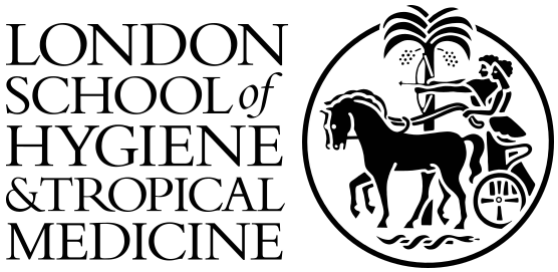

Big Data Analytics in Healthcare Internet of Things

Gunasekaran Manogaran,[1*] Daphne Lopez,[1] Chandu Thota,[2]
Kaja M. Abbas,[3] Saumyadipta Pyne,[4] and Revathi Sundarasekar[5]

[1] VIT University, School of Information Technology and Engineering, Vellore, India

[2] Albert Einstein Lab, Infosys Ltd, Hyderabad, India

[3] Department of Population Health Sciences, Virginia Tech, Blacksburg, USA

[4] Indian Institute of Public Health, Hyderabad, India

[5] Priyadarshini Engineering College, Vellore, India

[*] gunavit@gmail.com

Abstract

Nowadays, wearable medical devices play a vital role in many environments such as continuous health monitoring of individuals, road traffic management, weather forecasting, and smart home. These sensor devices continually generate a huge amount of data and stored in cloud computing. This chapter pro- poses Internet of Things (IoT) architecture to store and process scalable sensor data (big data) for healthcare applications. Proposed architecture consists of two main sub-architecture, namely, MetaFog-Redirection (MF-R) and Grouping & Choosing (GC) architecture. Though cloud computing provides scalable data storage, it needs to be processed by an efficient computing platforms. There is a need for scalable algorithms to process the huge sensor data and identify the useful patterns. In order to overcome this issue, this chapter proposes a scalable MapReduce-based logistic regression to process such huge amount of sensor data. Apache Mahout consists of scalable logistic regression to process large data in distributed manner. This chapter uses Apache Mahout with Hadoop Distributed File System to process the sensor data generated by the wearable medical devices.

Keywords

Internet of Things, Big data analytics, MetaFog-Redirection (MF-R), Grouping and Choosing (GC) architecture, Healthcare

1 Introduction

Data generation speed and the amount of data have increased over the past 20 years in different fields. A report published in 2011 by the International Data Corporation (IDC) states that the overall generated and stored data size in the globe was 1.8ZB (≈ 1021B), which enlarged by almost nine times within five years. Due to the enormous growth of world data, the name of big data is essentially used to express massive datasets. In general, big data analytics requires advanced tools and techniques to store, process, and analyze the large volume of data. Big data consists of large unstructured data that require advanced real-time analysis. Thus, nowadays, many of the researchers are interested in developing advanced technologies and algorithms to solve the issues when dealing with big data. In order to discover new opportunities and hidden values from big data, Yahoo developed the Hadoop-based tools and technologies to store and process the big data. Nowadays, private organizations are also interested in the high prospective of big data, and numerous government agencies declared vital ideas to speed up the big data research and applications. Two leading scientific journals such as *Nature and Science* also opened special issues to solve and discuss the challenges and impacts of big data. In recent years, big data plays a vital role in Internet companies such as Google, Facebook, and Twitter. For example, Google handles nearly 100 petabytes (PB) and Facebook produces log data of over ten petabytes per month. A modern Chinese company, Baidu, analyzes data of ten petabytes (PB), and Taobao, a subsidiary of Alibaba, produces data of ten petabytes (PB) for online trading per day.

"Big data" initially meant the volume, velocity, and variety of data that becomes tricky to analyze by using conventional data processing platforms and techniques. Nowadays, data production sources are improved rapidly, such as telescopes, sensor networks, high-throughput instruments, and streaming machines, and these environments generate massive amount of data. Nowadays, big data has been playing a crucial role in a variety of environments such as healthcare, business organization, industry, scientific research, natural resource management, social networking, and public administration. Big data can be categorized by 10Vs as follows:

Volume: The big volume indeed represents big data. Recently, the data generation sources are augmented, and it causes diversity of data such as text, video, audio, and large size images. In order to process the enormous amount of data, our conventional data processing platforms and techniques have to be enhanced.

Velocity: The rate of the incoming data has increased dramatically; this velocity indeed represents big data. The phrase velocity represents the data generation speed. The data explosion of the social media has changed and causes variety in data. Nowadays, people are not concerned in old post (a tweet, status updates, etc.) and notice most hot updates.

Variety: The variety of the data indeed represent big data. Nowadays, the collection of data types is also increased. For example, most organizations use the following type of data formats such as database, excel, and CSV, which can be stored

in a plain text file. Nevertheless, sometimes, the data may not be in the anticipated format, and it causes difficulties to process. In order to defeat this issue, the organization has to identify the data storage system which can analyze variety of data.

Value: The value of data indeed represents big data. Having continuous amounts of data is not helpful until it can be turned into value. It is essential to understand that it does not always mean there is value in big data. The benefits and costs of analyzing and collecting the big data are more important things when doing big data analytics.

Veracity: This veracity of data indeed represents big data. Veracity represents the data understandability; it doesn't represent data quality. It is significant that the association should perform data processing to prevent "dirty data" from accumulating in the systems.

Validity: It is essential to ensure whether the data is precise and accurate for future use. In order to take the right decisions in the future, the organizations should validate the data noticeably.

Variability: Variability refers to the data consistent and data value.

Viscosity: Viscosity is an element of velocity, and it represents the latency or lag time in data transmit between the source and destination.

