1.1 What Data, Methodologies, and Tools can enhance Childhood Infectious Disease Surveillance system in Tanzania

(i) Data Collection and Methods

The study found that the following data collection and methodologies are used in the CIDSS process:

Data Collection

Infectious disease report cases are collected in public hospitals, healthcare centers, and dispensaries through clinical-based healthcare data sources. When a child-patient affected with an infectious disease report and investigated by the medical doctor in a hospital and confirmed the infectious disease case, the Notification of Communicable Disease Report Form is filled by the medical doctor and sent to district hospitals. In the healthcare centers and dispensaries, the doctors prepare weekly return reports using Infectious Disease Week-Ending Report (IDWER) forms which are then submitted through postal services to the district hospitals. At the district hospitals, data is entered into the HMIS/DHIS system at the district and regional level for analysis and decision-making at the ministerial level. The approximate time taken to process, communicate, and information sharing for notification is about 13 days.

Alternatively, the community healthcare activists who are hired by the municipal in the district visit the child patients in their local environment and collect suspected and probable childhood-infectious disease cases using paper-based forms and send them to the hospitals in need. Once the infectious disease is confirmed, it is entered in the Infectious Disease Register for further processing. The interviews on the capability of the existing CIDSS to collect and process infectious disease report cases based on the local geographical areas, internet, or community case findings, they reported that:

> *Using the current system it is very hard to identify or run any query that could help to identify trends and magnitude of the infectious disease report cases from the specific local geographical area. This is because the system was planned to collect and process infectious disease report cases based on what has already been processed at lower levels of the healthcare system (Pediatric Senior Officer).*

For instance, in Temeke hospital, 56 535 data of infectious disease report cases from 2014 – 2018 were collected, but when the Pediatrician and medical data storage staff were asked if it could be possible to access data of the infectious disease from Mbagala or Kongowe area only for all these years, they replied that:

> *We cannot be able to get those data because the system was not set up to collect data from the specific local area but they can get access of those data if they go back and check from Mtuha paper-based documents which are very difficult to get all those data (Medical Data Storage reporting staff).*

The interview on the challenges of collecting and analyzing healthcare data by integrating healthcare datasets from other healthcare-related systems such as health insurance, public pharmacies, websites, social network, etc. The Pediatric officer reported that:

> *Using the DHIS system, we can analyze the disease report cases data collected and entered from the connected lower levels healthcare centers such as dispensaries, healthcare centers, district hospitals, and regional hospitals. The system does not offer the function of integration from other healthcare-related organizations. If we want to access data from other organizations like NHIF or other insurance organizations, the only way is to communicate with their agents working for insurance issues in the hospital*

Findings

The results in our research design show that many hospitals in Tanzania do not collect and analyze data of infectious diseases from big data sources which are outside of the formal structured data system. The results in questions 8-14 of the data analysis show that multiple healthcare data sources including Internet-based healthcare data, insurances, public pharmacies, websites, social networks, online free-text messages, RSS feeds, search engine queries, e-mail, mobile phone data, online communication archives, online government healthcare reports, expert opinion reports, web search queries, and online information aggregates are not included in the healthcare data collection and analysis as indicated in Appendix 2.

The results in questions 1-7 confirmed that the healthcare professionals have greater challenges to collect data on infectious diseases from the patients' environments as indicated in question 5. The majority of responses of more than 70% strongly agreed that inability to collect data from the patient's environment was a greater challenge as indicated on the histogram chart in Fig. 4. This proves that in the traditional CIDSS, multiple healthcare data sources are not included in the decision-making process.

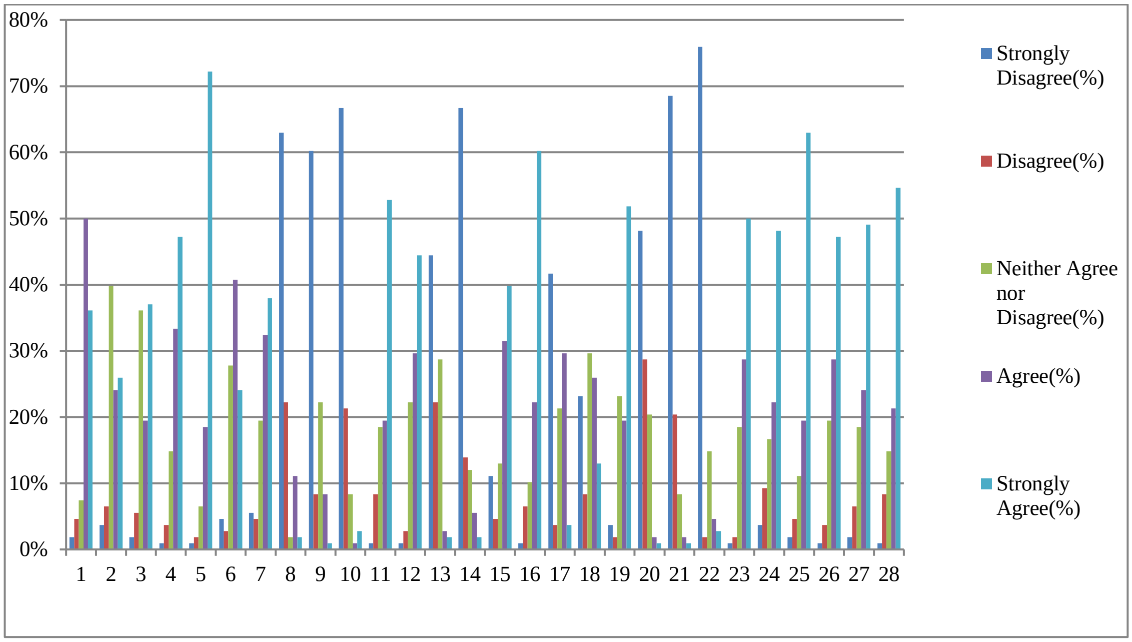


Figure 4: Histogram chart graph of the responses from the respondents

(ii) Information Communication Technology Infrastructure Tools

Data Collection

During observation, it was observed that some healthcare facilities use a fully paper-based system for disease surveillance. Those used Information Communication Technologies (ICTs) infrastructures for infectious disease surveillance mostly had single-processor computers with low memory and hard disk drives as indicated in Table 1. The ICT staff in healthcare centers do not has experience in healthcare big data and data-driven innovation.

Table 1: Information communication technology infrastructure facilities

SN	Hospitals	Computer Type	Processor Type	Random Access Memory (RAM)	Hard Disk Drive (HDD)
1	Health Facility 1	1-HP Desktop Computer	Intel Celeron	8 GB	350 GB
2	Health Facility 2	2-Dell Desktop Computer	Pentium III - Xeon	4 GB	500 GB
3	Health Facility 3	3 – Dell Servers 4– Dell Desktop Computers	Quad Core 2.0 Intel Xeon E -3 1230	4 GB	200 GB
4	Health Facility 4	2-HP Servers	Intel Celeron	4 GB	500 GB
5	Health Facility 5	3 –Dell Desktop Computers	Intel Celeron	4 GB	500 GB
6	Health Facility 5	2–Compaq Server Computer	Pentium III - Xeon	8 GB	200 GB

Researcher (2019)

However, the health facilities were in the process of replacing these technologies with network-attached storage (NAS) and direct-attached storage (DAS) technologies, which have the benefit of high-speed data collection but they are very expensive.

Findings

The results confirmed that 70 (65%), out of 108 from ICT departments respondents strongly agreed that they do not have experience in healthcare big data and data-driven innovation. They do not have experience with high-performance computational ICT infrastructures. 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. 3.

Over 75% (n=81) of the 108 respondents strongly agreed that the collection and analysis of infectious disease report cases from big data 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. 8. The 86(80%) of 108 respondents chosen mobile phone short-text messages as a good source of information. The 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 tool free-text application as the good source of healthcare data as indicated in Fig. 4.

4.1.2 Requirements Analysis

Requirement analysis involved determining the system user requirements for the development of the proposed BDAF-CIDSS in Tanzania. Functional and non-functional requirements were collected and analyzed from the responses of the questionnaires, system observations, and data mining. Functional requirements involve the properties and capabilities of the proposed framework to accommodate user requirements and non-functional requirements involve what the framework will provide to its users.

(i) Functional Requirements

From the requirements analysis, the following functional requirements were established based on the questionnaire responses, system observations, and data mining:

(a) The proposed BDAF-CIDSS should enable users to collect, process and analyze structured and unstructured (free-text) data, online news, e-mail, tables, video through mobile phones, web-based systems, and online data streaming.

(b) The BDAF-CIDSS should enable users to collect process and analyze tables, text, e-mail, and text messages.

(c) The BDAF-CIDSS should enable users to integrate and analyze healthcare data from HMIS/DHIS-2 and other healthcare-related database systems.

(d) The BDAF-CIDSS should enable users to analyze infectious disease report cases, counting the disease report cases events to person, place, and time and generates reports for decision-makers to characterize the level of distribution, spread, and cost of prevention and treatment to suggest determinants of disease transmission.

(ii) Non-Functional Requirements

The following non-functional requirements were also established:

(a) Structured and unstructured data of the patients collected from other sources should have controlled privacy and security. Not every person should be allowed to collect and access patient healthcare data. There must have access control levels in terms of data collection, processing, analysis, and report generation.

(b) Structured and unstructured data of the patients collected from other sources format should be easy to learn such as to identify disease type, person, place and time and use for processing, analysis, and interpretation.

(c) The BDAF-CIDSS platform should be cost-effective and affordable. The system should not involve high-cost infrastructures such as computers or servers with high specifications of Random Access Memory (RAM) and Processors which require a high-cost budget of more than 10 000 USD.

(d) The BDAF-CIDSS platform should be easily maintained, accessible and upgradable. It should be easy to install its software, configure and update. It should be easy to access and command for healthcare data processing and analysis.

(e) The BDAF-CIDSS platform should easily be connected with other healthcare-related data sources such as HMIS/DHIS-2.

4.1.3 Design of a Big Data Analytics Framework for Childhood Infectious Disease Surveillance System for Tanzania

(i) Framework Development

The proposed BDAF-CIDSS which is indicated in Fig. 6, was developed from the analyses and synthesis of the international literature and of experiences of implementing big data analytics technology in analyzing diverse mixes of datasets. To understand the big data analytics and categorization processes based on potential benefits of using big data analytics framework, various big data analytics literature were studied from big data technology benefits, the evolution of data in the enterprises, and connection of data across the organization. In the application of big data technology benefits, Big Data for Dummies book published by Hurwitz *et al.* (2013) was explored. The information extracted in big data technology including Hadoop MapReduce, Hadoop Distributed File System (HDFS), and Hadoop Ecosystem tools in big data management helped us to develop the big data analytics framework. The knowledge on integration of big data with the traditional database systems and how the extraction, transformation, and loading of health data changing the roles of data warehousing helped us to establish the relationship between the components in the developed framework. The evolution of data in the enterprises and connection of data across the organization from Simon (2013) helped to categorize the proposed components of the framework. Additionally, the knowledge on how to modify health business intelligence data to handle big data including data analytics algorithm and infrastructure support, structured and unstructured free-text data, streaming data and complex events data processing analytics using various approaches and techniques helped us to strengthen the operational benefits of the proposed framework. Big Data Analytics Framework for Healthcare System developed by Wang *et al.* (2018) was considered as a reference for the development of the framework. Wang framework is a dynamic big data framework for the healthcare system which is considered as a healthcare big data framework base for the general healthcare system which supports tracking and monitoring of infectious disease. It has incorporated the general known healthcare big data analytics variables approaches. Their objectives were to generalize the healthcare big data analytics approaches based on the fundamental variables. This approach is inefficient to the process of collection, processing, and analyzing the health database system, Internet-based and mobile phones healthcare data sources to be adopted in the Tanzania environment.

In this study, a framework was developed 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. Based on our research design, the development of our proposed BDAF-CIDSS framework was focused on the need to access and analyze additional information from areas not covered by the existing disease surveillance system. Therefore, the task of developing the framework was divided into the following stages:

Data Collection

During this research study, it was noted that in the traditional system, data collection is done through clinics and hospitals. In order to focus on areas that are not covered by the traditional system, the task was divided based on big data sources as follows:

- **Mobile Device Applications**

Mobile device applications involve the collection of health data through mobile phones, wearable devices, and sensors. Using this technology, it will be possible to track children’s movements through their parents to understand the path of infectious disease transmission. It involves understanding the suspected people who have been contacted with the parents and children. This will help to identify children and parents who have already been infected or contacted with the infected people before isolated and treated them in advance.

- **Social Media Networks**

Social media networks including Twitter, Facebook, WhatsApp, Online Websites, and others, can help to collect real-time free-text data for disease prevention. The use of social media networks allows large groups of people to create and share health information, experiences, medications, drug side effects, and health condition status online. The collection and analysis of the social network data collected in real-time will help physicians to gain a deeper understanding of the time and geographical location of childhood infectious disease transmission. For instance, counting the number of parents' health-related tweets on Twitter social media, specific-disease hashtag links, or daily visits to social media articles like Wikipedia articles related to infectious diseases will help physicians to predict and forecast childhood infectious disease transmission. Therefore, this study proposed social media to become one of the additional health data sources in the development of the new framework.

- **Electronic databases**

Electronic database systems involve all secure websites and online health systems. Traditional surveillance system uses an electronic database of HMIS / DHIS-2 to collect and analyze health data for infectious diseases. Standard disease specific-electronic database systems help to share infectious disease data worldwide. In the development of our framework, we proposed the use of additional electronic database systems including traditional health database systems such as (HMIS/DHIS-2), pharmacy databases, R&D institution databases, online magazines, donors online systems, immunization campaign online system, health insurance systems such as NHIF, Jubilee Insurance, Platinum health insurance, and other external health database systems such as PubMed, Medline, Github repository among others to improve traditional surveillance system. Since these health databases systems are very crucial in infectious disease surveillance, then electronic database systems was proposed to become one of the additional health data sources in the development of the new framework.

Data Acquisition

During this research, it was discovered that infectious disease data acquisitions are collected using HMIS and DHIS-2 systems. Few cases declared using mobile phones for data collection and analysis. The acquisition of these data varies because the traditional system uses numerical data structure for analysis. In addition to free-text data, emails, and other unstructured data, big data analytics technology using text-format analysis tools was considered for data storage and analysis. The application of data extraction, transformation, and load (ETL) big data tools using a NoSQL database engine was proposed for implementation in the framework.

Analysis

In the traditional system, the analysis of structured numerical data is always done using Relational Structured Query Language (R-SQL) analysis using Relational Database Management System (RDBMS) such as Microsoft Excel, Access, and few cases using MySQL database engine. Data in a numerical format including tables can be efficiently analyzed in the traditional system. In big data analytics, multiple health data with different data formats from different channels are involved. This data involves multiple unstructured free-text data, emails, logs, audio, video, event data streams, online transactions, search queries, among others. Therefore, the application of Hadoop big data analytics technology with MapReduce algorithm, and Map Reduce-Based Query Language (MRQL) together with NoSQL database engine tool was recommended for implementation in the framework.

Interpretation and Visualization

In the traditional system, health data interpretation and visualization are done using portable graphs plotted in extensible markup language (XML) and web visual analysis technology format. Using this technology it is possible to uncover relationships within few structured health data. This is very difficult when health data interpretation and visualization involves massive health datasets with mixed formats including geographical location data from Geographical Information System (GIS). Therefore, the application of big data analytics techniques using the MapReduce algorithm was recommended. The technology can help to process, analyze and interpret digital data using batch and distributed health data processing that can be transformed to a format acceptable with the GIS for efficient data visualization and information exploration.

The proposed BDAF-CIDSS involved interdependencies of the physical structure and dynamics of the healthcare system to healthcare delivery. It involved the following features; data capture, data acquisition, data analytics, and information exploration layers as indicated in Fig.5.

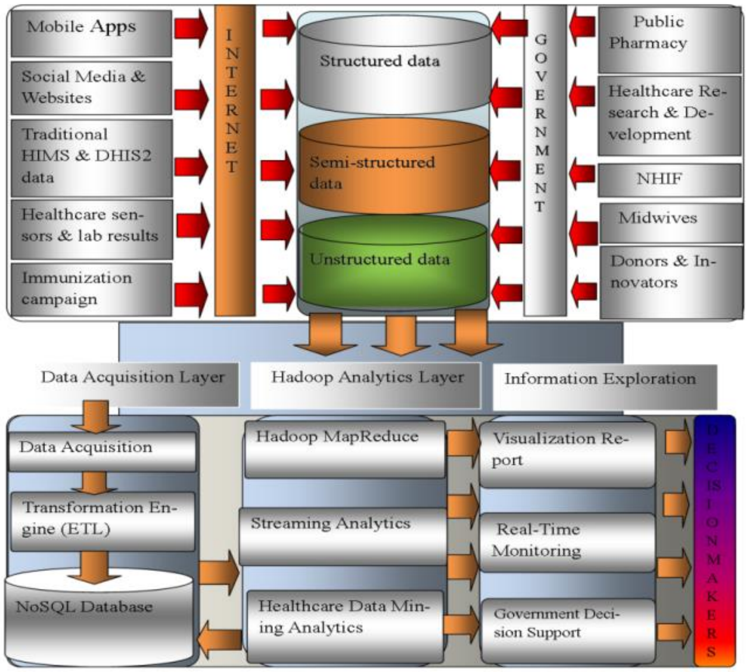


Figure 5: Proposed Big Data Analytics Framework for Childhood Infectious Disease Surveillance System for Tanzania for Tanzania environment

(ii) 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 sources such as HMIS/DHIS-2 system, health insurances, online healthcare information archives, network sensors, and public pharmacy. These clinical data can be collected through tables, csv files, json data files, and text file format and stored in relational databases depending on the content format such as MySql, Oracle, PostgreSQL, and others. Unstructured data such as text, e-mails, text messages, online text archives, and free-text messages data can be collected using various tools such as sqoop, flume, and Web scraping scripts. Since unstructured data cannot be processed using structured databases (Al-barhamtoshy & Eassa, 2014), then data will be stored in non-traditional databases which can handle unstructured data such as MongoDB Databases. The data will then transfer into the next layer which is the data acquisition layer for extraction and transformation.

This layer was accommodated in the simulation using various tools; Online Disease Case Definition Forms which used to capture details of the patient data (Patient Name, Address, Date Of Birth, Age, sex, location coordinates, laboratory results, case classification, hospitalized option, name of the notifier and telephone number of the notifier.) through mobile phones. These datasets were collected and submitted to the designated central database server for storage. Web

Crawler Script tool; was used to collects web-based healthcare news articles and free-text health data from the Internet-based system and stored in MySQL and MongoDB Databases engine. Hadoop Cloudera Eco System tool; was used to support the collection and analysis of structured and unstructured health big data such as tables, free-text, online archives, and search queries.