Virality: Virality represents the speed of the data send and receives from various sources.

Visualization: Visualization is used to symbolize the big data in a complete view and determine the hidden values. Visualization is an essential key to making big data an integral part of decision-making.

2 Overview of the Smarter HealthCare System in Internet of Things (IoT)

Nowadays, development in wireless communication technologies has changed the traditional communication methods. In the last decade, man-to-man communication and man-to-machine communication are most often used in communication environments. The push toward the network communications has increased, and machine-to-machine communication is recently used in many platforms. IoT provides the platform to service with supporting communication among physical objects and virtual representations. IoT consists of various tools and technologies such as controllers, sensors, or low-powered wired and wireless services. In other words, the Internet of Things (IoT) is the wired or wireless interconnected various physical devices used to observe, communicate, and transfer information with their external environment or internal states. Nowadays, wireless mobile sensor network (WMSN) is most often used in continuous monitoring of healthcare applications, where patients are monitored with the help of sensor devices. This section describes the research work related to the healthcare systems using medical sensor networks

and wearable sensor devices. Harvard Sensor Network Lab recently developed the CodeBlue project, which aims to monitor the patients (Malan et al. 2004; Lorincz et al. 2004). CodeBlue project, several medical sensors are fixed on the patient's body to sense the patients' health conditions. In addition, these medical sensors continuously sense the patient body and transmit the health conditions to the end-user devices (laptops, PDAs, and personal computers) using wireless technologies. These data are generally used for finding the useful patterns to protect the patients from emergency situations. The main function of CodeBlue is very simple: a medical professional or doctor issues a query for patient healthcare data using their personal digital assistant (PDA), which works based on the publish and subscribe architecture. Finally, the collected data from the medical sensors are publishing to a specific channel and end-user need to subscribe that channel by using their laptop and PDA (Kumar and Lee 2011). In addition, Wood et al. (2006) from the University of Virginia have developed the heterogeneous network architecture named AlarmNet (Wood et al. 2006). The goal of this project is to monitor the patient health in the home and assisted living environment. More similarly, AlarmNet consists of environmental sensor networks and body sensor networks to efficiently sense the specific data. Three network tiers are used in this framework to sense the specific data in home and assisted living environment. In the first tier, a patient wears a variety of body sensor devices such as accelerometer, ECG, and SpO2, which sense individual physiological (health) data, whereas in the second tier, environmental sensors such as dust, temperature, motion, and light (i.e., MicaZ boards) are fixed in the living space to sense the range of environmental conditions. Finally, in the third tier, an Internet Protocol (IP)-based network is made available which is comprised of Stargate gateways called AlarmGate. AlarmNet has used the body sensor devices to broadcast the individual physiological data from single hop to the second tier (i.e., nearest stationary sensor). Once the physiological data is received by the second tier, the stationary sensors forward the physiological data using shortest path first routing protocol (i.e., multi-hop communication) to the AlarmGate. The AlarmGate works as a gateway between the IP networks and wireless sensor nodes and is also attached to a back-end server. Ng et al. (2004) have developed the ubiquitous monitoring environment for wearable and implantable sensors (UbiMon) (Ng et al. 2004). This project is a type of body sensor network (BSN) architecture composed of implantable and wearable sensors using the wireless ad hoc network. The main goal of this project is to provide continuous monitoring of patient's health status and also predict the emergency conditions. In addition, Chakravorty (2006) have developed a mobile healthcare project called MobiCare. This project is used to provide continuous and timely monitoring of individual's physiological status.

MobiCare project possibly saves many patients' lives and quality of patient care. MobiCare project consists of wearable sensors such as ECG, SpO2, and blood oxygen to monitor the patients. This project timely senses the patient's body and transfers the health status to the MobiCare client. In order to send the BSN data to the server, MobiCare client uses HTTP POST protocol. In addition, MobiCare server is also used to perform off-line physiological analysis and supports to the medical staffs for patient care (Chakravorty 2006).

Blum and Magill (2010) have proposed a personalized ambient monitoring (PAM) project used to monitor the patient's mental health (Blum and Magill 2010). The goal of the PAM project is to monitor the day-to-day activity of patients with bipolar disorder (BP). Various Bluetooth protocols are used to join the mobile phones and body sensors; thereafter Bluetooth also connects the personal computers and mobile phones. The goal of the mobile phones is to aggregate the body sensors data and send it to the personal computers for storage and analysis.

3 Big Data in Healthcare

In recent decades, big data analytics also impact more in healthcare. Nowadays, healthcare systems are rapidly adopting clinical data, which will rapidly enlarge the size of health records that are accessible, electronically (Lopez and Sekaran 2016). Concurrently, fast progress and development have achieved in modern healthcare management system (Lopez et al. 2014). A recent study expounds six use cases of big data to decrease the cost of patients, triage, readmissions, adverse events, and treatment optimization for diseases affecting multiple organ systems (Bates et al. 2014). In yet another study, big data use cases in healthcare have been divided into number of categories such as clinical decision support (with a subcategory of clinical information), administration and delivery, consumer behavior, and support services (Lopez and Manogaran 2016). Jee et al. described how to reform the healthcare system based on big data analytics to choose appropriate treatment path, improvement of healthcare systems, and so on (Jee and Kim 2013; Lopez and Gunasekaran 2015; Manogaran and Lopez 2017). The above use cases have utilized the following big data in healthcare implementation (Chandu Thota, Gunasekaran Manogaran, Revathi Sundarsekar, V. Vijayakumar, "Big Data Security Framework for Distributed Cloud Data Centers," 2016). (1) Patient-centered framework is produced based on the big data framework to approximate the amount of healthcare (cost), patient impact (outcomes), and dropping readmission rates (Chawla and Davis 2013). (2) Virtual physiological human analysis framework is combined with big data analytics to create robust and valuable solutions in silico medicine (Viceconti et al. 2015).

3.1 Digital Epidemiology

Digital epidemiology enables real-time disease surveillance through novel analysis of digital data (Salathe et al. 2012). Meaningful understanding and analysis of digital sources, such as social media, are critical to improve real-time disease surveillance and enable significant public health solutions. Digital epidemiology complements traditional epidemiological studies, such as case records, case reports, ecological studies, cross-sectional studies, case control studies, cohort studies, randomized controlled trials, and systematic reviews and meta-analysis. While data of

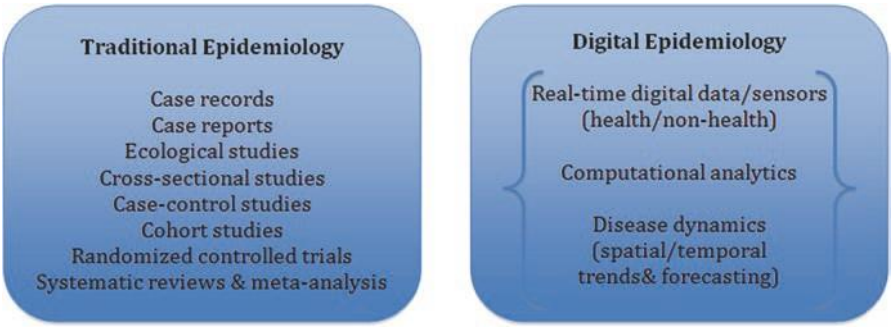

Fig. 10.1 Digital and traditional epidemiology. Digital epidemiology complements traditional epidemiology, by conducting computational analytics of real-time health- and non-health-related digital data and sensors to derive real-time estimates of disease dynamics

study participants are primarily collected by traditional epidemiological studies to address a research question of clinical and/or public health significance, digital epidemiology also makes use of data sources that are originally collected and/or generated for health- and non-health-related purposes. Figure 10.1 illustrates the methods of digital and traditional epidemiology.

ChatterGrabber is a social media surveillance software toolkit to identify potential health risks and disease outbreaks, by analyzing tweets for disease symptoms in specific locations (Schlitt et al. 2015). This software toolkit is used for disease surveillance in different applications, including the EpiDash application to monitor norovirus outbreaks. Google search queries have been analyzed for surveillance of infectious diseases, such as in Google Flu Trends and Google Dengue Trends for influenza and dengue, respectively (Ginsberg et al. 2009). These digital surveillance systems act as early warning systems for infectious disease outbreaks and complement traditional disease surveillance systems that have a lag time in collection and dissemination of estimates of disease burden. HealthMap provides real-time infectious disease surveillance by analyzing online sources of news (Google News), email list serves (ProMED-mail), and information provided by global health organizations (WHO, OIE, FAO) (Freifeld et al. 2008). It is an automated monitoring tool for infectious disease outbreaks affecting human and animal health.

3.2 FRED Software for Disease Modeling

FRED (a Framework for Reconstructing Epidemiological Dynamics) is a disease modeling software used to handle huge amount of data (synthetic population). FRED uses mitigation strategies, viral evolution, and personal health behaviors to model the disease outbreak (Grefenstette et al. 2013). FRED is an open source framework for epidemic modeling, rather than a model of a particular infectious disease. Geographic regions are used in FRED to represent every individual as