(iii) 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 and unstructured databases can be transformed into HDFS format for processing in a big data analytics platform. Since the incoming data from the above layer comes from various data sources, their characteristics are varying in terms of a communication channel, frequency, size, volume, and file format. Therefore, this layer is used in the extraction of data from original databases, transformed them into HDFS format, and load data into big data analytics tools. 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 datasets 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 is also done in unstructured data whereby data in the format of e-mail, weblogs, or text can be extracted and transformed into key-value pairs that can be read with the big data analytics tools. The transformed health data will then be transferred into the next layer which is the data analytics layer. In this research study, we accommodated this layer by installing the Hadoop Cloudera ecosystem tool. Sqoop and hive tools were used to transfer data from the RDBMS database of MySQL and transform them into HDFS.

(iv) Data Analytics Layer

The data analytics layer is the main layer for healthcare data processing and analysis. It accommodates three types of health big data processing and analysis: Hadoop Map-Reduce data processing, Data streaming, and In-Database Analytics. It 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 Map phase is the raw or unstructured (free-text) data, which is processed by split up into key-value pairs. And the output from the *Map function* is processed by the Map-Reduce framework before being sent to the *Reduce function*. Map Reduce processing provides the ability to process a large volume of structured and unstructured healthcare data in batch processing and massively parallel processing (Barkhordari & Niamanesh, 2018). Datastream processing help to process and analyze real-time and near real-time stream data. In real-time stream data processing, it could be possible to track healthcare data-in-motion such as rates of infectious disease spread, prediction of infectious disease outbreaks, respond to unexpected infectious disease outbreaks and quickly make an evidence-based decision for early infectious disease notification, prevention, and control. In-database analytics, we can perform a data mining approach using data mining techniques such as machine learning using various algorithms such as Naïve Bayes Classifier, K-means algorithm, Clustering, and Logistic regressions to parallelize processing and analyze large scale healthcare datasets of the infectious diseases (Imai *et al.*, 2015). In this research study, this layer was accommodated by developing modified Map-Reduce algorithm program design models for simulation.

Map-Reduce Algorithm Program; is a distributed data processing algorithm introduced by Google (Kijsanayothin *et al.*, 2019). It is mainly developed and implemented using a functional programming model. Map Reduce algorithm data processing use MRQL to support all sub-queries such as create tables, joins, group-by, union, and load a large volume of health data in big data technology platform. It works by breaking structured or unstructured data into two phases: Map phase and Reduce phase processing. Each phase has key-value pairs as input and output (Liu *et al.*, 2015). The input to our Map phase is the free-text messages or unstructured data, which is processed by split up into key-value pairs. And the output from the *Map function* is processed by the Map-Reduce framework before being sent to the *Reduce function*. The output information is the one transferred into the last layer which is the information exploration layer ready for health data visualization.

For instance: k and v stand for key and value pairs respectively in HDFS file system, then, in Map Reduce Algorithm Programming Model processing:

$\text{map}(k1,v1) \rightarrow \text{list}(k2,v2)$
$\text{reduce}(k2, \text{list}(v2)) \rightarrow \text{list}(k3,v3)$

(v) Information Exploration Layer

The information exploration layer aimed to provide output results based on data visualization and real-time monitoring reports. Infectious disease prevention and control visualization reports and real-time data streaming monitoring were used to perform active disease surveillance activities in day-to-day operations. This feature helped healthcare professionals to make early evidence decisions such as infectious disease alerts, warnings, and notifications to the citizens before infectious disease outbreaks.

4.1.4 Framework Use Case Diagram

Big data analytics framework can be implemented in the real world by setup the Hadoop Cluster using commodity computers cluster and configuring MasterNode with DataNodes to facilitate health big data processing. It can be implemented at the district level, regional level, up to the national level to enhance infectious disease surveillance. Framework use case diagram helps to represent how user's interaction with the actual performance of the proposed BDAF-CIDSS in real life can be conducted. The diagram helps to provide a simplified real picture to the stakeholders of how the framework can be implemented in real-world use cases to comply with the user requirements as indicated in Fig. 6. Based on the framework use case diagram, the diagram can be divided into four areas: (a) Healthcare big data sources (b) Data ingestion zone (c) Big data analytics zone, and (d) Big data application zone.

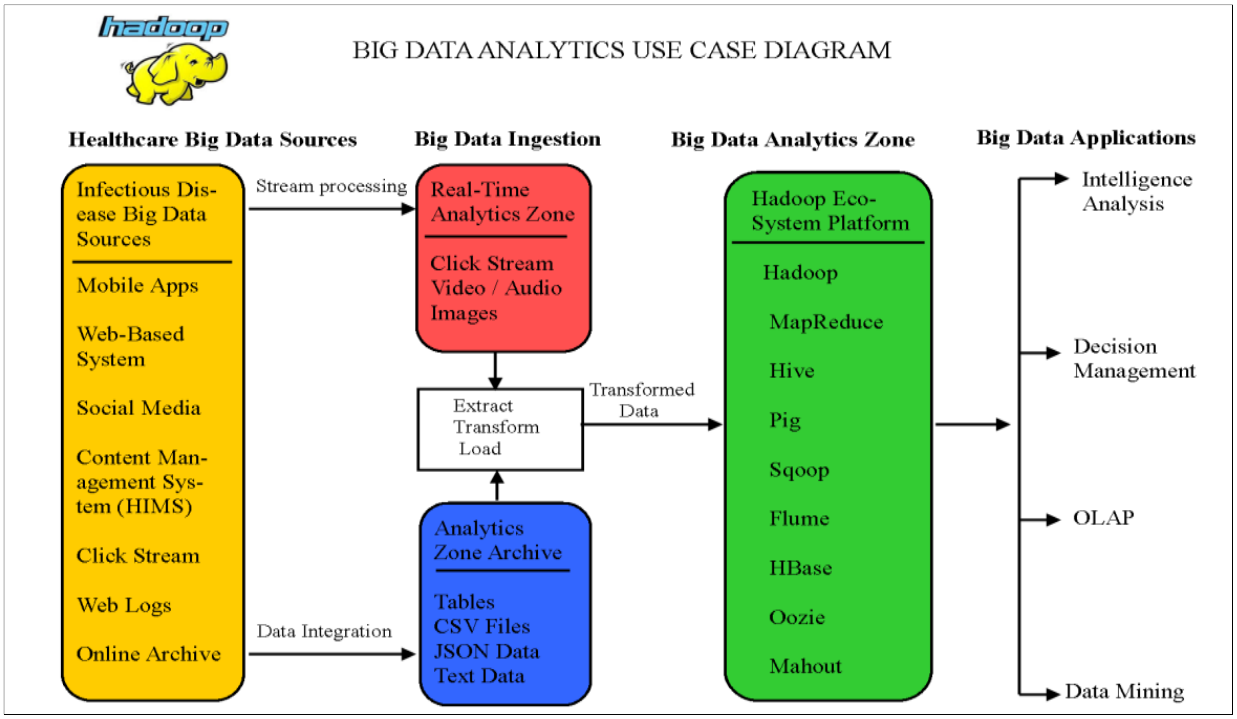


Figure 6: Use Case Diagram for the implementation of the Big Data Analytics Framework for Childhood Infectious Disease Surveillance System for Tanzania

(i) Healthcare Big Data Sources

Public health big data sources involve roles of data collection from the various healthcare data sources. It involves the collection of infectious diseases data from patients using various tools including mobile applications, 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 HMIS/DHIS-2, IDWE, and IDSR systems. Also, it has community healthcare activists who collect and submit data to the hospitals at the healthcare centers levels. These categories of users will be improved by assigned activities of collecting infectious disease data using digitized data systems through mobile applications and web-based systems instead of manual paper-based systems. Initially, doctors, IT personnel, laboratory scientists, and medical records personnel in the hospitals can help to collect infectious diseases data from multiple sources including pharmacies, social media, clickstreams, weblogs, and online archives.

(ii) Data Ingestion Zone

The data ingestion zone involves data integration and streaming processes. The IT personnel, doctors, medical records personnel, and laboratory scientists can help to conduct this process. It involves running queries and commands for extraction, transformation, and loading infectious disease data from the Internet-based data sources into the Hadoop platform. Structured databases such as tables, csv files, and json data from pharmacies, NHIF, immunization campaign systems, network sensors, and unstructured online free-text healthcare data streaming sources can be collected and integrated at this stage. The basic knowledge of R-SQL among healthcare professionals can help them to process these data using Hive tool which using MRQL that of HiveQL language similar to R-SQL language. Therefore, IT personnel will help to run these codes, queries, and commands with the directives from the healthcare professionals and doctors.

(iii) Big Data Analytics Zone

The big data analytics zone involves running healthcare data analytics using health big data sources. It involves executing public health data jobs using Hadoop MRQL 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.

(iv) 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. It involves data standardization, master data governance, policies, standards, 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 through its lifecycle from data archives to support business goals of preventing and control infectious diseases. These activities can be performed by the higher authority and decision-makers in healthcare organizations.

4.1.5 Data Flow Diagram

A data flow diagram is a graphical representation of the data flow in a proposed framework. It describes a process that involved a framework to transfer infectious diseases data from the input to the generation of the final report as presented in Fig. 7.

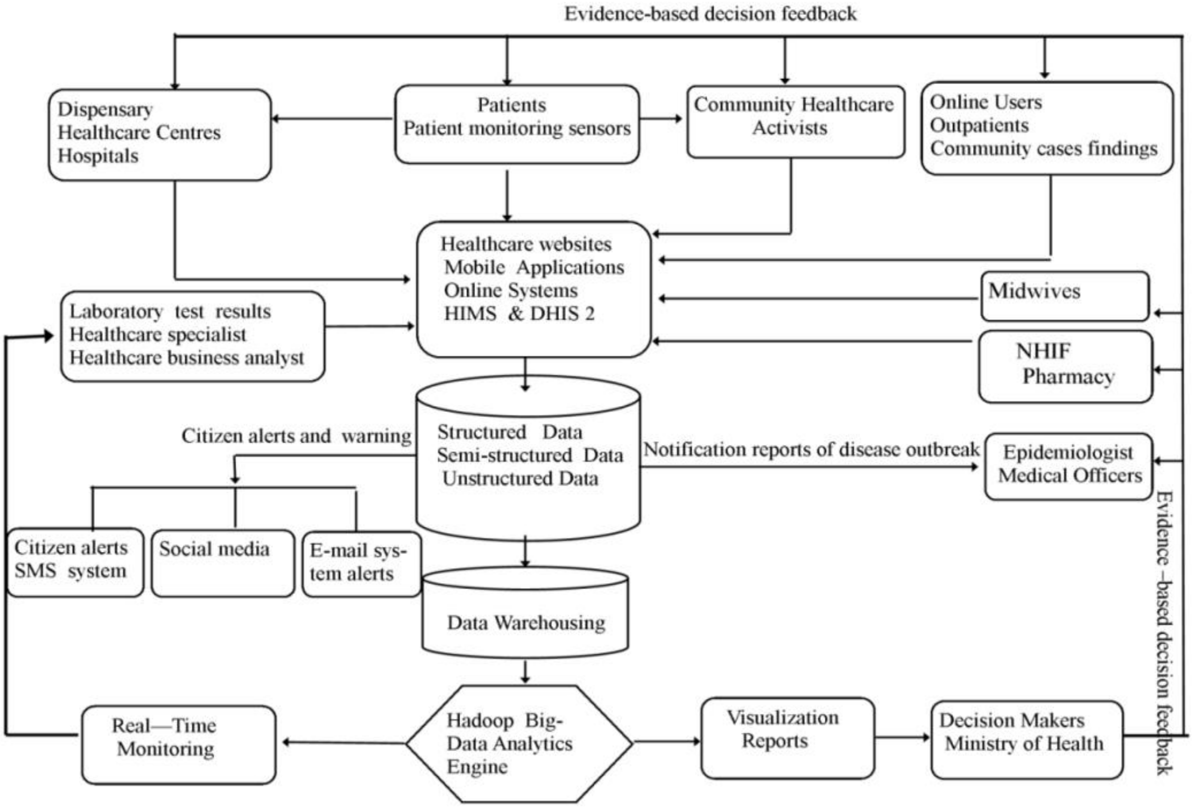


Figure 7: Data flow diagram for the proposed big data analytics framework for childhood infectious disease surveillance system for Tanzania

4.1.6 Framework Validation Approach

The proposed BDAF-CIDSS was validated using Open Data Kit (ODK) data collector for mobile data collection, WebCrawler script program, MySQL, MongoDB databases, and Hadoop Cloudera ecosystem for big data analytics technology software were used to test the implementation of the framework.

(i) Open Data Kit-Data Collector

The ODK data collector is community open-source mobile software for collecting, managing, and using data with the support of geo-locations, numerical and text data, images, audio, and video clips in resource-limited settings environments. The reason why ODK data collector tool was opted for data collection was that in resource-limited settings environments or countries like Tanzania, data collection through mobile applications has limited settings in message length and submission of geo-location added to the records but with ODK data collector these services can be extended to accommodate users. Also, digital data collection with ODK data collector has been supported by many world humanitarian network organizations including WHO in Nigeria, the CDC, USAID, and Red Cross organizations (Bokonda *et al.*, 2020). In addition, the ODK data collector allows the collection of digital data without an internet connection and submits the data when internet connectivity is available.

(ii) Web Crawler Program

The Web Crawler is a computer program which used to search and fetch relevant information from the Web. It starts at a root page and then following the hyperlinks within it to explore other pages on the same website until all pages are fetched. A Web crawler is using by scientists to search important information from the Internet search engines using specified domain indexes. Full-text document of priority infectious diseases information keywords was collected using web scraping tool presented in Appendix 8, which then analyzed using Map-Reduce algorithm in Hadoop Cloudera Express platform.

(iii) Hadoop Cloudera Express Manager

Hadoop Cloudera is an Apache open-source software developed by Google. It is a framework written in Java that allows distributed processing of a large volume of data across commodity computers cluster using a simple programming model called Message Passing Interface (MPI). Hadoop is a popular choice when you need to filter, sort, pre-process and process large amounts of healthcare data. Hadoop runs applications using HDFS and Map Reduce algorithms where data is processed parallel with others. By configuring a cluster with MasterNode, NameNode, DataNode, SecondaryNameNode, JobTracker, and Task Tracker, Hadoop can process huge amounts of health big data in parallel. It is developed to perform statistical analysis on huge amounts of data and designed to scale up from a single commodity computer or server to thousands of interconnected computer machines working in parallel as indicated in Fig. 8.

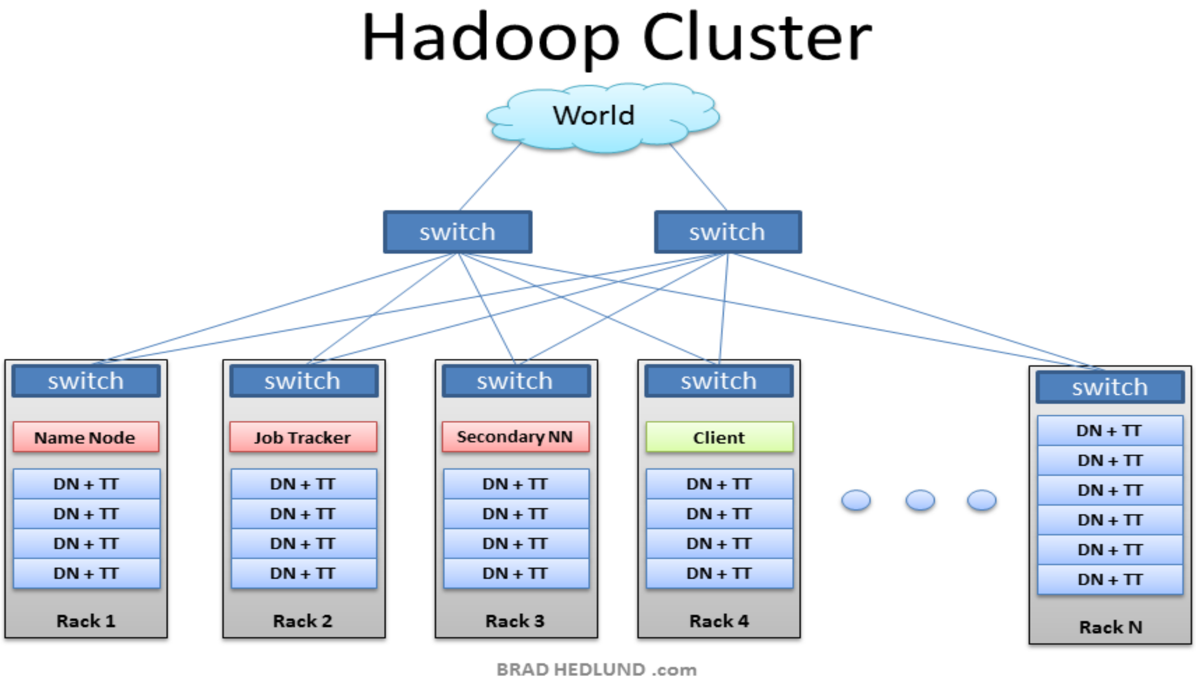


Figure 8: Hadoop cluster system architecture (Brad Hedlund.com)

(iv) Map-Reduce Algorithm in Hadoop Cluster

To take advantage of using Hadoop technology for parallel processing in healthcare big data, it needs to translate our R-SQL healthcare data query into an MRQL job. After some processing, these data can be processed into cluster machines. Map Reduce in Hadoop works by breaking data processing into two phases: Map phase and Reduce phase. Each phase has key-value pairs as input and output. The input to 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 Map-Reduce framework before being sent to the *Reduce function* as presented in Fig. 9. This processing sorts and groups the key-value pairs by key for final output.

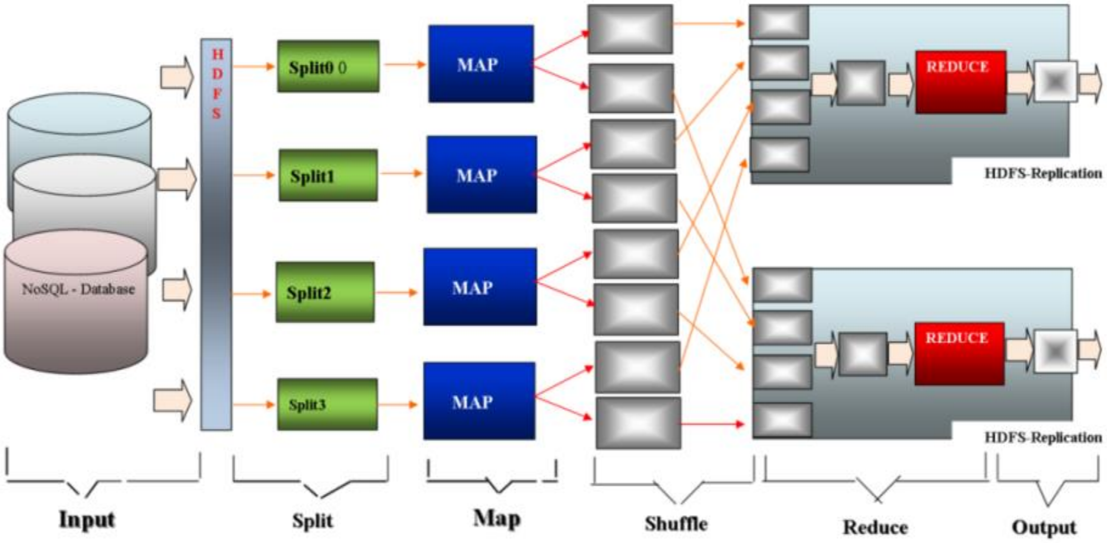


Figure 9: Map-reduce algorithm data flow diagram in hadoop platform

Hadoop runs Jobs by dividing healthcare data jobs into tasks, of which there are two types: *Map tasks* and *reduce tasks.* The tasks are scheduled using Hadoop Yarn and run on the compute nodes on the cluster. If a task fails, it will be automatically rescheduled to run on different compute nodes. Hadoop divides the input data to a Map-Reduce job into fixed-size pieces called *input splits* which run the user-defined map function for each record in the split. Having many splits means the time taken to process each split is small compared to the time to process the whole input of the infectious disease data.