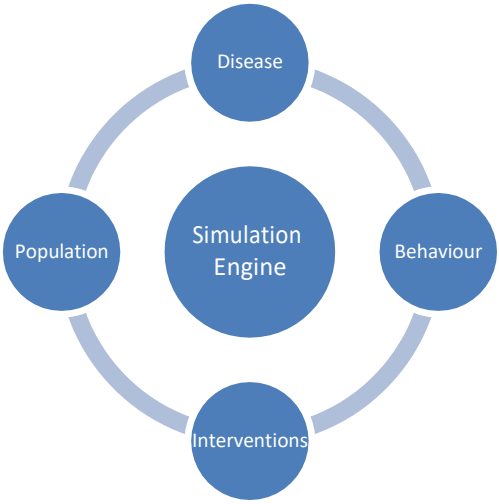

Fig. 10.2 Framework of the FRED

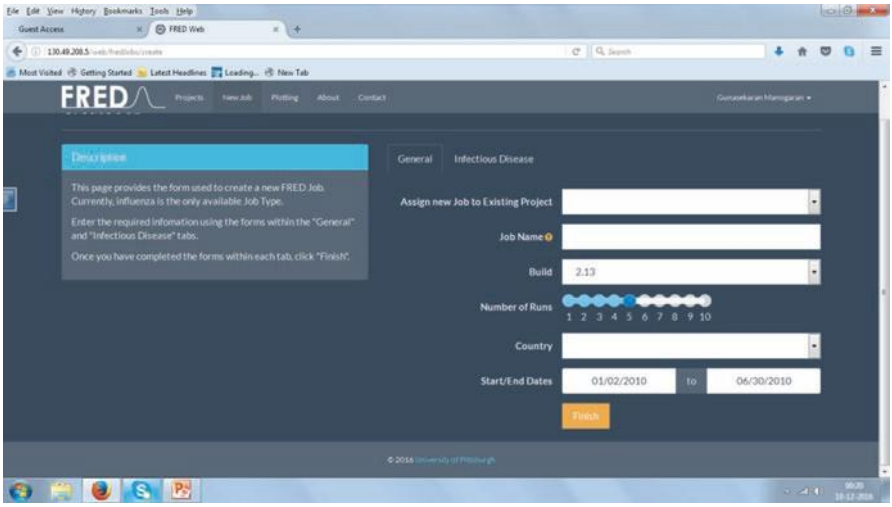

Fig. 10.3 Project creation in FRED

agent. Each agent has a set of sociodemographic characteristics and daily behaviors that include age, sex, employment status, occupation, and household location and membership in a set of social contact networks. This synthetic population data is used to model the disease outbreak in FRED. The FRED framework is depicted in Fig. 10.2. Figure 10.3 represents the graphical user interface for the FRED software. Figure 10.4 represents the output variables for disease modeling of H1N1 influenza. Figures 10.5 and 10.6 represent the susceptible, exposed, infected, and recover (SEIR) model simulation results of H1N1 influenza.

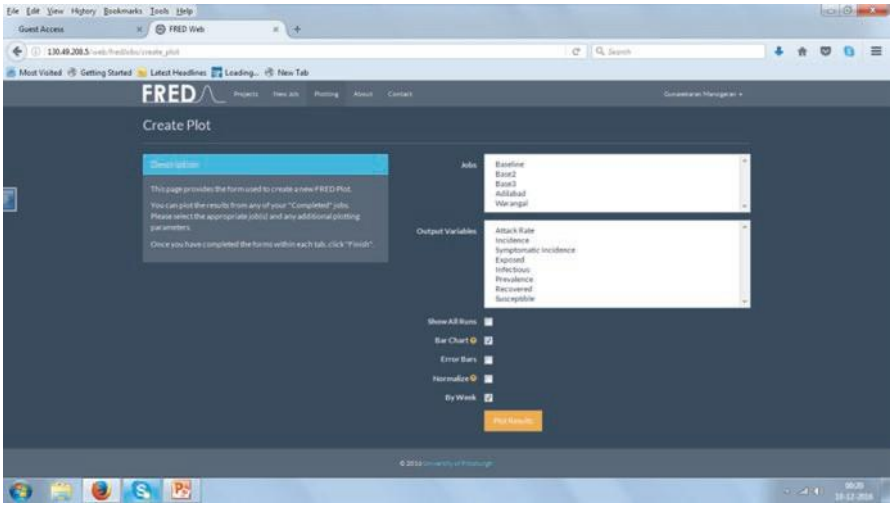

Fig. 10.4 Output variable specification in FRED

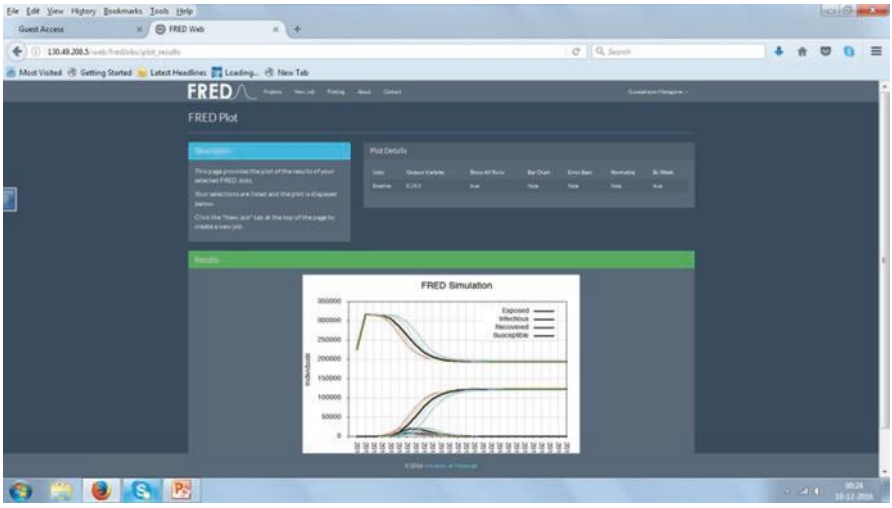

Fig. 10.5 SEIR model simulation

3.3 NoSQL Databases for Big Data in Healthcare

NoSQL database is used to store huge volume of data in distributed manner. A NoSQL database does not follow any relational schema. NoSQL databases can be classified into four types such as key-value stores, column family database stores, document stores, and graph stores. The difference between EHR requirements and NoSQL database features is depicted in Table 10.1.

3.3.1 Key-Value Stores for Storing Big Clinical Data

Key-value databases store the data based on the key and value pairs. In general, key-value databases assume that the path is the key and the contents are the file. Key-value databases are applicable only for small applications not for complex applications. Key-value database for storing clinical data is shown in Fig. 10.7.

3.3.2 Column Family Stores for Storing Big Clinical Data

Column family database stores the huge data into rows as collections of columns. All the rows in this database consist of number of columns. Column family database for storing clinical data is shown in Fig. 10.8.

3.3.3 Document Stores for Storing Big Clinical Data

A document store databases are used to store huge data related to document format. This type of database is usually used to store semi-structured data. Document store database for storing clinical data is shown in Fig. 10.9.

3.3.4 Graph Stores for Storing Big Clinical Data

Graph databases consist of connections, or edges, between nodes. It is a type of NoSQL database that uses graph theory to store, map, and query relationships. Graph database for storing clinical data is shown in Fig. 10.10. Figure 10.11 shows the various NoSQL databases, and Table 10.2 compares the various types of NoSQL databases.

4 Proposed Architecture for Healthcare Internet of Things

Electronic medical record (EMR) is comprised of patient health-related information. The following information generally available in all EHRs are laboratory results, billing data, medication records, and test details. In most of the cases, laboratory results and billing data are available as structured “name-value pair” data. Recently, more number of researches is trying to develop big data-based electronic phenotype algorithms to identify diseases from the EHR. Laboratory data and vital signs are mostly in the structured format. It follows coding scheme to store the huge amount of lab-related data. Nowadays, many dictionaries and various algorithms are developed to reduce the complexity of laboratory data.

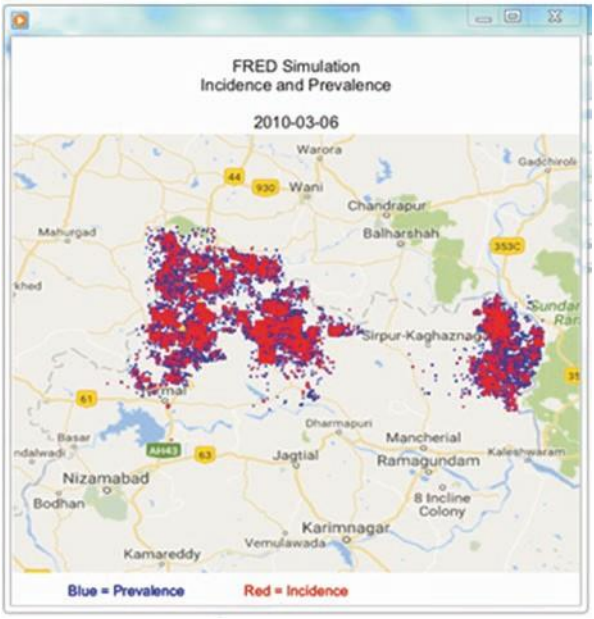

Fig. 10.6 Geospatial visualization of FRED result

Table 10.1 Difference between EHR requirements and NoSQL database features

EHR requirement	NoSQL database feature
Healthcare data size increasing over time, so it needs scalability	Automatic scaling is available in NoSQL databases
Structure of data is varied over time, so it needs new solutions	NoSQL databases allow unstructured or semi-structured data to be stored easily
Healthcare data should always be accessible for continuity of healthcare services	High availability is one of the main feature of NoSQL databases
Generally healthcare data is added continuously	Eventual consistency suggested by NoSQL database architecture is considered acceptable for EHR use cases
Healthcare data should be available to multiple locations which require a high-performance system for high-speed data access	NoSQL databases offer higher performance in many use cases

In order to store the huge amount of healthcare data, proposed architecture has used NoSQL-based database. High availability of healthcare information has led to increase the accuracy and overall quality of healthcare delivery. Nowadays, the size and structure of healthcare data are increasing dramatically. Hence, relational database management system is not suitable for storing such huge size of data. Researchers develop a number of big data technologies to solve such issues. NoSQL databases have significant advantages such as auto scaling, better performance, and high availability which address the limitations of relational databases in distributed

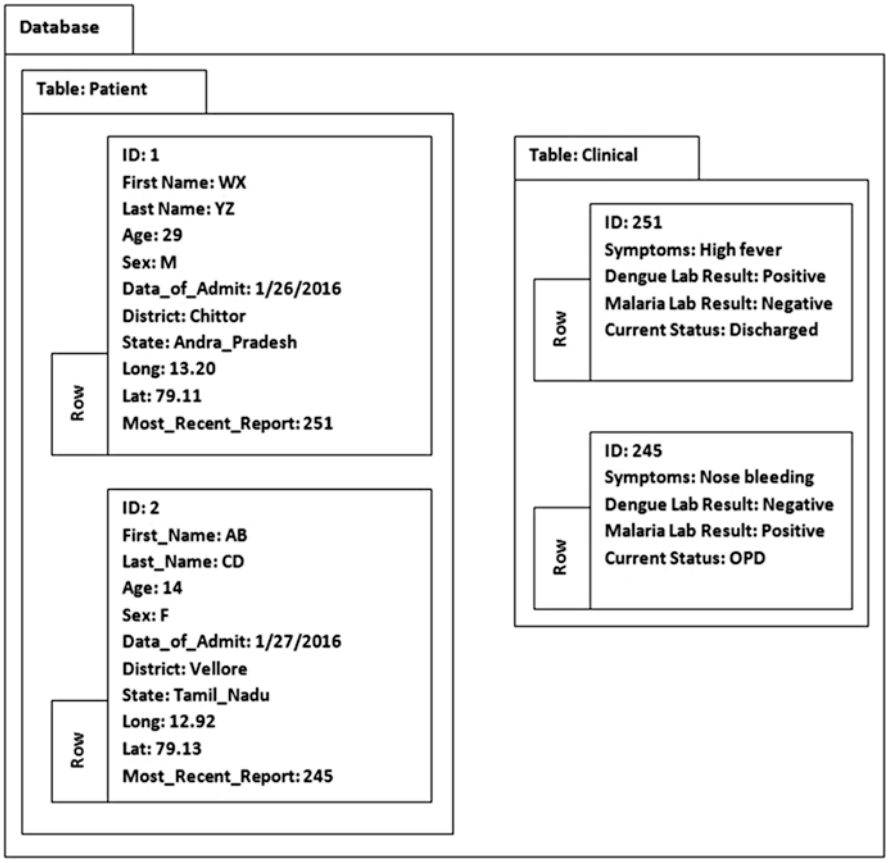

Fig. 10.7 Key-value database structure for storing clinical data

healthcare systems. Scalable sensor data processing architecture is proposed in this chapter to store and process body sensor data for healthcare applications (Fig. 10.12). In this proposed architecture, electronic medical records are collected through clinical test, and the results are stored into cloud storage (Amazon S3). MapReduce implementation of online stochastic gradient descent algorithm is used in the logistic regression to develop the prediction model. Prior electronic medical records are used to train the logistic regression model. After completion of training process, the prediction model will use the current sensor data (blood pressure, blood sugar level, and heart rate) of the patient to predict the heart disease status.