4.1.7 Proposed System Design Architecture

The proposed BDAF-CIDSS system design architecture helps to guide healthcare professionals to implement the proposed BDAF-CIDSS framework using available cost-effective system infrastructures to collect and analyze infectious disease report cases from multiple channels of structured and unstructured datasets in the real world. The system design architecture helps the physicians to plan, select, and implements infrastructures based on the key organizational structure of the proposed BDAF-CIDSS framework. It involves four layers including; data capture layer which involves mobile applications, social media networks, electronic databases systems, and Web application servers. The data acquisition layer involves the extraction, transformation, and load engine functions which allow data integration from the multiple health data formats such as xml, csv, PDF, json, table, and text. Also, it allows the function of real-time data streaming from social media and clicks streaming. Data analytics layer involves Hadoop Cluster system which configured with MapReduce algorithm system that consists of MasterNode, NameNode, DataNode and HDFS file system with NoSQL database engine. The data exploration layer involves a health data visualization tool that supports big data interpretation and visualization as presented in Fig. 10.

This heterogeneous framework design architecture consists of the following integrated systems smart mobile application, Web-based system, Web scripting program, and Hadoop big data analytics technology. Infectious diseases data can be collected using Android or other mobile applications through the Infectious disease case definition forms and Web scripting program for unstructured data. The forms and spider scraping program were prepared to demonstrate the data collection process in the framework. The forms used to collect structured data were filled and uploaded into the ODK Data Collector which was then stored in the MySQL Database. The Spider scraping program which used to collect unstructured data such as free-text data stored in the MongoDB database. The PyCham integrated development environment (IDE) was used to run Scrapy Python program scripts. Hadoop Cloudera express was set up to pull healthcare data from the remote MySQL and MongoDB database using Apache Sqoop which converted structured data, semi-structured data, and text-format data into HDFS format for further processing in the Hadoop Cloudera platform. Apache Sqoop is a Hadoop ecosystem tool that transfers data between the Hadoop platform and RDBMS servers such as MySQL, Oracle, and others. It is used to import and transform data from RDBMS into the HDFS, and vice versa. Sqoop tool function is very crucial in this framework as it helps to import infectious disease data from other healthcare-related databases such as HMIS/DHIS-2 system, health insurance system, public pharmacies, websites, social networks, network sensors systems, and others. The data is then stored in the NoSql Hadoop data warehousing server for healthcare decision-making analysis. At this stage, different data analysis algorithms like Map Reduce, Naïve Bayes Classifier, K-Mean, or Logistic Regression, can be deployed to conduct big data analytics such as disease prediction analytics.

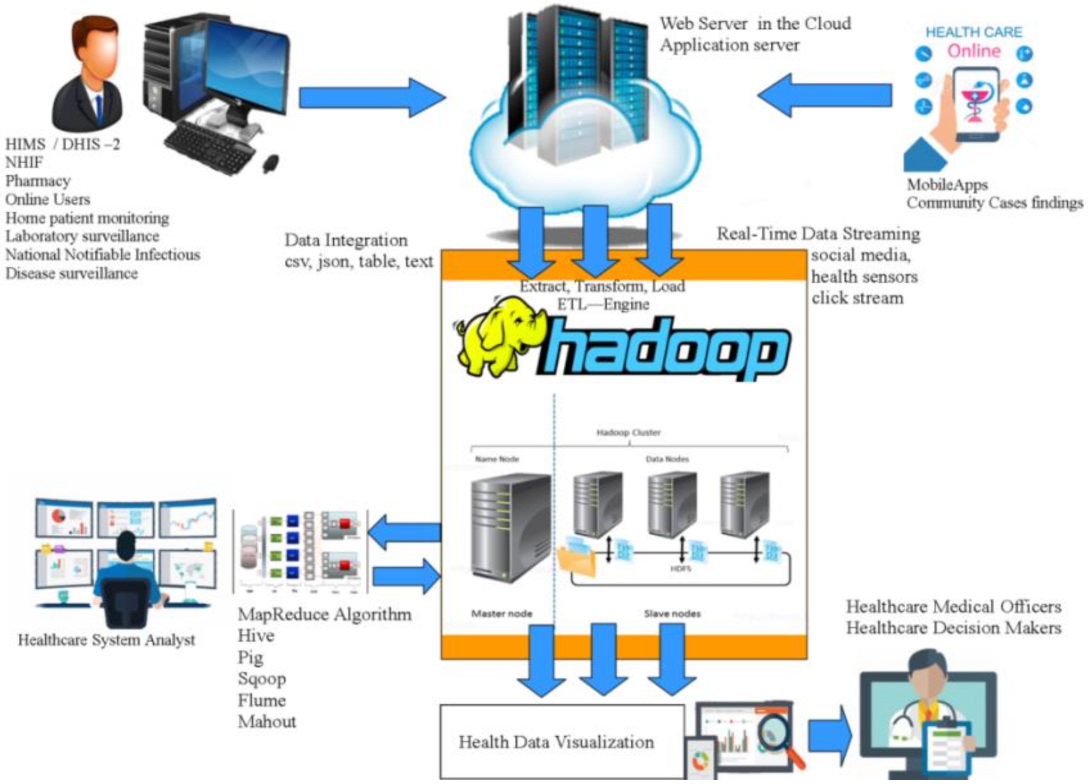


Figure 10: Proposed big data analytics framework for childhood infectious disease surveillance system design architecture

4.1.8 System Settings Approach

In this research design, main concentration was the collection and analysis of structured and unstructured healthcare datasets which is the key activity in any infectious disease surveillance system, and run big data analytics using Map-Reduce algorithm for evidence-based decision. To achieve these objectives the following activities were conducted:

(i) Online Disease Case Definition Form

The online infectious disease case definition form was customized using a free online open data kit build system at opendatakit.org as presented in Appendix 5. The form was then uploaded and configured into the ODK data collector mobile application as indicated in Appendix 6 to facilitate community case findings data collection simulation. The system provides functions to configure server type, server URL, username, and password. It is also possible to configure the user interface, form management, and user device identity to control the system users.

(ii) Web Crawler Program

A Web Crawler scrappy program was developed and installed using Python Programming language to collect online free-text unstructured data and full-text document data from the Internet. The scraping spider script program contained an item, MongoDB pipeline, and spider python script to extract health data from the Internet as presented in Appendix 8. The program was set up to monitor relevant and useful infectious disease online news sources from the Internet. The free-text healthcare data was extracted from the Web-based on a tag name, title, and an attribute of that tag in a format of dictionary and stored them into the MongoDB database. Various online healthcare news database sources including Google Scholars, Daily News online magazine, and others were used to gather free-text documents for demonstration purposes. The aim was to use the framework to conduct healthcare information mined from the Internet-based healthcare data sources to identify news articles that contain medical-significance-related information of the key infectious diseases. The web crawling keywords were pneumonia, hepatitis, measles, diarrhea, and ARI.

(iii) MongoDB Database

The MongoDB database was set up for simulation to store and retrieve data from various online systems. The online infectious disease free-text healthcare data based on a tag name, title, and attribute in a format of the dictionary was collected and stored into the MongoDB database for big data analytics.

(iv) Hadoop Cloudera Express

The commodity computers cluster was installed and configured for virtualization functions. The cluster was virtualized using a VMware workstation. The CentOS operating system was installed in the computers cluster. Apache Hadoop Cloudera Express as a big data analytics ecosystem was installed to support big data-intensive distributed applications. It consisted of 2 Nodes, one was designed as MasterNode with the following services NameNode and JobTracker. Another one was acted as DataNode with Task Tracker services. The HDFS file system was configured in all compute nodes. Map Reduce as a core function of Hadoop technology for data writing in parallel processing on the Hadoop cluster was configured. In this study, apache Hadoop cluster consisted of a single MasterNode and one slave Task Tracker node and the following services HBase, HDFS, Hive, Impala, Hue, Oozie, Spark, Yarn, Sqoop, Pig, and Flume.

Hadoop Cloudera Express Setup

Hadoop Cloudera Express was installed on Windows PC to perform healthcare big data analytics operations. The section below provides details of the installation setup:

- **Virtual Machine Installation**

    (i) Step1: Use commodity computers, check for prerequisites on the computer system minimum of 8 GB (at least) of RAM is required for Hadoop Cloudera, but 16 GB is highly recommended.

    (ii) Step 2: Download the open-source virtual machine VMware software player or workstation (EMC Corporation, 2004)

    (iii) Step 3: Download Hadoop Cloudera Express (Cloudera Inc, 2008) quick start virtual machine

    (iv) Step 4: Install virtual machine in windows PC and set up its parameters

    (v) Step 5: Install Hadoop Cloudera Express quick start virtual machine

- **Load Healthcare data in Hadoop Cloudera Express**

To load healthcare data into Hadoop Cloudera Express, two option steps can be followed:

(i) Option 1: Loading data from the local machine. The following commands were used to load healthcare data from the local machine into Hadoop Cloudera express HDFS:

```
$> hdfs dfs –copyFromLocal /home/...     user/hadoop/cloudera/
```

(ii) Option 2: Loading data from the remote/ local database server: Sqoop tool in Hadoop Cloudera express was used to load data into HDFS from the remote database server.

Sqoop is a Hadoop ecosystem tool designed to transfer data between the Hadoop platform and relational database servers such as MySQL, MySQL, Oracle, and others. It is used to import and transform data from RDBMS into the HDFS, and vice versa. Sqoop tool function is very crucial in this framework as it helps to import infectious disease data from other healthcare-related databases such as NHIF database system, public pharmacies, websites, social networks, network sensors systems, and others.

In order sqoop tool to import data from RDBMS to HDFS and vice versa, the sqoop import and export commands must be executed as follows:

Commands:

```
$ sqoop import --connect jdbc: mysql: //..ip_address/database_name  --table_name   -- username
--password     ---target_dir
```

(v) Healthcare Data Acquisition for Validation

Since healthcare big data requires data for many years (2010 - 2018) which was hard to collect because of the confidentiality and integrity, the healthcare datasets were generated using a mockaroo (Mockaroo, 2002). Infectious disease case definition online form in Appendix 5, was developed and used for dummy data collection. The healthcare data was generated and made realistic as possible to support the validation testing and simulation. The 2GB healthcare data sets were generated which contains 10 thousand data of patient records collected over 8 years from 2010 – 2019.

Also, with the growth of the Internet, users are often facing difficulties to search for priority relevant information on the Web due to information overload. A typical real-world Internet-free text dataset contains a large number of noisy, redundant, and uninterested concept features. To avoid this problem, a web crawler scrappy program was used to generate web-based healthcare data. The priority keyword features of pneumonia, hepatitis, measles, diarrhea, and ARI were used to search online healthcare text document data to improve the accuracy of data collection.

4.1.9 Validation Experiment Study

In this framework validation experiment, different infectious disease data analytics use cases scenarios have been tested to perform healthcare data analytics to support evidence-based decisions. The following use cases scenario were validated:-

(i) Use Case-Scenario I

The use-case scenario I which is integrating healthcare datasets from various healthcare-related data sources has been demonstrated. Integrating healthcare datasets from various multiple data sources and analyzing them all together for evidence-based decision-making was one of the great challenges facing the traditional system. In the traditional system, it was very difficult to integrate health datasets from other health-related data sources. Currently, in the traditional system, each organization has its independent system of tracking infectious disease report cases. Hospitals conduct infectious disease surveillance using a paper-based system, NHIF also has its way of tracking costs incurred by the patients on treating the infectious diseases, local pharmacies also have their way of tracking drugs provided to the patients suffering from infectious diseases symptoms. These systems cannot communicate with each other even though they deal with the same common function of prevention and control of infectious diseases. These independent systems hinder the effective performance of the traditional disease surveillance system.

In this research design, we developed the Map-Reduce algorithm to enable healthcare professionals to integrate healthcare datasets from various healthcare-related data sources such as NHIF, pharmacy, home patient monitoring systems and integrated with data from the hospitals. The aim was to track patient against the repetition of drug use for the same infectious disease which in turn develop drug resistance adverse effects. The purpose was to identify how many times a particular patient has visited hospitals and the amount spent on the same infectious disease case and the type of drugs provided. This simulation helps healthcare professionals improve drug risk and cost implication management for evidence-based decisions. Drug resistance adverse effects may catalyze infectious disease transmission from one person to another in the local environments as well as increase the cost burden to the government.

In this first use case scenario, the Map-Reduce algorithm indicated in Appendix 9 which is presented on the reduce-side joins design model, and which involves joins of multiple datasets from different healthcare data sources was customized. The multiple structured data was developed from a hospital and National Health Insurance (NHIF) datasets and develop Map Reduce algorithm for integrating multiple datasets using the reduce-side joins Map-Reduce algorithm. The MapReduce algorithm and description for the implementation of the system was described as follows:

Map-Reduce algorithm design model

The Reduce-side joins MapReduce algorithm design model was developed as follows:

```
  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 treated
                              2.  Increase counter by 1
                              3.  Accumulate amount spent, then
                              4.  Get Total Amount.
   // Else,
          Store variable for future assignment;
   End Task:
```

The following data field format was used to generate dummy data in json format from (Mockaroo LLC, 2002): NHIF Datasets: Patient_id, Patient_Name, Drug_description, Amount, Disease_case_description and the Hospital datasets json format involved: Patient_id, First_Name, Last_Name, Address, Gender, Age, Disease_case_diagnosed as indicated on hospital dummy data sample generated in Mockaroo.com in Appendix 7.

In this design, two input files of healthcare datasets were developed from hospital and NHIF datasets input files. The Reduce-Side Joins Map-Reduce algorithm program was developed using

Java programming language, then loaded and transformed them into HDFS file system for processing. In this Map-Reduce algorithm processing, the following functions were observed:

The Mapper and Reducer functions for hospital and NHIF database details are as follows:

- **The Mapper Function Phase**

In the Mapper function phase:

(i) The system read the two input files by taking one tuple at a time

(ii) Then, the system tokenized each word in that tuple and fetches the Patient ID along with the name of the patient, and fetches the amount value instead of the name in NHIF details.

(iii) The patient ID is the key of the key-value pair that our Mapper would generate repeatedly from hospital details.

(iv) We added tags "hosp" to indicate the input tuple of the patient from the hospital and "nhif" to indicate the input tuple of the patient from the NHIF details respectively.

(v) Our Mapper from the hospital details produced the following key-value pairs: Key – Value pair [Patient_ID, hosp name]: example: [1, hosp John], [2, hosp Rose].etc

(vi) The Patient_ID would be our key of the key-value pair that our Mapper would generate repeatedly from hospital and NHIF details database tables.

(vii) The output for Mapper from the NHIF details would be the following key-value pairs: Key –Value pair [Patient_ID, nhif name]: example: [1, nhif John], [2, nhif Rose].etc

- **Sorting and Shuffling Phase**

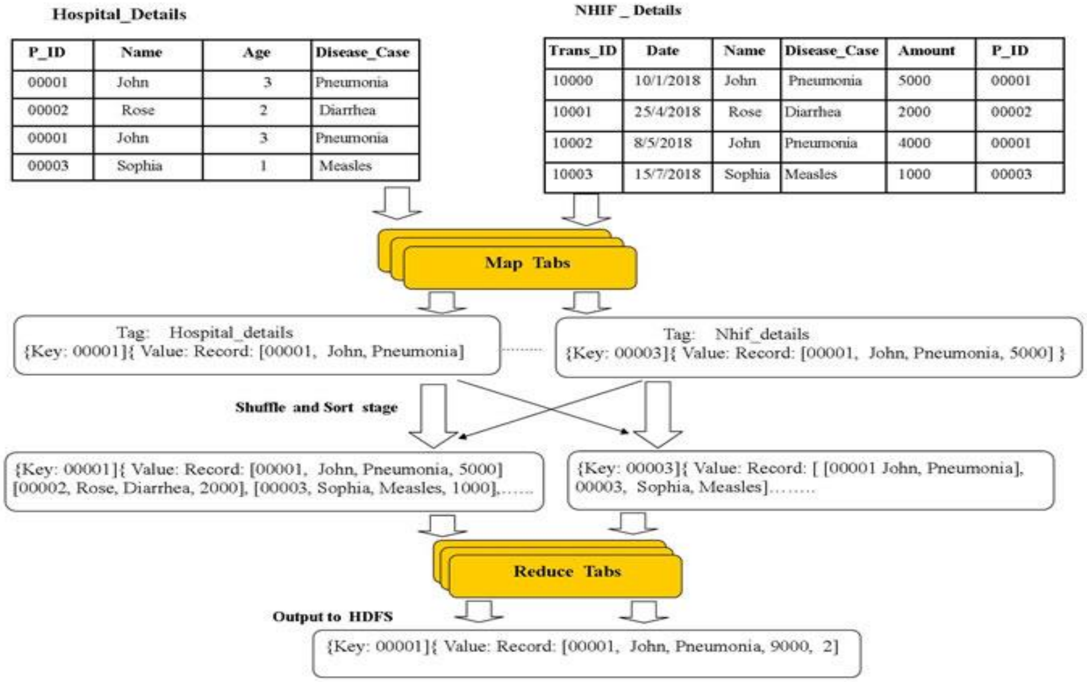
The sorting and shuffling phase would generate an array list of values corresponding to each key and the output would be the following:

Key – list of values: {Patient_ID1 – [(hosp name1), (nhif amount1), (nhif amount2), (nhif amount3)….]}...

Now, the Map-Reduce algorithm would call reduce() method { reduce(Text key, Iterable <Text> values, Context context)} for each unique join key (Patient_ID) and the corresponding list of nhif values. Then the reducer would perform join operations to present the list of values for output.

- **Reducer Function Phase**

In each of the Reducer, there should have a key and a list of values where the key belongs to Patient_ID and the list of values would have the amount from NHIF and hospital details. Then these lists of values were looped through the Reducer phase to check if the value belongs to hospital details or NHIF details. If it was NHIF details, the counter value was increased by one to calculate the frequency of visits by the patient and accumulate the amount to calculate the total amount spent by the particular patient in a particular infectious disease. If the value belongs to the hospital details table the values were stored in a string variable for future assignment. The whole Reduce-Side Joins Map-Reduce algorithm processing was conducted in a series of job processing as indicated in Fig.11. Finally, the reducer generated the output which was stored in the HDFS file system for visualization by the healthcare decision-makers.



Hospital_Details

P_ID	Name	Age	Disease_Case
00001	John	3	Pneumonia
00002	Rose	2	Diarrhea
00001	John	3	Pneumonia
00003	Sophia	1	Measles

NHIF _ Details

Trans_ID	Date	Name	Disease_Case	Amount	P_ID
10000	10/1/2018	John	Pneumonia	5000	00001
10001	25/4/2018	Rose	Diarrhea	2000	00002
10002	8/5/2018	John	Pneumonia	4000	00001
10003	15/7/2018	Sophia	Measles	1000	00003

Map Tabs
Tag: Hospital_details
{Key: 00001]{ Value: Record: [00001, John, Pneumonia]
Tag: Nhif_details
{Key: 00003]{ Value: Record: [00001, John, Pneumonia, 5000] }
Shuffle and Sort stage
{Key: 00001]{ Value: Record: [00001, John, Pneumonia, 5000]
[00002, Rose, Diarrhea, 2000], [00003, Sophia, Measles, 1000],......
{Key: 00003]{ Value: Record: [[00001 John, Pneumonia],
00003, Sophia, Measles]........
Reduce Tabs
Output to HDFS
{Key: 00001]{ Value: Record: [00001, John, Pneumonia, 9000, 2]

Figure 11: Map-Reduce Algorithm data processing

The implementation of the MapReduce algorithm in the system

To run the above Map-Reduce algorithm program, the following commands were executed:

(i) Change directory into the Hadoop workspace directory:

```
$ >  cd  /home/cloudera/workspace
```

(ii) List files available in the HDFS directory

```
workspace$ >  hdfs dfs –ls
```

(iii) Execute Map Reduce program

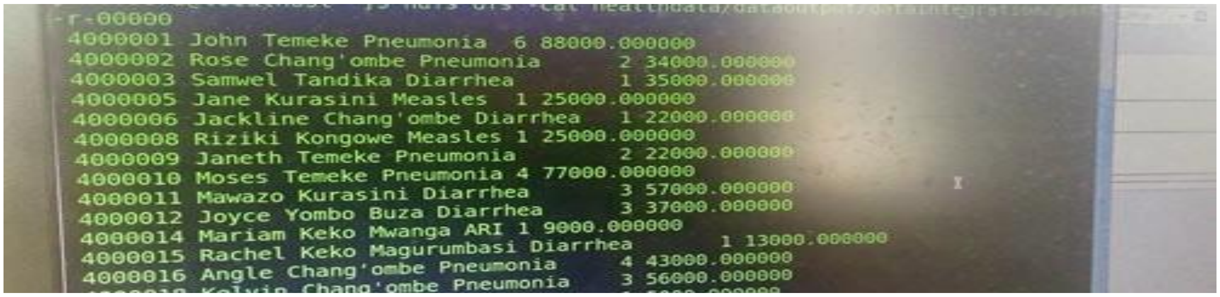
workspace$> hadoop jar MapReduce.jar DriverClass input   output

(iv) Access the results/ output:

workspace$> hdfs dfs –cat HDFS directory/output/part-00000

Or, can also be executed by clicking the following directory user/cloudera/…../output/part-00000 in Hadoop Cloudera express home page as indicated on Fig. 12.

Output results:



```
-r-00000
 4000001 John Temeke Pneumonia  6 88000.000000
 4000002 Rose Chang'ombe Pneumonia      2 34000.000000
 4000003 Samwel Tandika Diarrhea        1 35000.000000
 4000005 Jane Kurasini Measles  1 25000.000000
 4000006 Jackline Chang'ombe Diarrhea   1 22000.000000
 4000008 Riziki Kongowe Measles 1 25000.000000
 4000009 Janeth Temeke Pneumonia        2 22000.000000
 4000010 Moses Temeke Pneumonia 4 77000.000000
 4000011 Mawazo Kurasini Diarrhea       3 57000.000000
 4000012 Joyce Yombo Buza Diarrhea      3 37000.000000
 4000014 Mariam Keko Mwanga ARI 1 9000.000000
 4000015 Rachel Keko Magurumbasi Diarrhea       1 13000.000000
 4000016 Angle Chang'ombe Pneumonia     4 43000.000000
         Kelvin Chang'ombe Pneumonia    3 56000.000000
```

Figure 12: Output results part-00000 file from the Hadoop Cloudera System

(ii) Use Case-Scenario II

The use case scenario II which is counting and identifying the number of infectious disease report cases per local geographic area from 2010 – 2019 has been demonstrated. One of the great functions of an infectious disease surveillance system is to identify the number of infectious disease report cases at the local geographical areas to identify trends of the disease for early warning, alerts, quick response, and government intervention. This function is currently done in the traditional system using a paper-based system to count the number of infectious disease report cases using paper-based documents. This is very difficult to identify trends of the disease for many previous years example cases from 2010 – 2019.

In this validation testing, we demonstrated the capability of the big data analytics technology to count the number of infectious disease report cases based on local geographical areas for many previous years using the better and effective technique. To count the number of infectious diseases to identify trends of the disease, the Map-Reduce algorithm indicated in Appendix 10 was developed and implemented as presented on the word count algorithm design model for simulation.

The MapReduce algorithm and description for the implementation of the system was described as follows:

MapReduce Algorithm Design Model

The InfectiousDiseaseReportCases WordCount algorithm design model was developed as follows:

```
     Procedure: Count Number Of InfectiousDiseaseReportCases WordCount
     Input:  InfectiousDiseaseReportCases 2010 – 2019 datasets
     Output: Number Of Counts of InfectiousDiseaseReportCases for Each Local 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), DiseaseReportCase1) (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 for each Area_ID, Area_Name, Disease_Name, key to sum up the DiseaseReportCases count
   //If the key has more value:
```

```
                1. Count number of DiseaseReportCase
                2. Increase counter by 1
                3. Accumulate the number of DiseaseReportCases, then
                4. Get Area_Name, Disease_Name, and Number of
                   DiseaseReportCases.
    // Else,
            Store variable for future assignment;
    End Task:
```

The implementation of the MapReduce algorithm in the system

In this simulation, the following steps were followed:

(i) Create Map-Reduce program using eclipse which is installed together in Hadoop Cloudera Express and create Map Reduce algorithm program file.

(ii) Create MapReduce program using Java programming language

(iii) Convert the MapReduce program into a jar file and export it into the workspace directory.

(iv) Check the file directory in HDFS:

$> hdfs dfs –ls

(v) Create folder DiseaseCasesCount:

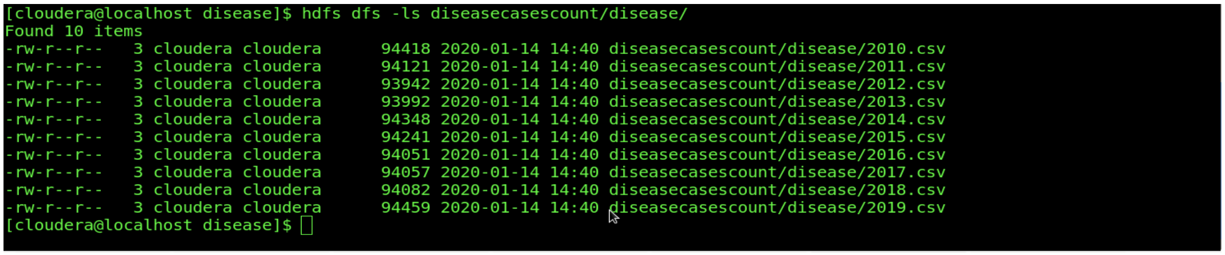
$> hdfs dfs –mkdir diseasecasescount

(vi) Copy disease report cases files from the local file system (2010 – 2019.csv files) to HDFS:

*$> hdfs dfs –copyFromLocal /home/cloudera/Documents/diseasecases/**

(vii) Verify the presence of all source files in the HDFS file system as indicated in Fig. 13:

$> hdfs dfs –ls diseasecasescount/



```
[cloudera@localhost disease]$ hdfs dfs -ls diseasecasescount/disease/
Found 10 items
-rw-r--r--   3 cloudera cloudera      94418 2020-01-14 14:40 diseasecasescount/disease/2010.csv
-rw-r--r--   3 cloudera cloudera      94121 2020-01-14 14:40 diseasecasescount/disease/2011.csv
-rw-r--r--   3 cloudera cloudera      93942 2020-01-14 14:40 diseasecasescount/disease/2012.csv
-rw-r--r--   3 cloudera cloudera      93992 2020-01-14 14:40 diseasecasescount/disease/2013.csv
-rw-r--r--   3 cloudera cloudera      94348 2020-01-14 14:40 diseasecasescount/disease/2014.csv
-rw-r--r--   3 cloudera cloudera      94241 2020-01-14 14:40 diseasecasescount/disease/2015.csv
-rw-r--r--   3 cloudera cloudera      94051 2020-01-14 14:40 diseasecasescount/disease/2016.csv
-rw-r--r--   3 cloudera cloudera      94057 2020-01-14 14:40 diseasecasescount/disease/2017.csv
-rw-r--r--   3 cloudera cloudera      94082 2020-01-14 14:40 diseasecasescount/disease/2018.csv
-rw-r--r--   3 cloudera cloudera      94459 2020-01-14 14:40 diseasecasescount/disease/2019.csv
[cloudera@localhost disease]$ 
```

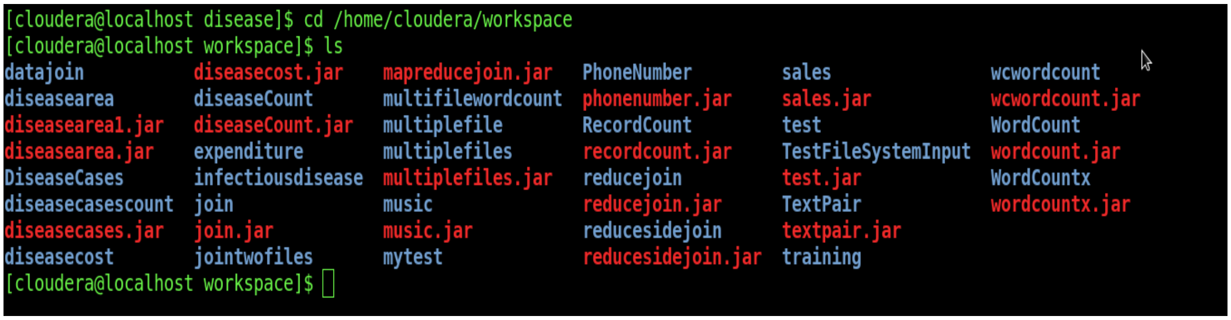
Figure 13: Infectious disease report case csv files in Hadoop platform

(viii) Go to workspace folder and run Hadoop jar class files for Map Reduce algorithm:

$> cd /home/cloudera/workspace

(ix) Check existing jar files for running our created Map Reduce program in the workspace directory as indicated in Fig.14:

Workspace$> hdfs dfs –ls



```
[cloudera@localhost disease]$ cd /home/cloudera/workspace
[cloudera@localhost workspace]$ ls
datajoin           diseasecost.jar    mapreducejoin.jar   PhoneNumber        sales              wcwordcount
diseasearea        diseaseCount       multifilewordcount  phonenumber.jar    sales.jar          wcwordcount.jar
diseasearea1.jar   diseaseCount.jar   multiplefile        RecordCount        test               WordCount
diseasearea.jar    expenditure        multiplefiles       recordcount.jar    TestFileSystemInput wordcount.jar
DiseaseCases       infectiousdisease  multiplefiles.jar   reducejoin         test.jar           WordCountx
diseasecasescount  join               music               reducejoin.jar     TextPair           wordcountx.jar
diseasecases.jar   join.jar           music.jar           reducesidejoin     textpair.jar
diseasecost        jointwofiles       mytest              reducesidejoin.jar training
[cloudera@localhost workspace]$
```

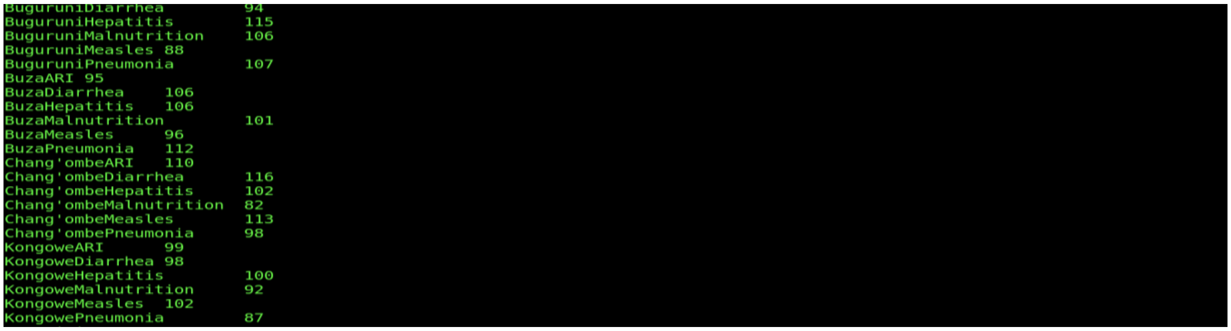
Figure 14: Hadoop Jar files running Map-Reduce algorithm in Hadoop platform

(x) Run the Map-Reduce program using jar file to process disease report cases files:

Workspace$>hadoop jar diseasearea.jar DiseaseCasesArea diseasecasescount diseasecasescount/output

(xi) Check for the analytic results indicated in Fig. 15. From the created part-00000 file using commands:

workspace$> hdfs dfs –cat diseasecasescount/output/part-0000



```
BuguruniDiarrhea        94
BuguruniHepatitis       115
BuguruniMalnutrition    106
BuguruniMeasles 88
BuguruniPneumonia       107
BuzaARI 95
BuzaDiarrhea    106
BuzaHepatitis   106
BuzaMalnutrition        101
BuzaMeasles     96
BuzaPneumonia   112
Chang'ombeARI   110
Chang'ombeDiarrhea      116
Chang'ombeHepatitis     102
Chang'ombeMalnutrition  82
Chang'ombeMeasles       113
Chang'ombePneumonia     98
KongoweARI      99
KongoweDiarrhea 98
KongoweHepatitis        100
KongoweMalnutrition     92
KongoweMeasles  102
KongowePneumonia        87
```

Figure 15: Healthcare big data analytic result processed by Map-Reduce algorithm

(iii) Use Case-Scenario III

The use case scenario III which is the collection and analyzing infectious disease big data from the online news archives has also been demonstrated. Another great challenge of the traditional system is to collect and analyze online infectious disease datasets from online websites, social media, and online healthcare news archives. The function of using online information aggregates search is currently not done in the traditional system which hinders data coverage and completeness on the evidence-based decision making.

In this validation testing, we demonstrated the capability of the system to collect and analyze healthcare online news archives datasets. In this simulation, we collected 12 text document files using a web crawler script program from the healthcare news archives from the internet including Google scholars. The goal was to use the framework to conduct healthcare information mined from the Internet to identify news articles that contain medical significance-related information of the key infectious diseases. The keywords for distributed cache were pneumonia, hepatitis, measles, diarrhea, and acute respiratory infection. In our simulation, the Map-Reduce algorithm indicated in Appendix 11 was developed and implemented as presented on distributed cache Map-Reduce algorithm design model. The MapReduce algorithm and description for the implementation of the system was described as follows:

Distributed Cache MapReduce Algorithm Design Model

The Distributed Cache Map-Reduce algorithm design model was developed as follows:

```
Procedure: Distributed Cache Map Reduce Algorithm
Input:  12 – files of Text documents with 1 – Keyword file datasets (contains: Pneumonia, Hepatitis, Measles, Diarrhea, ARI, and Malnutrition Keywords)
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
                     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:
```

The implementation of the MapReduce algorithm in the system

In the implementation of the MapReduce algorithm in the system the following steps were followed:

(i) Create MapReduce algorithm program using Java programming language

(ii) Convert the MapReduce program into a jar file and export it into the Hadoop workspace directory.

(iii) Check the file directory in HDFS:

```
$> hdfs dfs –ls
```

(iv) Create folder Words Matching:

$> hdfs dfs –mkdir wordsmatching

(v) Copy the online healthcare news archive files from the local or remote file system or internet (12- text files) to HDFS:

$>hdfs dfs –copyFromLocal /home/cloudera/Documents/wordsmatching/*

(vi) Verify the presence of all source files in the HDFS file system:

$> hdfs dfs –ls wordsmatching/

(vii) Go to workspace folder and run Hadoop jar class files for Map Reduce algorithm:

$> cd /home/cloudera/workspace

(viii) Check existing jar files for running our Map Reduce program in the workspace directory:

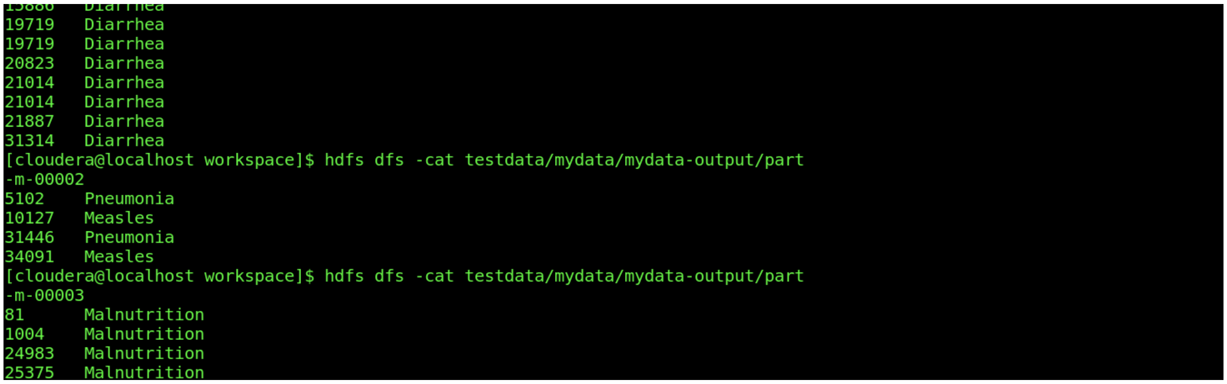
Workspace$> hdfs dfs –ls

(ix) Run the Map Reduce program using jar file to process our online healthcare news archive files:

Workspace$> hadoop jar keywords.jar DiseaseCasesArea diseasecasescount diseasecasescount/output

(x) Check the analytic results from the created part-0000 file using commands as indicated in Fig.16:

workspace$> hdfs dfs –cat wordsmatching/output/part-0000



```
15886   Diarrhea
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
-m-00003
81      Malnutrition
1004    Malnutrition
24983   Malnutrition
25375   Malnutrition
```

Figure 16: Infectious disease keywords appearance from the online news archive files

In these results, out of 12 collected online healthcare news archives, 8 news articles were found to contain significance-related information on various diseases including diarrhea, pneumonia, measles, and hepatitis. This will help healthcare professionals make a decision on which article should be critically read to collect and analyze significant information on specific infectious diseases.