The proposed architecture is used for the personal health monitoring of individuals. Whenever the respiratory rate, heart rate, blood pressure, body temperature, and blood sugar exceed its normal value, then the device sends an alert message with clinical value to the doctor using wireless network through fog computing. After successfully identified the authorized user, health data securely transfer to different data centers provided by different cloud data service providers. Meta Cloud Data

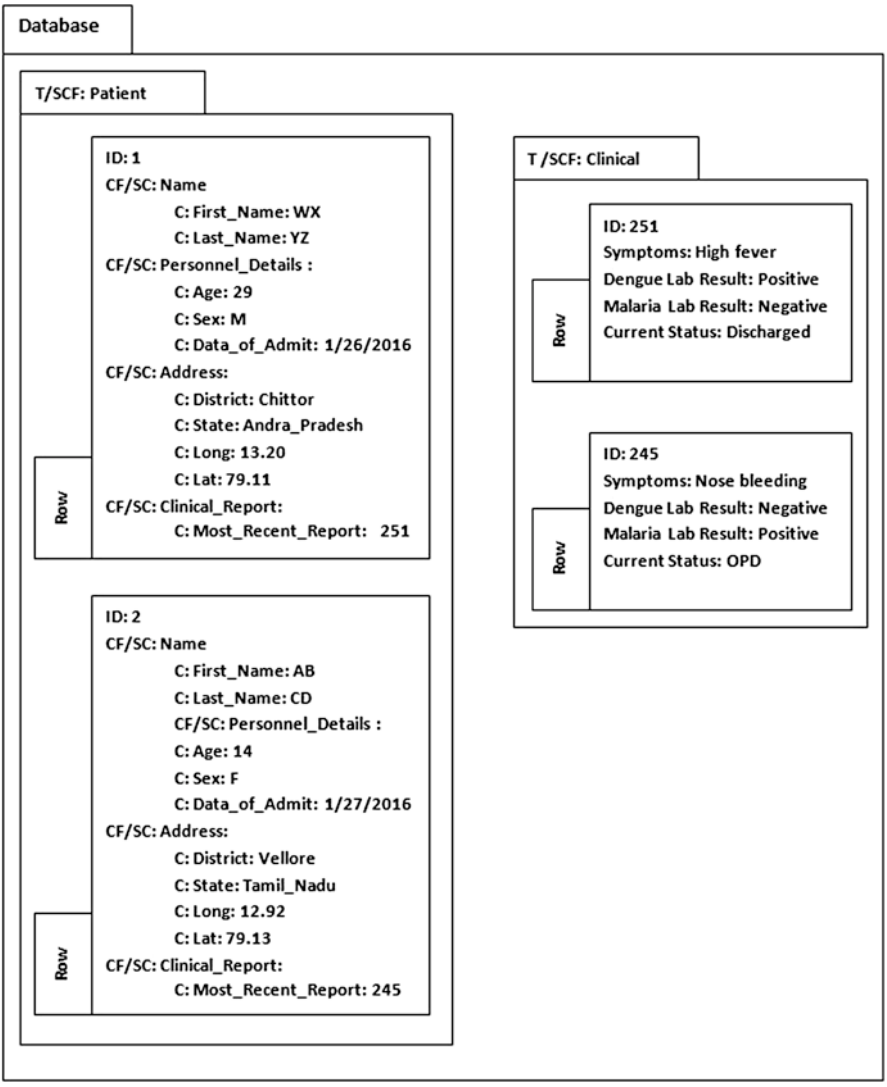

Fig. 10.8 Column family database structure for storing clinical data

Storage architecture is used to transfer the data from applications to cloud data centers and cloud data centers to applications.

Once the data is transferred from applications to the cloud data centers, it is required to be stored efficiently. Nowadays, data generation sources are increased such as high-throughput instruments, telescopes, sensor networks, and streaming machines, and these environments produce huge amount of data. In order to solve