4.1.10 Framework Evaluation

To evaluate the performance of the proposed BDAF-CIDSS, healthcare stakeholders and users were also involved. The evaluation was planned and conducted to acquire users' perspectives and satisfaction, and also to verify if the proposed model is complying with users and technology requirements.

The evaluation testing was conducted to address the following questions:

(a) Is the developed framework complying with the users' requirements?

(b) Is the developed framework complying with the available big data technology?

(c) Does the proposed framework align with the users' requirements?

(d) Does the proposed framework accommodate all healthcare functions and other related healthcare stakeholders?

To address the above questions, we planned and conducted two types of evaluation methods:

(i) Performance-Based Comparison Evaluation

In this evaluation, we went through a synthetic study of the big data analytics technology based on the Map-Reduce algorithm to present technical perspectives and their advantages. Then the conciseness of the traditional healthcare data analytics system were compared based on the RDBMS. In this comparison analysis evaluation, we concentrated on the MRQL against the traditional R-SQL. In big data analytics, MRQL is a fundamental syntax of structured query language which is similar to the traditional R-SQL language used to run queries in big data analytics technology. It supports all sub-queries such as create tables, joins, group-by, union, and load large-scale healthcare data in big data analytics. It involves languages such as Hive Query Language (HiveQL), JSON Query Language (JAQL), Pig Latin, and others. In the evaluation much concentration was on HiveQL to evaluate its performance on large scale data analysis compared to traditional R-SQL on the following metrics:

Increasing Input Data Size and Running-Time Execution

In this evaluation experiment, we performed healthcare data input measurement using MRQL and R-SQL using a fixed size computer cluster of one Name Node and two Data Nodes. In Use Case Scenario II, we used infectious disease report cases per local geographic areas WordCount program from 2010-2019 for comparison. The healthcare datasets were doubled for each experiment and conduct a Join benchmark to get the results of processing time between the two systems as indicated in Fig. 17.

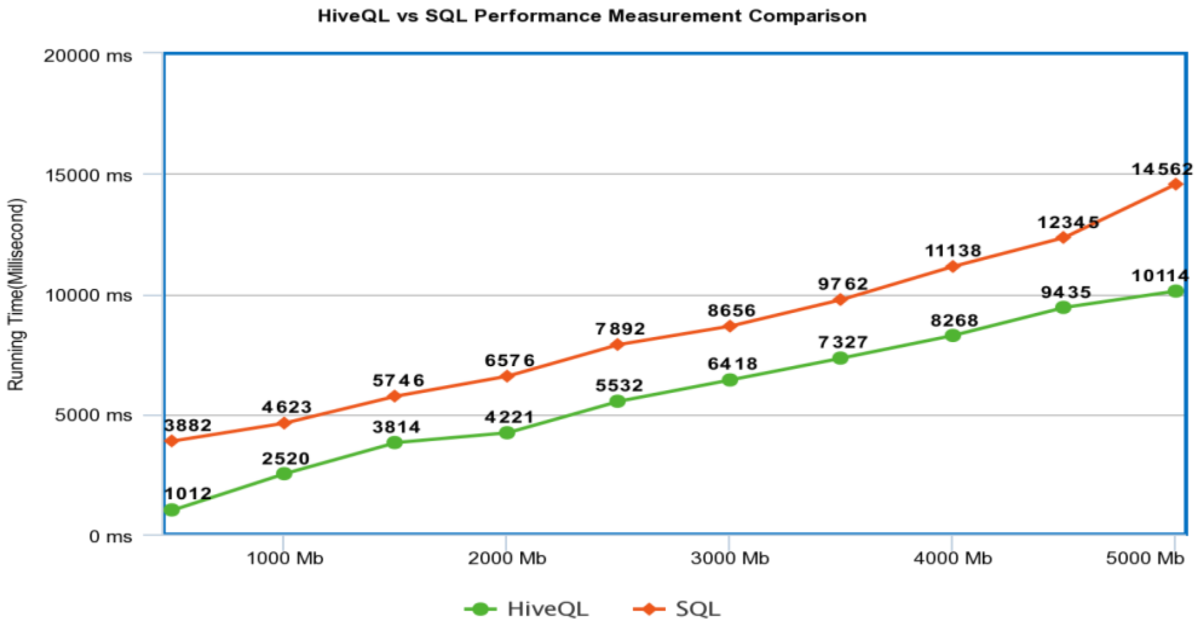


Figure 17: HiveQL vs R-SQL Performance measurement comparison

In this experiment, it was observed that the running time between these two technologies increases as the input size of the healthcare datasets was increased. The traditional R-SQL language system processing has the highest running time and the MRQL has the lowest running time of about 56% quicker than the traditional SQL language system. This can be seen from our experiment results in Fig. 18.

Increasing Processing Units

In this experiment, much concentration was placed on the processing units by increasing the number of compute nodes. The size of input datasets is fixed which is all infectious disease report cases per local geographic areas from 2010 – 2019. The result of this experiment is presented in Fig. 18.

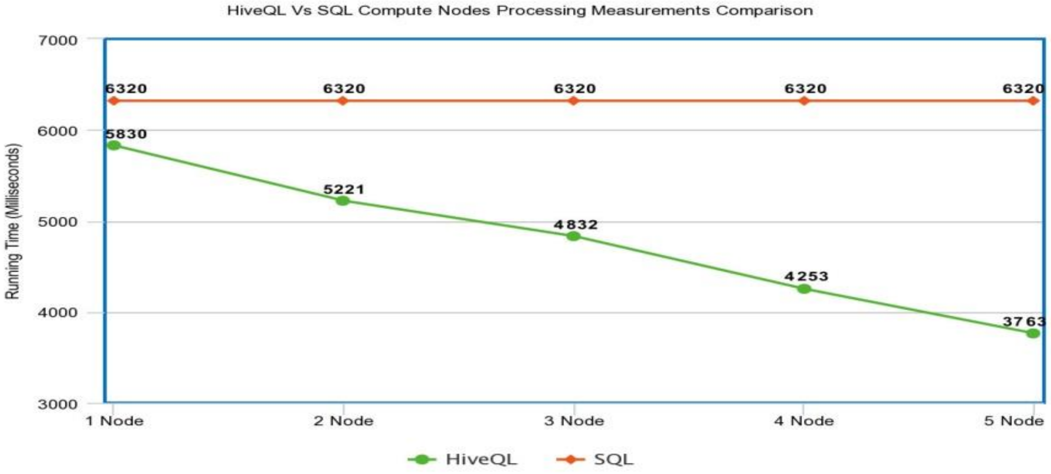


Figure 18: HiveQL vs R-SQL Compute nodes processing measurement comparison

In this experiment, it was observed that the traditional R-SQL has the challenge of using additional compute nodes for processing. Because R-SQL is an intensive memory processing and cannot be applicable for multi-processing in large-scale data operations. It is not possible to transmit large-scale healthcare data using the R-SQL language system into the available memories of the processing units. While the MRQL can break the large task of data and distributes them to massively parallel processing and query available processing unity for execution. It provides supports for creating tables and load command which reads and moves massively healthcare data from relational database and data stored locally into HDFS and distributes them into the available memories of the processing units. This experiment proves that the additional number of processing units has no benefit to the R-SQL language system which achieves no changes in execution performance. This is quite different from the MRQL system which achieves more changes on execution by decreasing running time because of the fastest performance as the number of processing units was increased.

Healthcare Data Transformation

In this experiment, it was examined how the system offers opportunity to transform and extract useful healthcare information from unstructured healthcare data. In Use Case Scenario III, 12 online text information archives were extracted from the healthcare websites and processing them using the R-SQL language system and MRQL user-defined program.

In this experiment, it was observed that the MRQL in big data analytics offers more support for data processing. It offers an opportunity to write user-defined programs or run queries using other languages such as HiveQL, Pig Latin, and others and executed on top of the native Map-Reduce algorithm program to produce the expected results from the unstructured healthcare datasets. But, the traditional R-SQL Language system has no support in processing unstructured healthcare datasets.

Fault tolerance

In this evaluation, it was examined how two systems behave when a fault occurs. The two systems were switched and start executing the jobs and made interruptions in the middle of the execution.

In this experiment, it was observed that in a traditional R-SQL language system once the interruption occurs the system stop processing. This means you need to restart the system to continue with the execution of the jobs. The system does not support fault tolerance. This is quite different from the MRQL system whereby when the fault occurs in one DataNode the system continues with its execution in other compute nodes and provides the final aggregates of the expected results.

(ii) Evaluation Using Survey Questionnaire

The questionnaires were developed and distributed to 20 healthcare representative staff from Temeke, Mwananyamala, Ilala, Mawenzi, and Mount Meru Hospitals who were selected randomly indicated in Appendix 4. The staff was given a chance to use the system to conduct health big data analytics based on the proposed use-cases scenarios. Eight members of IT departments, six members of pediatric and six members of health data management staff were given a chance to use the system. Out of the 20 issued questionnaires, 18 were collected and qualified for analysis. Microsoft Excel was used to analyze the responses for the results. The survey questionnaires had 12 questions on a 5 point on Likert scale of (1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly Disagree). Appendix 4, presents the summary of user's responses table on framework validation testing as indicated in Fig. 19

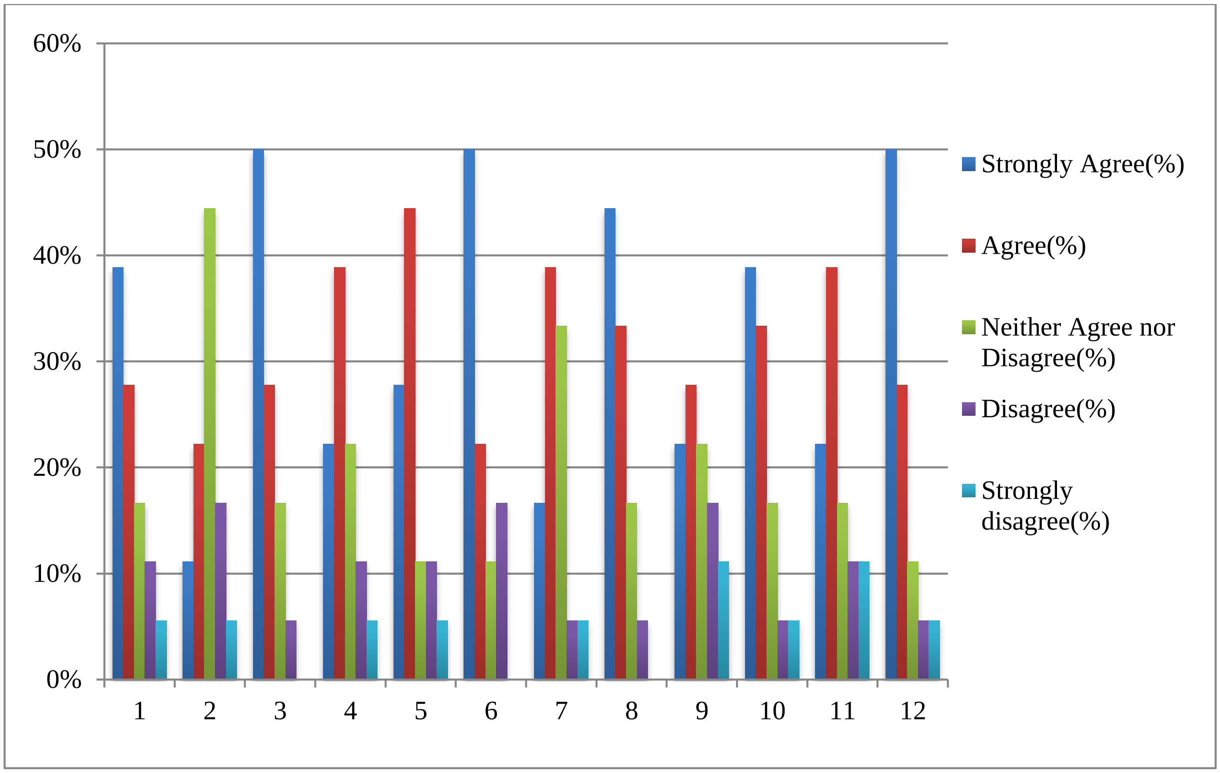


Figure 19: The histogram graph of the responses after the validation experiment

Based on the framework evaluation results the majority of the respondents show the following results:

Outcome of Question Three

Question three was asked about the framework system coverage of all that has been presented in the proposed BDAF-CIDSS. The 50% of the respondents strongly agreed, 28% agreed, 16% were neutral, those who disagreed were only 5% and there were no strongly disagreed as presented in a histogram in Fig. 19.

Outcome of Question Six

Question six asked if the framework system integrated all patients' data in one area for easy processing and analysis. The respondents who strongly agreed were 50%, 22% agreed, neutral was 11% and those who disagreed were 17% as presented in the histogram in Fig. 19.

Outcome of Question Seven

Question seven was asked if the framework system captured the traditional and non-traditional infectious disease data. The number of respondents who strongly agreed were 18%, 39% agreed, neutral was 35%, disagreed and strongly disagreed were 5% respectively as presented in the histogram in Fig. 19.

Outcome of Question Eight

Question eight was asked if the framework system captured text, e-mails, and tables. The number of respondents who strongly agreed were 44% out of all responses, 33% agreed, 16% were neutral, and 5% disagreed as presented in the histogram in Fig. 19.

Outcome of Question 10

Question 10 was asked if the framework system might improve infectious disease data analysis for evidence-based decisions. The number of respondents who strongly agreed were 39%, 34% agreed, 16% were neutral and 5% disagreed as presented in the histogram in Fig. 19.

Outcome of Question 12

Question 12 asked if the framework system reduces the time of the data collection process from the outpatients and community level cases findings. The number of respondents who strongly agreed were 50%, 28% agreed, 11% were neutral, 5% disagreed and 5% strongly disagreed as presented in the histogram in Fig. 19.

In this evaluation, 22% of the majority of users agreed that the proposed big data analytics framework will help healthcare professionals to view and progressively monitor infectious disease-reported cases.

4.2 Discussion

The BDAF-CIDSS has been designed for patients, the community, other healthcare stakeholders, healthcare professionals, and decision-makers to meet their specific needs in Tanzania to prevent and control infectious diseases affecting children 0-5 years of age. The framework has been developed to overcome the following healthcare issues that prevail in the traditional CIDSS:

4.2.1 Data Collection

The framework has been developed to accommodate data collection and analysis process from various healthcare stakeholders. It involves collection of the patients data through Internet-based and mobile apps sources using monitoring sensors, community through social media, websites, public pharmacies, laboratory test results, healthcare insurances, and others. From the framework validation experiment and performance evaluation results, the collection of web-based free-text data and mobile phone data can improve the traditional CIDSS in Tanzania. The best healthcare data formats and system integration using big data analytics technology achieve better and sophisticated data collection and analysis than the traditional system. In data collection, the framework can be widely implemented using a mobile application, SMS messages, online healthcare system, social network, blogging, Internet protocol address, weblogs, healthcare monitoring sensors, and healthcare websites that 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 the Internet, social networks, community-level case findings, hospitals, and dispensaries. Thus the quality and data coverage will be improved by involving all infectious diseases report cases from all traditional and nontraditional data sources.

4.2.2 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, the infectious disease notification alerts including warning, notification messages, and disease outbreaks notifications will be sent to the citizens and healthcare professionals through text messages, e-mail systems, social network pop-ups to take quick action as presented on the data flow diagram. As the proposed framework helps to collect infectious diseases report cases from the Internet, online web-based systems, social media, and mobile phones data sources, this will help to improve data quality and coverage to produce accurate information for early disease detection which will help healthcare professionals take appropriate action to save people's lives on time.

4.2.3 Healthcare Information Analysis

The BDAF-CIDSS tool has been developed to facilitate passive and active infectious disease surveillance using real-time infectious diseases data analysis, infectious diseases report cases monitoring and prediction, and data visualization. Having an integrated commodity computer clusters with big data analysis technology makes it easier to perform various types of data analysis in the public health sector using various algorithms. From our framework validation experiment and performance evaluation results, the use of the proposed BDAF-CIDSS in disease surveillance will help to solve technical and computational challenges that face traditional CIDSS 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.

4.2.4 Evidence-Based Decision Making

The CIDSS conducted in most African countries is conducted in the condition of resource-limited settings in which often suffer 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 (Akinnagbe *et al.*, 2018). Using the proposed framework, an 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 between healthcare professionals and other stakeholders to prevent the emerging and re-emerging of infectious diseases.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

In this study, we developed and described the general BDAF-CIDSS based on Internet-based healthcare data sources and mobile phone healthcare data monitoring. The experimental study was based on Internet-based unstructured free-text, online health databases, and mobile phone structured and unstructured health data monitoring. For Internet-based unstructured data, Web Crawler scrapping program was set up and used, and Hadoop technology to gather online healthcare news articles and free-text data, and on structured data, auto-generated dummy healthcare data was used from Mockaroo (Mockaroo, 2002). Then the Map-Reduce algorithm design model was used to break healthcare data into Map phase and Reduce phase to extract a combination of Key-Value pairs for input and output. Then this study went through a synthetic study to perform short-term validation of the two systems' performance: Big data analytics technology based on MRQL algorithm and the RDBMS based on R-SQL. The goal was to compare system performance between traditional surveillance systems and big data analytics surveillance systems. In this comparison analysis evaluation, more concentration was placed on the MRQL against the traditional R-SQL. The results showed that the collection of web-based, online health databases, and mobile phone healthcare data and analyze those using big data analytics improve infectious disease surveillance data analysis for decision-making. Either, the power of big data analytics technology to analyze and interpret a high volume of healthcare data from Internet-based systems such as online information aggregates, free-text mapping, and classification, search engine queries, online news articles, online government technical reports, RSS feeds, and social network surveillance is more powerful than the traditional system. Furthermore, the computational capability of big data analytics to handle a high volume of healthcare data is more powerful, tolerant, and scalable. It offers more opportunities to write user-defined programs for healthcare data processing than the traditional system.

The BDAF-CIDSS focused mainly on the performance of the traditional CIDSS in Tanzania. The 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. The framework works with existing R-SQL databases and analytics tools. A distributed implementation of the framework 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 the input to output framework. This approach can be used in similar environments worldwide, particularly in developing countries, whereby many of the countries have similar conditions of not paid attention to the infectious disease data quality, coverage, and representativeness. Whether the infectious disease surveillance endpoint is situational awareness, outbreak detection, and control or identifies the infectious disease estimation trends, infectious disease data quality, coverage, and completeness is the key factor during each stage. The assessment of infectious diseases data quality, coverage, and completeness using a diverse mix of data sources and efficient analytical methods is crucial to facilitate evidence-based decision-making to improve people's health in the country. This approach can play a unique role in Tanzania 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 promote health behaviors at the community level 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 BDAF-CIDSS for guidance to build a systematic infectious disease surveillance system that monitors Internet-based healthcare data, online health databases, and mobile phone data sets for infectious disease surveillance in the Tanzanian context, which is more critical in today's digital data revolution. With such a framework, we can systematically collect infectious disease data from the Internet, online health databases, and mobile phones through web-based mapping, online free-text documents, RSS feeds, search engine queries, online information aggregates, social networks, blogging, and local infectious disease cases, thus providing accurate and timely information to decision-makers. It is believed that such a framework is very important to patients, researchers, epidemiologists, decision-makers, the Ministry of Health, and other public healthcare providers. Second, the techniques and methods used are based on big data analytics using the Map-Reduce algorithm which has been reported as the best performing algorithm in big data analytics. It allows distributed and parallel processing of large-scale datasets across commodity computers cluster which can easily be applied in resource-limited setting counties like Tanzania to improve high-performance computation to reduce the cost of ICT infrastructures.

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 Internet-based healthcare data sources, access to such information is limited, technical, costly, security and legal concerned and even impossible for many research society. 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. The online healthcare data needs web queries, information aggregates, free-text filtering, and search engines based technical surveillance. This depends on the availability of sufficient Web-Internet access to generate signals for data response. Furthermore, while there is a lot of disease surveillance data streams algorithm developed by researchers, there is a need for developing multi-variable analytic algorithms which can analyze multiple healthcare data streams based on different groups, targets and stages of the disease.

5.2 Recommendations

An integrated BDAF-CIDSS using big data analytics for controlling infectious diseases affecting children has been proposed which is robust and fast controlling the infectious diseases affecting children in the country. The system architecture and use case diagram have been designed with minimum implementation to improve the performance of the traditional CIDSS to collect and analyze data for proper evidence-based decision-making. The framework facilitates infectious disease data integration to improve data coverage and completeness for evidence-based decision-making for early detection, prevention, and control of infectious disease outbreaks. It ensures the infectious disease cases collection, processing, and analysis to cope with today’s world of technology and comply with the IHR, 2005. It supports online and social media healthcare data surveillance, public pharmacies, home patient monitoring, and integrating with other healthcare-related stakeholders into the mainstream of the CIDSS. It consolidates early infectious disease notifications, warnings, and alerts to the citizens through epidemiologists and managerial inputs to prevent infectious disease threats. It supports infectious disease data flow from the Internet-based and community-level case findings at the local environments up to the central part in the hospitals and provides feedback in reversed direction. Initially, the proposed BDAF-CIDSS is limited to infectious diseases affecting children. However, the framework can easily be expanded to other communicable and non-communicable disease surveillance systems in the country. It can accommodate more other user function tools in big data analytics techniques such as machine learning using mahout big data mining tool to analyze and predict infectious diseases using algorithms like Naïve Bayes Classifier, K–Means Algorithm, Dirichlet, and Logistic Regressions. All these algorithms work better with the Map-Reduce function paradigm in infectious disease predictions.

REFERENCES

Ahmed, W., Jagsi, R., Gutheil, T. G., & Katz, M. S. (2020). Public disclosure on social media of identifiable patient information by health professionals: Content analysis of twitter data. *Journal of Medical Internet Research*, *22*(9), 1–12. https://doi.org/10.2196/19746

Akinnagbe, A., Peiris, K. D. A., & Akinloye, O. (2018). *Prospects of Big Data Analytics in Africa Healthcare System*. *10*(6), 114–122. https://doi.org/10.5539/gjhs.v10n6p114

Al-barhamtoshy, H. M., & Eassa, F. (2014). *A Data Analytic Framework for Unstructured Text*. https://doi.org/10.13140/2.1.4330.0485

Alaoui, I. El, Gahi, Y., Messoussi, R., Chaabi, Y., Todoskoff, A., & Kobi, A. (2018). A novel adaptable approach for sentiment analysis on big social data. *Journal of Big Data*, 5(12),1-18. https://doi.org/10.1186/s40537-018-0120-0

Baechle, C., & Agarwal, A. (2017). A framework for the estimation and reduction of hospital readmission penalties using predictive analytics. *Journal of Big Data*, 4(37), 1–15. https://doi.org/10.1186/s40537-017-0098-z

Bansal, S., Chowell, G., Simonsen, L., Vespignani, A., & Viboud, C. (2016). Big data for infectious disease surveillance and modeling. *Journal of Infectious Diseases*, *214*(Suppl 4), S375–S379. https://doi.org/10.1093/infdis/jiw400

Barkhordari, M., & Niamanesh, M. (2018). Chabok: A Map: Reduce based method to solve data warehouse problems. *Journal of Big Data*, 5(40), 1-25. https://doi.org/10.1186/s40537-018-0144-5

Bennett, C., & Doub, T. (2011). *Data Mining and Electronic Health Records: Selecting Optimal Clinical Treatments in Practice*. 2010, 313–318. http://arxiv.org/abs/1112.1668

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., & Mandl, K. D. (2008). Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Medicine*, *5*(7), 1019–1024. https://doi.org/10.1371/journal.pmed.0050151

Butler, D. (2013). When Google got flu wrong. *Nature*, *494*(7436), 155–156. https://doi.org/10.1038/494155a

Care, Q. H. (2020). *Health Sector Strategic Plan*. https://www.google.com

Chan, E. H., Sahai, V., Conrad, C., & Brownstein, J. S. (2011). Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *PLoS Neglected Tropical Diseases*, *5*(5), 1-6. https://doi.org/10.1371/journal.pntd.0001206

Chandak, M. B. (2016). Role of big - data in classification and novel class detection in data streams. *Journal of Big Data*, 3(5), 1-9. https://doi.org/10.1186/s40537-016-0040-9

Cheng, C. K. Y., Ip, D. K. M., Cowling, B. J., Ho, L. M., Leung, G. M., & Lau, E. H. Y. (2011). Digital dashboard design using multiple data streams for disease surveillance with influenza surveillance as an example. *Journal of Medical Internet Research*, *13*(4),1-9. https://doi.org/10.2196/jmir.1658

Chipwaza, B., Sumaye, R. D., Weisser, M., Gingo, W., Yeo, N. K. W., Amrun, S. N., Okumu, F. O., & Ng, L. F. P. (2021). Occurrence of 4 Dengue Virus Serotypes and Chikungunya Virus in Kilombero Valley, Tanzania, during the Dengue Outbreak in 2018. *Open Forum Infectious Diseases*, *8*(1), 1–6. https://doi.org/10.1093/ofid/ofaa626

Chowell, G., Cleaton, J. M., & Viboud, C. (2016). Elucidating transmission patterns from internet reports: Ebola and middle east respiratory syndrome as case studies. *Journal of Infectious Diseases*, *214*(Suppl 4), S421–S426. https://doi.org/10.1093/infdis/jiw356

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534. https:// doi. org/ 10. 1016/ S1473-3099(20)30120-1

Dredze, M., Cheng, R., Paul, M. J., & Broniatowski, D. (2014). *HealthTweets.org: A platform for public health surveillance using Twitter*. https://www.google.com

Erraguntla, M., Zapletal, J., & Lawley, M. (2019). Framework for Infectious Disease Analysis: A comprehensive and integrative multi-modeling approach to disease prediction and management. *Health Informatics Journal*, *25*(4), 1170–1187. https:// doi. org/ 10. 1177/ 1460458217747112

Feliciano, J. T., Salmi, L., Blotner, C., Hayden, A., Nduom, E. K., Kwan, B. M., Katz, M. S., & Claus, E. B. (2020). Brain tumor discussions on twitter (#BTSM): Social network analysis. *Journal of Medical Internet Research*, *22*(10), 1–8. https://doi.org/10.2196/22005

Ferlie, S. (2001). A Framework for a Systems Approach to Health Care Delivery. http:// www.nap.edu/catalog/11378.html

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. https://doi.org/10.1038/nature07634

Heisey-Grove, D. M., McClelland, L. E., Rathert, C., Tartaglia, A., Jackson, K., & DeShazo, J. P. (2020). Associations between Patient Health Outcomes and Secure Message Content Exchanged between Patients and Clinicians: Retrospective Cohort Study. *Journal of Medical Internet Research*, *22*(10), 1–14. https://doi.org/10.2196/19477

Herland, M., Khoshgoftaar, T. M., & Bauder, R. A. (2018). Big Data fraud detection using multiple medicare data sources. *Journal of Big Data*, 5(29), 1–21. https:// doi. org/ 10. 1186/ s40537 -018-0138-3

HSSP VI. (2021). *Health Sector Strategic Plan*. https://www.google.com

Hu, H. A. N., Wen, Y., Member, S., & Chua, T. (2014). Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *Access, 2*, 652–687. https:// doi. org/ 10. 1109/ ACCESS. 2014. 2332453

Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2013). *Big Data for Dummies.* https://www.google.com

Imai, C., Armstrong, B., Chalabi, Z., Mangtani, P., & Hashizume, M. (2015). Time series regression model for infectious disease and weather. *Environmental Research*, *142*, 319–327. https://doi.org/10.1016/j.envres.2015.06.040

Jeon, J., Baruah, G., Sarabadani, S., & Palanica, A. (2020). Identification of risk factors and symptoms of COVID-19: Analysis of biomedical literature and social media data. *Journal of Medical Internet Research*, *22*(10), 1–10. https://doi.org/10.2196/20509

Jia, Q., Guo, Y., Wang, G., & Barnes, S. J. (2020). Big data analytics in the fight against major public health incidents (Including COVID-19): A conceptual framework. *International Journal of Environmental Research and Public Health*, *17*(17), 1–21. https://doi.org/10.3390/ijerph17176161

Kalabamu, F., & Maliki, S. (2021). Use of Haematological Changes as a Predictor of Dengue

Infection among Suspected Cases at Kairuki Hospital in Dar Es Salaam, Tanzania: A Retrospective Cross Sectional Study. East African Health Research Journal, 5(1), 91–98. https://doi.org/10.24248/eahrj.v5i1.655

Kamel Boulos, M. N., Resch, B., Crowley, D. N., Breslin, J. G., Sohn, G., Burtner, R., Pike, W. A., Jezierski, E., & Chuang, K. Y. S. (2011). Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: Trends, OGC standards and application examples. *International Journal of Health Geographics*, *10*, 1–29. https://doi.org/10.1186/1476-072X-10-67

Kamradt-Scott, A. (2019). The International Health Regulations (2005). *International Organizations Law Review*, *16*(2), 242–271. https://doi.org/10.1163/15723747-01602002

Keller, M., Blench, M., Tolentino, H., Freifeld, C. C., Mandl, K. D., Mawudeku, A., Eysenbach, G., & Brownstein, J. S. (2009). Use of unstructured event-based reports for global infectious disease surveillance. *Emerging Infectious Diseases*, *15*(5), 689–695. https://doi.org/10.3201/eid1505.081114

Kijsanayothin, P., Chalumporn, G., & Hewett, R. (2019). On using MapReduce to scale algorithms for Big Data analytics: A case study. *Journal of Big Data*, 6(105) 1–20. https://doi.org/10.1186/s40537-019-0269-1

Kwok, S. W. H., Vadde, S. K., & Wang, G. (2021). Tweet topics and sentiments relating to COVID-19 vaccination among Australian twitter users: Machine learning analysis. *Journal of Medical Internet Research*, *23*(5), 1-16. https://doi.org/10.2196/26953

Li, W. Y., Chiu, F. C., Zeng, J. K., Li, Y. W., Huang, S. H., Yeh, H. C., Cheng, B. W., & Yang, F. J. (2020). Mobile health app with social media to support self-management for patients with chronic kidney disease: Prospective randomized controlled study. *Journal of Medical Internet Research*, *22*(12), 1–15. https://doi.org/10.2196/19452

Liu, X., Wang, X., Matwin, S., & Japkowicz, N. (2015). Meta-MapReduce for scalable data mining. *Journal of Big Data*, 2(14), 1-23. https://doi.org/10.1186/s40537-015-0021-4

Loola, B. P., Ouazzani, T. K., & Souissi, N. (2020). *Mobile Data Collection Using Open Data Kit*, *1*, 543–550. https://doi.org/10.1007/978-3-030-36778-7_60

Mangu, C. D., Manyama, C. K., Sudi, L., Sabi, I., Msila, H., & Nyanda, E. (2016). *Emerging* *viral infectious disease threat : Why Tanzania is not in a safe zone*. *18*(3), 1–14.

MoHCDGEC. (2017). *Tanzania National Road map of eHealth Investment 2017-2023*. Tanzania Digital Health Investment Road Map. https://www.google.com

Mwanyika, G. O., Mboera, L. E. G., Rugarabamu, S., Ngingo, B., Sindato, C., Lutwama, J. J., Paweska, J. T., & Misinzo, G. (2021). Dengue virus infection and associated risk factors in africa: A systematic review and meta-analysis. *Viruses*, *13*(4), 1–17. https://doi.org/10.3390/v13040536

Nagwani, N. K. (2015). Summarizing large text collection using topic modeling and clustering based on MapReduce framework. *Journal of Big Data*, 2(6) 1–18. https://doi.org/10.1186/s40537-015-0020-5

Naveen, A., Antarip, B., Sumit, D., Saurav, N., & Rajiv, P. (2012). *The Abzooba Smart Health Informatics Platform (SHIP): From Patient Experiences to Big Data to Insights.* https://www.google.com

Nkowane, B. M. (2019). *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*. https://pdf.usaid.gov/pdf_docs/PA00TKWC.pdf

Osadchiy, V., Jiang, T., Mills, J. N., & Eleswarapu, S. V. (2020). Low testosterone on social media: Application of natural language processing to understand patients' perceptions of hypogonadism and its treatment. *Journal of Medical Internet Research*, *22*(10), 1–16. https://doi.org/10.2196/21383

Oussous, A., Benjelloun, F., Ait, A., & Belfkih, S. (2018). Big Data technologies : A survey. *Journal of King Saud University - Computer and Information Sciences*, *30*(4), 431–448. https://doi.org/10.1016/j.jksuci.2017.06.001

Pal, G., Hong, X., Wang, Z., Wu, H., Li, G., & Atkinson, K. (2019). Lifelong Machine Learning and root cause analysis for large - scale cancer patient data. *Journal of Big Data*, 6(108), 1-29. https://doi.org/10.1186/s40537-019-0261-9

Paolotti, D., Carnahan, A., Colizza, V., Eames, K., Edmunds, J., Gomes, G., Koppeschaar, C., Rehn, M., Smallenburg, R., Turbelin, C., Van Noort, S., & Vespignani, A. (2014). Web-based participatory surveillance of infectious diseases: The Influenzanet participatory surveillance experience. *Clinical Microbiology and Infection*, *20*(1), 17–21. https://doi.org/10.1111/1469-0691.12477

Pavlin, B., & Cic, M. P. H. F. (n.d.). *WHO | Disease early warning, alert and response in emergencies*. http://www.who.int/features/2016/disease-early-warning-response/en/

Pivette, M., Mueller, J. E., Crépey, P., & Bar-Hen, A. (2014). Drug sales data analysis for outbreak detection of infectious diseases: A systematic literature review. *BMC Infectious Diseases*, *14*(1). 1-14. https://doi.org/10.1186/s12879-014-0604-2

Priyanka, K., & Kulennavar, N. (2014). A survey on big data analytics in health care. *International Journal of Computer Science and Information Technologies*, *5*(4), 5865–5868. https://doi.org/5: 5865-5868

Reid, P. P., Compton, W. D., & Jerome, H. (2005). *Better Delivery System*. https://www.google.com

Salathé, M. (2016). Digital pharmacovigilance and disease surveillance: Combining traditional and big-data systems for better public health. *Journal of Infectious Diseases*, *214*(Suppl 4), S399–S403. https://doi.org/10.1093/infdis/jiw281

Salathé, M., & Khandelwal, S. (2011). Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. *PLoS Computational Biology*, *7*(10), 1-7. https://doi.org/10.1371/journal.pcbi.1002199

Sanders, R., Araujo, T. B., Vliegenthart, R., van Eenbergen, M. C., van Weert, J. C. M., & Linn, A. J. (2020). Patients' Convergence of Mass and Interpersonal Communication on an Online Forum: Hybrid Methods Analysis. *Journal of Medical Internet Research*, *22*(10), 1–14. https://doi.org/10.2196/18303

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). 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*, *6*(5). 1-10. https://doi.org/10.1371/journal.pone.0019467

Simon, P. W. and. (2013). Too Big to Ignore: The Business Case for Big Data. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699.

Sobowale, K., Hilliard, H., Ignaszewski, M. J., & Chokroverty, L. (2020). Real-time communication: Creating a path to COVID-19 public health activism in adolescents using social media. *Journal of Medical Internet Research, 22*(12). 1-9. https://doi.org/10.2196/21886

Swinkels, H. M., Kuo, M., Embree, G., Stone, J., Trerise, S., Brisdon, S., Louie, K., Asplin, R., Stiller, A., Abraham, T., Gill, I. S., Rice, G., Andonov, A., Henry, B., & Buxton, J. A. (2014). Hepatitis a outbreak in British Columbia, Canada: The roles of established surveillance, consumer loyalty cards and collaboration, February to May 2012. *Eurosurveillance*, *19*(18), 8–15. https://doi.org/10.2807/1560-7917.ES2014.19.18.20792

Torabzadehkashi, M., Rezaei, S., Heydarigorji, A., Bobarshad, H., & Alves, V. (2019). Computational storage: an efficient and scalable platform for big data and HPC applications. *Journal of Big Data,* 6(100), 1-29. https://doi.org/10.1186/s40537-019-0265-5

Vatrapu, R., Mukkamala, R. R. A. O., Hussain, A., & Flesch, B. (2016). Social Set Analysis : A Set Theoretical Approach to Big Data Analytics. *IEEE Access*, *4*, 2542–2571. https://doi.org/10.1109/ACCESS.2016.2559584

Version, S. D. (2017). *The United Republic of Tanzania the National Health Policy 2017*, *October*. https://www.google.com

Wang, Y., Kung, L. A., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, *126*, 3–13. https://doi.org/10.1016/j.techfore.2015.12.019

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science*, *338*(6104), 267–270. https://doi.org/10.1126/science.1223467

Yuan, Q., Nsoesie, E. O., Lv, B., Peng, G., Chunara, R., & Brownstein, J. S. (2013). *Monitoring Influenza Epidemics in China with Search Query from Baidu*. *8*(5), 1-7. https://doi.org/10.1371/journal.pone.0064323

APPENDICES

Appendix 1: Questionnaire for data collection

Survey questionnaire for the proposed BDAF-CIDSS in Tanzania.

This survey questionnaire exercise has three main objectives:

1. To understand the performance and its potential barriers of the existing traditional infectious disease surveillance and responses system framework.

2. To understand the use of healthcare Big data analytics in decision-making, healthcare business processes and emerging business models, and

3. To gather insights on the characteristics of a functioning data ecosystem in healthcare centers and identify existing or potential barriers to the development of data-driven healthcare sectors in Tanzania.

The questionnaire takes about twenty minutes to complete. The outcome will contribute to the development of the Healthcare Big data analytics framework for data-driven surveillance on Infectious diseases affecting children in Tanzania.

In case you or your organization wishes to submit any additional input please send it to mdoe.mwamnyange@nm-aist.ac.tz

Thank you for your time and input.

Section A: Information about the organization or Staff

Section B: Major challenges of the prevention and control the infectious diseases

□ Mwananyamala Hospital

□ Ilala Hospital

□ Temeke Hospital

□ Mawenzi Hospital

□ Mount Meru Hospital

□ Mbeya Referral Hospital

	Not at all true	Very lightly true	Very true	Quite true	Com true
Q 1. Inadequate of the infrastructures / facilities	□	□	□	□	□

Q.2. The access and quality of the healthcare centers' services?

□ Not at all true □ Very lightly true □ Very true □Quite true □Completely true

Q.3 From what sources does your organization collect, or expects to collect, clinics healthcare data?

For each row, please tick only one box. To suggest other sources, please use the space provided.

	Collect now	Expects to collect in 5 years	No plans to collect	Do not know
Log	□	□	□	□
Text	□	□	□	□
Events	□	□	□	□
Emails	□	□	□	□
Social media	□	□	□	□
Tele-medical Sensors	□	□	□	□
Open data/Public Sector Information (PSI)	□	□	□	□
Phone usage	□	□	□	□
External feeds	□	□	□	□
RFID scans or POS data	□	□	□	□
Earth Observation and Space	□	□	□	□
Other Geospatial	□	□	□	□
Free-form text	□	□	□	□
Audio	□	□	□	□
Still images/videos	□	□	□	□

	Not at all true	Very lightly true	Very true	Quite true	Completely true
Q.4 Inability to collect infectious diseases data from the patient's environments.	□	□	□	□	□
Q.5 Poor quality of food, water, and housing services	□	□	□	□	□

Q.6. People's culture

□ Not at all true □ Very lightly true □ Very true □Quite true □Completely true

Q.7. The insufficient number of healthcare staff.

□ Not at all true □ Very lightly true □ Very true □Quite true □Completely true

Section C: Infectious disease data collection and analysis

	Not at all true	Very lightly true	Very true	Quite true	Completely true
Q.8 Do you have an electronic system to receive self-produced information from the patients	□	□	□	□	□
Q.9 Do you collect infectious disease data through web, text, emails, phone, images, and video?	□	□	□	□	□

Q.10. Do you provide general health clinics, illness, electronic appointments, and service consultation by phone, text messages, e-mail, online communication, and social media?

□ Not at all true □ Very lightly true □ Very true □ Quite true □Completely true

Q.11. It would be interesting to collect and analyze infectious disease data through mobile phones.

□ Not at all true □ Very lightly true □ Very true □ Quite true □Completely true

Q.12. It would be interesting to collect and analyze infectious disease data through social media.

□ Not at all true □ Very lightly true □ Very true □ Quite true □Completely true

Section D: Involvement of citizen/public in infectious disease prevention and control

	Not at all true	Very lightly true	Very true	Quite true	Completely true
Q.13 Do you involve citizen/public in infectious disease prevention and control?	□	□	□	□	□
Q.14 Do you use social media to collect infectious disease data?	□	□	□	□	□

Q.15. Do you collect and integrate infectious disease data from other related healthcare data e.g NHIF, pharmacies, etc?

□ Not at all true □ Very lightly true □ Quite true □ Completely true

Q.16. Does your organization share healthcare data with other entities e.g government, citizens, donors, etc?

□ Not at all true □ Very lightly true □ Quite true □ Completely true

Section E: Organization experience in Healthcare Big Data Analytics and Data-driven

	Not at all true	Very lightly true	Very true	Quite true	Completely true
Q.17 Does your organization has experience in Healthcare Big data or data analytics?	□	□	□	□	□
Q.18 It is difficult to obtain infectious disease data from other healthcare-related data sources	□	□	□	□	□

Q.19. Information on infectious diseases from other sources would have improved disease prevention and control.

□ Not at all true □ Very lightly true □ Quite true □ Completely true

Q.20. Do you collect data on infectious diseases through the web, text, e-mails, social media, mobile phones, external feeds, images, and video?

□ Not at all true □ Very lightly true □ Quite true □ Completely true

Q.21. Does your organization have Healthcare Big data Framework for processing, analyzing, and transforming very large healthcare datasets?

□ Not at all true □ Very lightly true □ Quite true □ Completely true

Q.22. Which departments in your organization are involved in using healthcare Big data technologies and data analytics?

Please tick all that apply. To suggest other departments, please use the space provided.

□ IT

□ Human Resources

□ Operations

□ Customer Service

□ Executive Management

Section F: The use of Healthcare Big Data technology

	all true	lightly true		true	true
Q.23 Do you have experience with Healthcare Big data technology?	□	□	□	□	□
Q.24 If available, I would participate in training about Healthcare Big data technology	□	□	□	□	□

Q.25. I use phone calls, social media, sms messages mainly for personal purposes.

□ Not at all true □ Very lightly true □ Quite true □ Completely true

Q.26. Phone calls, sms messages are often used for healthcare activities-related purposes.

□ Not at all true □ Very lightly true □ Quite true □ Completely true

Q.27. I sometimes use the internet on my phone

□ Not at all true □ Very lightly true □ Quite true □ Completely true

Q.28. I believe access to infectious disease information through mobile phones, websites, e-mails, content management systems, and social media would improve healthcare services to the citizens.

□ Not at all true □ Very lightly true □ Quite true □ Completely true

Q.29. I would be interested in participating in a project which tests Healthcare Big data using Big data analytics technology. These services might include data collection, processing, and analysis.

□ Not at all true □ Very lightly true □ Quite true □ Completely true

Q.30. In your organization, healthcare data collection, storage, and analysis are:

□ In-house □ Outsourced □ Both

Q.31. Does your organization have a website with information on healthcare services? Yes / No

If Yes, What information and functions does the website offer?

□ Information on the healthcare services provided by the organization? Yes / No.

□ Information on the organization's units (contact information, location). Yes / No.

□ Search functions for units or service providers providing various services. Yes / No.

□ Function for sending patient feedback on the care. Yes / No.

Thank you for taking the time to complete this questionnaire! Your contribution is very important to understand the role and impacts of healthcare big data on Infectious Diseases affecting Children.

Should you have any question, please contact:

Mr. Mdoe Mwamnyange

E-mail: mdoe.mwamnyange@nm-aist.ac.tz

Appendix 2: Summary of Survey Responses

SN	QUESTIONS	NT	VLT	QT	VT	CT	%
	Major challenges of infectious diseases prevention and control						
1.	Inadequate of the infrastructure / facilities	2	5	8	54	39	100%
2.	The access and quality of the healthcare centers' services	4	7	43	26	28	100%
3.	Difficulties to reach remote areas	2	6	39	21	40	100%
4.	Inability to collect infectious diseases data from the patient's environments	1	4	16	36	51	100%
5.	Poor quality of food, water and housing services	1	2	7	20	78	100%
6.	People's culture	5	3	30	44	26	100%
7.	Insufficient number of healthcare staff	6	5	21	35	41	100%
	Infectious disease data collection and analysis						
8.	Do you have an electronic system to receive self produced information from the patients	68	24	2	12	2	100%
9.	Do you collect and analyze infectious disease data through web, text, emails, phone, images, and video?	65	9	24	9	1	100%
10.	Do you provide general health clinics, illness, electronic appointments, and service consultation by phone, text messages, e-mail, online communication, and social media?	72	23	9	1	3	100%
11.	It would be interesting to collect and analyze infectious disease data through mobile phones.	1	9	20	21	57	100%
12.	It would be interesting to collect and analyze infectious disease data through social media	1	3	24	32	48	100%
	Involvement of citizen / public in infectious disease prevention and control						
13.	Do you involve citizen / public in infectious disease prevention and control?	48	24	31	3	2	100%
14.	Do you use social media to collect infectious disease data?	72	15	13	6	2	100%
15.	Do you collect, integrate and analyze infectious disease data from other related healthcare data e.g. NHIF, pharmacies, government and R&D reports etc?	12	5	14	34	43	100%
16.	Does your organization share healthcare data with other entities such as the government, citizens, donors, etc?	1	7	11	24	65	100%
	Organization experience in Healthcare Big data and data-driven innovation						
17.	Does your organization have experience in Healthcare Big data or data analytics?	45	4	23	32	4	100%
18.	It is difficult to obtain infectious disease data from other healthcare related data sources	25	9	32	28	14	100%
19.	Information of infectious diseases from Internet-based sources would have improved disease prevention and control	4	2	25	21	56	100%
20.	Do you collect data on infectious diseases through the web, text, e-mails, social media, mobile phones, external feeds, images, and video?	52	31	22	2	1	100%
21.	Does your organization have Healthcare Big data Framework for processing, analyzing, and transforming very large healthcare datasets?	74	22	9	2	1	100%
	The use of Healthcare Big data technology						

22.	Do you have experience with Healthcare Big data technology?	82	2	16	5	3	100%
23.	If available, I would participate in training about Healthcare Big data technology	1	2	20	31	54	100%
24.	I make phone calls, social media, text messages, Internet mainly for personal purposes	4	10	18	24	52	100%
25.	Phone calls, text messages are often used for healthcare activities related purposes	2	5	12	21	68	100%
26.	I sometimes use Internet on my phone	1	4	21	31	51	100%
27.	I believe access to infectious disease information through mobile phones, websites, e-mails, content management system and social media would improve healthcare services to citizens	2	7	20	26	53	100%
28.	I would be interested in participating in a project which tests Healthcare Big data using Big data analytics technology. These services might include data collection, processing, and analysis.	1	9	16	23	59	100%

Table 2: Summary of survey responses

Appendix 3: User Validation Test Questionnaire

User validation test questionnaire for the proposed BDAF-CIDSS.

This user validation testing questionnaire exercise has three main objectives:

1. To test and validate the performance of the proposed Big data analytics framework for Childhood Infectious disease surveillance and response system.

2. To test and validate the use of the proposed Big data analytics framework for Childhood Infectious disease surveillance and response system in real-world simulation cases related to traditional infectious disease surveillance and response system, and

3. To gather insights on the actual performance of the proposed Infectious disease surveillance and response system in the healthcare data ecosystem in healthcare centers and identifies existing or potential barriers to the development of data-driven healthcare sectors in Tanzania.

The questionnaire takes about ten minutes to complete. The outcome will contribute to the development of the Healthcare Big data analytics framework for data-driven surveillance on Infectious diseases affecting children in Tanzania.

In case you or your organization wishes to submit any additional input please send it to mdoe.mwamnyange@nm-aist.ac.tz

Thank you for your time and input.

Section A: Information about the organization or Staff

Organization / Hospital / Staff from:

□ Mwananyamala Hospital

□ Ilala Hospital

□ Temeke Hospital

□ Mawenzi Hospital

□ Mount Meru Hospital

□ Mbeya Referral Hospital

□ Healthcare Executive Staff

□ Clinic

□ IT department

□ Data Storage department

□ Other:

Section B -1: The Framework system performance

	Strongly Agree	Agree	Neutral	Disagree	Strongly disagree
Q.1 Framework validation system interactive?	□	□	□	□	□
Q.2 Does the Framework validation system user-friendly?	□	□	□	□	□

Q.3. Does the Framework validation system cover what was presented in the Framework?

□ Strongly Agree □ Agree □ Neutral □ Disagree □Strongly disagree

Q.4. Does the Framework easy to learn?

□ Strongly Agree □ Agree □ Neutral □ Disagree □Strongly disagree

Q.5. Does the framework captures most of the infectious disease report cases from patients?

□ Strongly Agree □ Agree □ Neutral □ Disagree □Strongly disagree

Q.6. The Framework has integrated all patients' data in one area for easy processing and analysis?

□ Strongly Agree □ Agree □ Neutral □ Disagree □Strongly disagree

Section B-2: The Framework system performance with related to the traditional CIDSS

	Strongly Agree	Agree	Neutral	Disagree	Strongly disagree
Q.7 Framework system captured traditional and non-traditional infectious disease data?	□	□	□	□	□
Q.8 Framework system testing captured weblogs, text, e-mails, and tables?	□	□	□	□	□

Q.9. Does the framework improves infectious disease data monitoring?

□ Strongly Agree □ Agree □ Neutral □ Disagree □Strongly disagree

Q.10. The Framework system testing might improve infectious disease data analysis for evidence-based decisions?

□ Strongly Agree □ Agree □ Neutral □ Disagree □Strongly disagree

Q.11. The Framework will improve the performance of the traditional infectious disease surveillance system?

□ Strongly Agree □ Agree □ Neutral □ Disagree □Strongly disagree

Q.12. Does the Framework system testing reduce the time of data collection from the outpatients and community level case findings?

□ Strongly Agree □ Agree □ Neutral □ Disagree □Strongly disagree

Thank you for taking the time to complete this questionnaire! Your contribution is very important to understand the role and impacts of healthcare big data on Infectious Diseases affecting Children.

Appendix 4: Summary of Users' Responses

SN	Questions	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	%
1.	Framework validation system interactive?	7	5	3	2	1	100%
2.	Does the Framework system user-friendly?	2	4	8	3	1	100%
3.	Does the Framework validation system cover what was presented in the Framework?	9	5	3	1	-	100%
4.	Does the Framework easy to learn?	4	7	4	2	1	100%
5.	Framework captures most of the infectious disease report cases from patients?	5	8	2	2	1	100%
6.	The Framework has integrated all patients' data in one area for easy processing and analysis?	9	4	2	3	-	100%
7.	Framework system captured traditional and non-traditional infectious disease data?	3	7	6	1	1	100%
8.	Framework system testing captured weblogs, text, e-mails, and tables?	8	6	3	1	-	100%
9.	The framework improves infectious disease data monitoring?	4	5	4	3	2	100%
10.	The framework system testing might improve infectious disease data analysis for an evidence-based decision?	7	6	3	1	1	100%
11.	The framework will improve the performance of the traditional infectious disease surveillance system?	4	7	3	2	2	100%
12.	Does the framework system testing reduce the time of data collection and processing from the outpatients and community level case findings?	9	5	2	1	1	100%

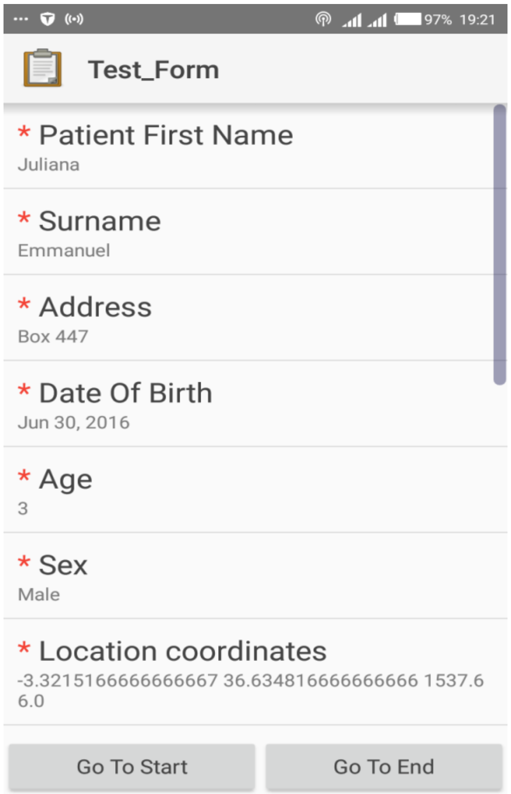
Table 3: Summary of users' responses during the framework validation experiment

Appendix 5: Online Disease Case Definition Form

Data Field	**Field description**
Patient_First_Name	First name of the patient
Surname	Last name of the patient
Address	Address / Area of the patient
Date_Of_Birth	Date of birth of the patient
Age	Age of the patient between 0-5 years
Sex	Gender of the patient
Location_coordinates	Area location where the case was collected
Infectious_Disease	Suspected Infectious disease cases: options are: Pneumonia, Diarrhea, Hepatitis, Measles, and ARI.
Date_Of_Diagnosis	Date of diagnosis if conducted
Laboratory_results	Laboratory results tested if any
Case_Classification	Suspected disease case classification: scope: probable, possible and confirmed.
Hospitalized	Information of hospitalization if any
Name_Of_Notifier	Name of the officer who recorded the information
Telephone	Telephone number of the officer recorded information

Table 4: Online disease case definition form

Appendix 6: Disease Case Definition Open Data Kit Mobile



Test_Form

* Patient First Name
Juliana

* Surname
Emmanuel

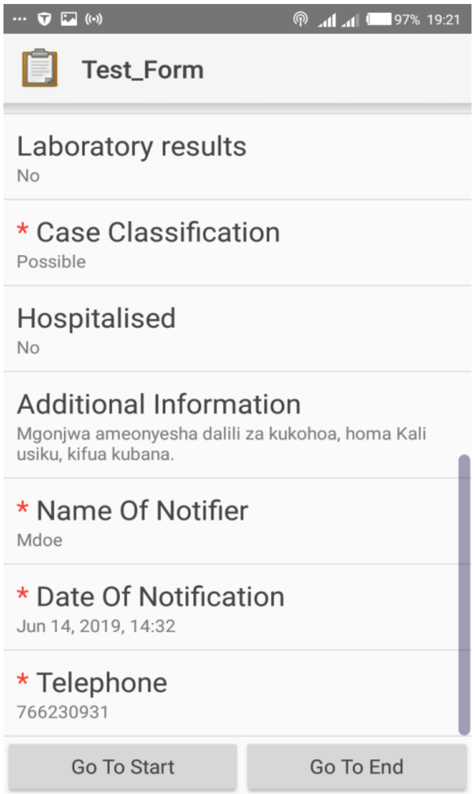
* Address
Box 447

* Date Of Birth
Jun 30, 2016

* Age
3

* Sex
Male

* Location coordinates
-3.3215166666666667 36.634816666666666 1537.6 6.0



Test_Form

Laboratory results
No

* Case Classification
Possible

Hospitalised
No

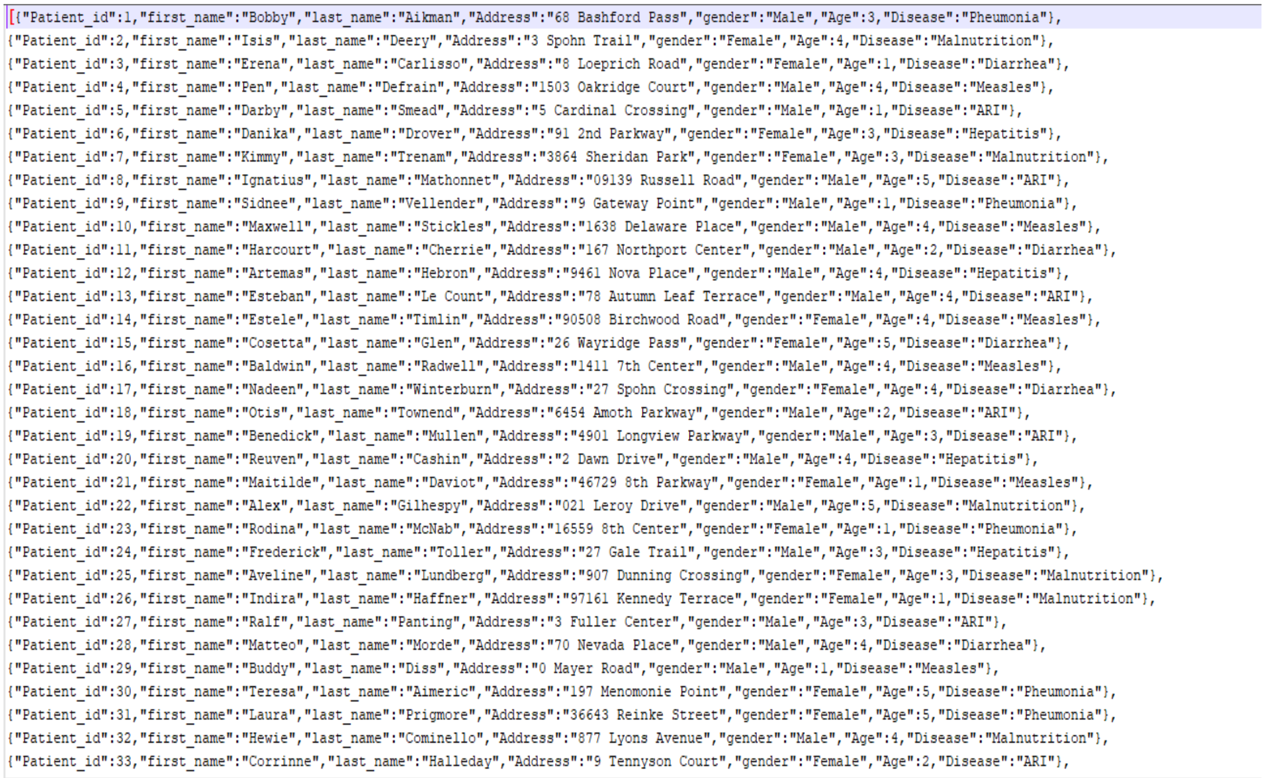
Additional Information
Mgonjwa ameonyesha dalili za kukohoa, homa Kali usiku, kifua kubana.

* Name Of Notifier
Mdoe

* Date Of Notification
Jun 14, 2019, 14:32

* Telephone
766230931

Appendix 7: Hospital Data Sample from Mockaroo.com



```
[{"Patient_id":1,"first_name":"Bobby","last_name":"Aikman","Address":"68 Bashford Pass","gender":"Male","Age":3,"Disease":"Pheumonia"},
{"Patient_id":2,"first_name":"Isis","last_name":"Deery","Address":"3 Spohn Trail","gender":"Female","Age":4,"Disease":"Malnutrition"},
{"Patient_id":3,"first_name":"Erena","last_name":"Carlisso","Address":"8 Loeprich Road","gender":"Female","Age":1,"Disease":"Diarrhea"},
{"Patient_id":4,"first_name":"Pen","last_name":"Defrain","Address":"1503 Oakridge Court","gender":"Male","Age":4,"Disease":"Measles"},
{"Patient_id":5,"first_name":"Darby","last_name":"Smead","Address":"5 Cardinal Crossing","gender":"Male","Age":1,"Disease":"ARI"},
{"Patient_id":6,"first_name":"Danika","last_name":"Drover","Address":"91 2nd Parkway","gender":"Female","Age":3,"Disease":"Hepatitis"},
{"Patient_id":7,"first_name":"Kimmy","last_name":"Trenam","Address":"3864 Sheridan Park","gender":"Female","Age":3,"Disease":"Malnutrition"},
{"Patient_id":8,"first_name":"Ignatius","last_name":"Mathonnet","Address":"09139 Russell Road","gender":"Male","Age":5,"Disease":"ARI"},
{"Patient_id":9,"first_name":"Sidnee","last_name":"Vellender","Address":"9 Gateway Point","gender":"Male","Age":1,"Disease":"Pheumonia"},
{"Patient_id":10,"first_name":"Maxwell","last_name":"Stickles","Address":"1638 Delaware Place","gender":"Male","Age":4,"Disease":"Measles"},
{"Patient_id":11,"first_name":"Harcourt","last_name":"Cherrie","Address":"167 Northport Center","gender":"Male","Age":2,"Disease":"Diarrhea"},
{"Patient_id":12,"first_name":"Artemas","last_name":"Hebron","Address":"9461 Nova Place","gender":"Male","Age":4,"Disease":"Hepatitis"},
{"Patient_id":13,"first_name":"Esteban","last_name":"Le Count","Address":"78 Autumn Leaf Terrace","gender":"Male","Age":4,"Disease":"ARI"},
{"Patient_id":14,"first_name":"Estele","last_name":"Timlin","Address":"90508 Birchwood Road","gender":"Female","Age":4,"Disease":"Measles"},
{"Patient_id":15,"first_name":"Cosetta","last_name":"Glen","Address":"26 Wayridge Pass","gender":"Female","Age":5,"Disease":"Diarrhea"},
{"Patient_id":16,"first_name":"Baldwin","last_name":"Radwell","Address":"1411 7th Center","gender":"Male","Age":4,"Disease":"Measles"},
{"Patient_id":17,"first_name":"Nadeen","last_name":"Winterburn","Address":"27 Spohn Crossing","gender":"Female","Age":4,"Disease":"Diarrhea"},
{"Patient_id":18,"first_name":"Otis","last_name":"Townend","Address":"6454 Amoth Parkway","gender":"Male","Age":2,"Disease":"ARI"},
{"Patient_id":19,"first_name":"Benedick","last_name":"Mullen","Address":"4901 Longview Parkway","gender":"Male","Age":3,"Disease":"ARI"},
{"Patient_id":20,"first_name":"Reuven","last_name":"Cashin","Address":"2 Dawn Drive","gender":"Male","Age":4,"Disease":"Hepatitis"},
{"Patient_id":21,"first_name":"Maitilde","last_name":"Daviot","Address":"46729 8th Parkway","gender":"Female","Age":1,"Disease":"Measles"},
{"Patient_id":22,"first_name":"Alex","last_name":"Gilhespy","Address":"021 Leroy Drive","gender":"Male","Age":5,"Disease":"Malnutrition"},
{"Patient_id":23,"first_name":"Rodina","last_name":"McNab","Address":"16559 8th Center","gender":"Female","Age":1,"Disease":"Pheumonia"},
{"Patient_id":24,"first_name":"Frederick","last_name":"Toller","Address":"27 Gale Trail","gender":"Male","Age":3,"Disease":"Hepatitis"},
{"Patient_id":25,"first_name":"Aveline","last_name":"Lundberg","Address":"907 Dunning Crossing","gender":"Female","Age":3,"Disease":"Malnutrition"},
{"Patient_id":26,"first_name":"Indira","last_name":"Haffner","Address":"97161 Kennedy Terrace","gender":"Female","Age":1,"Disease":"Malnutrition"},
{"Patient_id":27,"first_name":"Ralf","last_name":"Panting","Address":"3 Fuller Center","gender":"Male","Age":3,"Disease":"ARI"},
{"Patient_id":28,"first_name":"Matteo","last_name":"Morde","Address":"70 Nevada Place","gender":"Male","Age":4,"Disease":"Diarrhea"},
{"Patient_id":29,"first_name":"Buddy","last_name":"Diss","Address":"0 Mayer Road","gender":"Male","Age":1,"Disease":"Measles"},
{"Patient_id":30,"first_name":"Teresa","last_name":"Aimeric","Address":"197 Menomonie Point","gender":"Female","Age":5,"Disease":"Pheumonia"},
{"Patient_id":31,"first_name":"Laura","last_name":"Prigmore","Address":"36643 Reinke Street","gender":"Female","Age":5,"Disease":"Pheumonia"},
{"Patient_id":32,"first_name":"Hewie","last_name":"Cominello","Address":"877 Lyons Avenue","gender":"Male","Age":4,"Disease":"Malnutrition"},
{"Patient_id":33,"first_name":"Corrinne","last_name":"Halleday","Address":"9 Tennyson Court","gender":"Female","Age":2,"Disease":"ARI"},
```

Table 5: Hospital data sample from mockaroo.com

Appendix 8: Scrapy Spider Python Program Script

Item.py Script

```
import scrappy


class InfectiousDiseaseDataItem(scrapy.Item):
    # define the fields for your item here like:
    title = scrapy.Field()
    author = scrapy.Field()
    tag = scrapy.Field()
```

MongoDB_Pipeline.py Script

```
import pymongo


class InfectiousDiseaseDataPipeline:


def __init__(self):
        self.conn = pymongo.MongoClient( 'localhost', 27017)
        db = self.conn['infectiousdiseasedata']
        self.collection = db['infectious_disease_tb']


    def process_item(self, item, spider):
        self.collection.insert (dict(item))


        return item
```

Spider.py Script

```
import scrapy
from ..items import InfectiousDiseaseDataItem


class InfectiousDiseaseDataSpider(scrapy.Spider):
    name = 'infectious_disease_data_spider'
```

```
start_urls = ['http://…………………………………………./']
def parse(self, response):
    items = InfectiousDiseaseDataItem()

    all_div_infectious_disease_data = response.css('div.disease')

         for disease in all_div_infectious_disease_data:

        title = disease.css('span.text::text').extract()
        author = disease.css('.author::text').extract()
        tag = disease.css('.tag::text').extract()

        items['title'] = title
        items['author'] = author
        items['tag'] = tag
        yield items
```

Appendix 9: Use Case Scenario I Java Code Snippet

Map Reduce Algorithm Java Code Snippet:

Reduce-Side Join Class

```
import java.io.IOException;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.MultipleInputs;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

 public class ReduceJoin {
 public static class HospitalMapper extends Mapper <Object, Text, Text, Text>
         {
 public void map(Object key, Text value, Context context)
 throws IOException, InterruptedException
         {
         String record = value.toString();
         String[] parts = record.split(",");
         context.write(new Text(parts[0]), new Text("hosp   " + parts[1]));
   }
 }
```

NHIF Mapper Class

```
 public  static class NhifMapper extends Mapper <Object, Text, Text, Text>
         {
 public void map(Object key, Text value, Context context)
 throws IOException, InterruptedException
{
 String record = value.toString ();
 String [] parts = record.split (",");
 context.write(new Text(parts[2]), new Text("nhif   " + parts[3]));
    }
 }
```

Reduce Join Reducer Class

```
 public static class ReduceJoinReducer extends Reducer <Text, Text, Text, Text>
```

```
    {
public void reduce(Text key, Iterable<Text> values, Context context)
throws IOException, InterruptedException
{
        String name = "";
        double total = 0.0;
        int count = 0;
        for (Text t : values)
{
    String parts[] = t.toString().split("  ")
if (parts[0].equals("nhif"))
    {
        count++;
        total += Float.parseFloat(parts[1]);
}
        else  if (parts[0].equals("hosp"))
    {
            name = parts[1];
        }
    }
    String str = String.format("%d %f", count, total);
    context.write(new Text(name), new Text(str))
}
 }
```

Driver Map Reduce Code

```
public static void main(String[] args) throws Exception {
Configuration conf = new Configuration();
Job job = new Job(conf, "Reduce-side join");
job.setJarByClass(ReduceJoin.class);
job.setReducerClass(ReduceJoinReducer.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(Text.class);

MultipleInputs.addInputPath(job, new Path(args[0]),TextInputFormat.class, HospitalMapper.class)
MultipleInputs.addInputPath (job, new Path(args[1]),TextInputFormat.class, NhifMapper.class);
Path outputPath = new Path(args[2]);
FileOutputFormat.setOutputPath(job, outputPath);
outputPath.getFileSystem(conf).delete(outputPath);
System.exit (job.waitForCompletion (true) ? 0 : 1);

    }
```

```
}
```

Appendix 10: Use Case Scenario II Java Code

Map Reduce Program Java code snippet:

DiseaseCasesDriver Class

```java
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.FileInputFormat;
import org.apache.hadoop.mapred.FileOutputFormat;
import org.apache.hadoop.mapred.JobClient;
import org.apache.hadoop.mapred.JobConf;
import org.apache.hadoop.mapred.TextInputFormat;
import org.apache.hadoop.mapred.TextOutputFormat;

public class DiseaseCasesDriver {
        public static void main(String[] args) {
                JobClient my_client = new JobClient();
                // Create a configuration object for the job
                JobConf job_conf = new JobConf(DiseaseCasesDriver.class);
                // Set a name of the Job
                job_conf.setJobName("DiseaseCasesPerArea");
                // Specify data type of output key and value
                job_conf.setOutputKeyClass(Text.class);
                job_conf.setOutputValueClass(IntWritable.class);
                // Specify names of Mapper and Reducer Class
                job_conf.setMapperClass(DiseaseCasesMapper.class);
                job_conf.setReducerClass(DiseaseCasesReducer.class);
                // Specify formats of the data type of Input and output
                job_conf.setInputFormat(TextInputFormat.class);
                job_conf.setOutputFormat(TextOutputFormat.class);
                // Set input and output directories using command line arguments,
                //arg[0] = name of input directory on HDFS, and arg[1] =  name of output directory to be created
to store the output file.
                FileInputFormat.setInputPaths(job_conf, new Path(args[0]));
                FileOutputFormat.setOutputPath(job_conf, new Path(args[1]));
                my_client.setConf(job_conf);
                try {
                        // Run the job
                        JobClient.runJob(job_conf);
                } catch (Exception e) {
                        e.printStackTrace();
        }
    }
}
```

DiseaseCasesMapper Class

```
import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.*;

public class DiseaseCasesMapper extends MapReduceBase implements Mapper<LongWritable, Text, Text,
IntWritable> {
    private final static IntWritable one = new IntWritable(1);
    public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable> output, Reporter
reporter) throws IOException {
        String valueString = value.toString();
        String[] AreaDiseaseCases = valueString.split(",");
        String[] DiseaseCases = valueString.split(",");
        output.collect(new Text(AreaDiseaseCases[3]  +  DiseaseCases[7]), one);
    }
}
```

DiseaseCasesReducer Class

```
import java.io.IOException;
import java.util.*;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.*;
public class DiseaseCasesReducer extends MapReduceBase implements Reducer<Text, IntWritable, Text,
IntWritable> {
    public void reduce(Text t_key, Iterator<IntWritable> values, OutputCollector<Text,IntWritable> output,
Reporter reporter) throws IOException {
        Text key = t_key;
        int frequencyForDiseaseCasesReport = 0;
        while (values.hasNext()) {
            // replace type of value with the actual type of our value
            IntWritable value = (IntWritable) values.next();
            frequencyForDiseaseCasesReport += value.get();
```

```
                    }
                    output.collect(key, new IntWritable(frequencyForDiseaseCasesReport));
          }
}
```

Appendix 11: Use Case Scenario III Java Code

Map Reduce Code snippet

```
import java.io.BufferedReader;
import java.io.FileReader;
import java.io.IOException;
import java.net.URI;
import java.util.HashSet;
import java.util.Set;
import java.util.regex.Pattern;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.filecache.DistributedCache;
import org.apache.hadoop.fs.FileSystem;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
import org.apache.hadoop.util.Tool;
import org.apache.hadoop.util.ToolRunner;

public class LinesWithMatchingWordsJob implements Tool {
        private Configuration conf;
        public static final String NAME = "linemarker";
        public void setConf(Configuration conf) {
                this.conf = conf;
        }
        public Configuration getConf() {
        return conf;
        }
        public static void main(String[] args) throws Exception {
        if(args.length != 2) {
        System.err.println("Usage: linemarker <input> <output>");
        System.exit(1);
                }
        ToolRunner.run(new LinesWithMatchingWordsJob( new Configuration()), args);
        }
        public LinesWithMatchingWordsJob(Configuration conf) {
```

```
    this.conf = conf;
    }
    public int run(String[] args) throws Exception {
        DistributedCache.addCacheFile(new
Path("/user/cloudera/keywords/news_keywords.txt").toUri(), conf);
        Job job = new Job(conf, "Line Marker");
        job.setInputFormatClass(TextInputFormat.class);
        job.setOutputFormatClass(TextOutputFormat.class);
        job.setMapperClass(LineMarkerMapper.class);
        job.setNumReduceTasks(0);
        job.setOutputKeyClass(LongWritable.class);
        job.setOutputValueClass(Text.class);
        job.setJarByClass(LinesWithMatchingWordsJob.class);
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job,new Path(args[1]));
        return job.waitForCompletion(true) ? 1 : 0;
        }
    public static class LineMarkerMapper extends Mapper<LongWritable, Text, LongWritable, Text> {
        private Pattern space_pattern = Pattern.compile("[ ]");
        private Set<String> keywords = new HashSet<String>();

        @Override
        protected void setup(Context context) throws IOException, InterruptedException {
        URI[] uris = DistributedCache.getCacheFiles(context.getConfiguration());
        FileSystem fs = FileSystem.get(context.getConfiguration());
            if(uris == null || uris.length == 0) {
    throw new IOException("Error reading file from distributed cache. No URIs found.");
        }
        String localPath = "./keywords.txt";
        fs.copyToLocalFile(new Path(uris[0]), new Path(localPath));
        @SuppressWarnings("resource")
        BufferedReader reader = new BufferedReader(new FileReader(localPath));
        String word = null;
            while((word = reader.readLine()) != null) {
                keywords.add(word);
            }
        }
        @Override
protected void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
        String[] tokens = space_pattern.split(value.toString());
        for(String token : tokens) {
            if(keywords.contains(token)) {
```

```
                        context.write(key, new Text(token));
                        }
                    }
                }
            } }
```

RESEARCH OUTPUTS

The following are the list of research outputs:

(i) Modified MapReduce Algorithm Code Design Model

(ii) Developed Big Data Analytics Framework for Childhood Infectious Disease Framework

(iii) Publication

Mwamnyange, M., Luhanga, E., & Thodge, S. R. (2021). Big Data Analytics Framework for Childhood Infectious Disease Surveillance and Response System using Modified MapReduce Algorithm: A case Study of Tanzania. *International Journal of Advanced Computer Science and Applications*, 12 (3), 373-385.

(iv) Poster Presentation