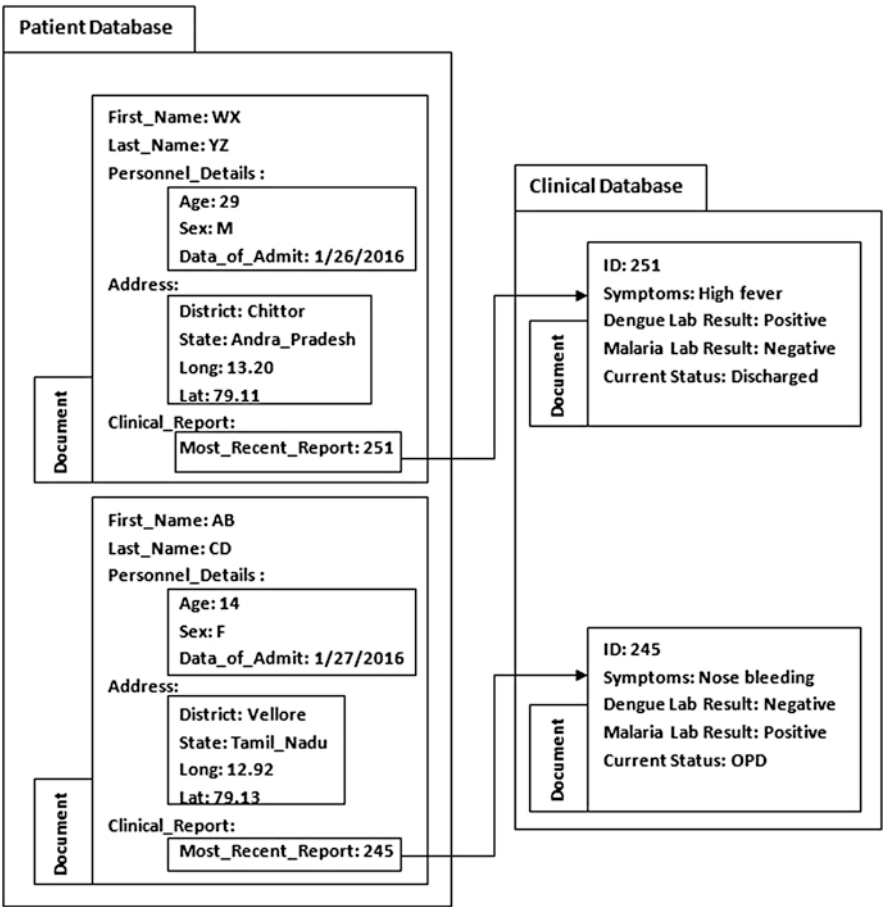

Fig. 10.9 Document store database structure for storing clinical data

this issue, Hadoop Distributed File System (HDFS) is used in this phase to store such huge amount of data. This phase also categorizes the data into different levels and stores them into different data centers.

Grouping & Choosing (GC) architecture is embedded with MetaFog-Redirection architecture to secure integration of fog to cloud computing and protect big data against intruder (Manogaran et al. 2017a). Clinical data is stored in multiple cloud data centers based on importance and categorization (Manogaran et al. 2017b). Data categorization is classified into three levels such as sensitive, critical, and normal. Each categorized data is supposed to be stored in different data center. Proposed architecture is used to redirect the user request to the appropriate data center.

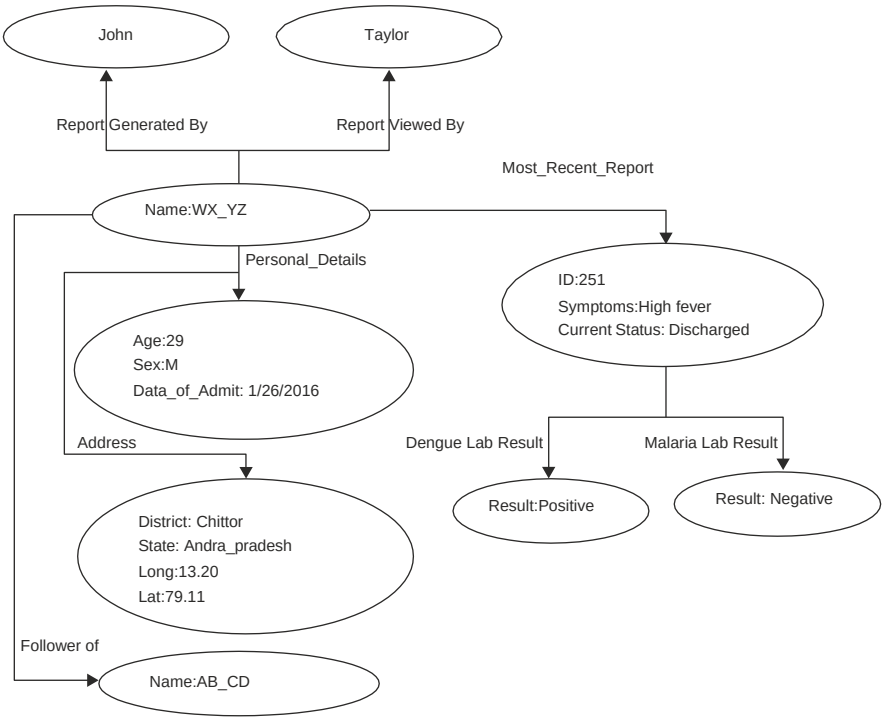

Fig. 10.10 Graph database structure for storing clinical data

MetaFog-Redirection (MF-R) architecture with Grouping & Choosing (GC) architecture (Fig. 10.13) is proposed in this chapter. The goal of the proposed GC architecture is to integrate the fog computing with cloud computing. The following integrations are performed in the GC architecture; it includes application integration, data transfer from fog servers to cloud data centers, and security mechanisms for communication between fog computing and cloud computing.

5 Big Data Analytics Using MapReduce

Logistic regression is used in the proposed MetaFog architecture to predict the disease based on the historical records. Logistic regression is often used in dataset where the dependent variable is dichotomous. Logistic regression is used in this chapter to develop the prediction model and find the relationship between dependent and independent variables. MapReduce implementation of stochastic gradient descent with logistic regression is shown below:

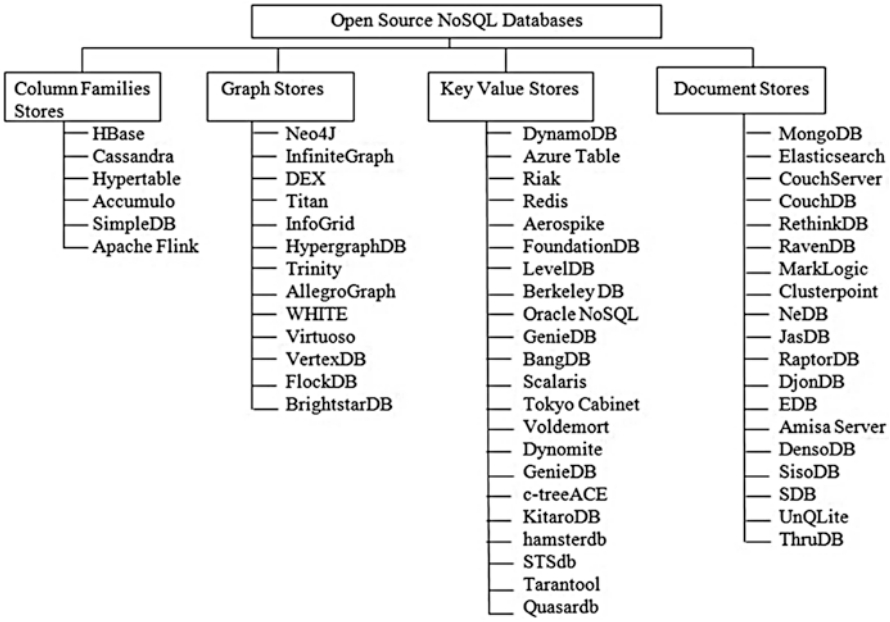

Fig. 10.11 Types of open source NoSQL databases

```
Input: Set V ={(a_i ϵ R^d ; b_i ϵ R)}^n_{i=1} of n clinical records
    θ^t ϵ R^d = Weights at time t
    n = Learning rate
Output: ^(t+1) R^d = Weights at time t+1
Step 1: Map Algorithm:
class MAPPER
     method INITIATIVE
    double θ=1;
            double n=0.1;
    # key k; value v;
method MAP (string; dounle v);
    begin
    (a_i ; b_i) ← (k; v)
            r=rand( ) #number of reducer
      EmitIntermediate (r;( a_i ; b_i );
    end
Step 2: Reduce Algorithm:
class REDUCER
     # key k; value v;
method REDUCE (string k;double v [1;2;...;r])
    begin
```

Table 10.2 Different NoSQL databases and its comparison

SNo	NoSQL DB	Data model	Usage	Strength	Weakness	Example
1	Key-value stores	Collection of key-value pairs	Briskly changing data and high availability (e.g., Stock Market Analysis)	Fast lookups	Stored data has no schema	Riak, Redis, Azure, Table Storage, Amazon, Simple DB
2	Column family stores	Column families	Read/write extensive applications (e.g., social networking)	Fast lookups	Very low-level API	HBase, Cassandra
3	Document stores	Collection of key-value connections	Working with occasionally changing/consistent data (e.g., CRM systems)	Incomplete data tolerant	Query performance, no standard query syntax	CouchDB, MongoDB
4	Graph stores	Property graphs—nodes	Spatial data storage (e.g., geographical information system)	Graph algorithm—shortest paths, connected ness, etc.	Not easy to cluster, travers whole graph to get answer	InfoGrid, Infinite Graph, Neo4J

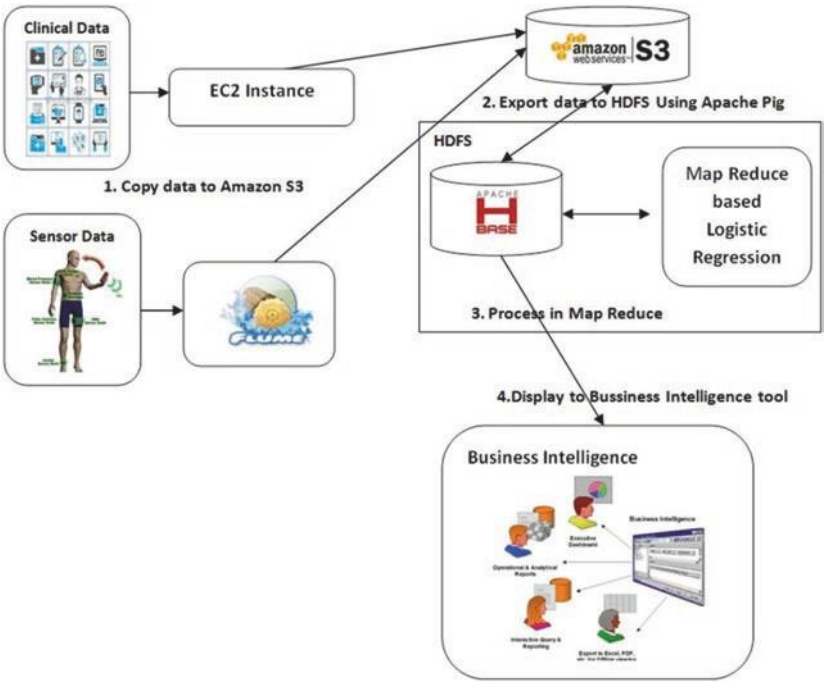

Fig. 10.12 Scalable sensor data processing architecture in the cloud

$$k^{q^{(t+1)}} \neg q^{t};$$

for each $(a_i; b_i)$ $\epsilon v[1;2;\ldots;r]$ *do*

$$k^{q^{(t+1)}} \neg k^{q^{(t+1)}} - n\tilde{\mathrm{N}}F\left(k^{q^{(t+1)}}\right);$$

end

$$return\left(string\, k; double\, k^{q^{t+1}}\right);$$

end

Set $V = \left\{\left(a \square R^{d}; b \square R\right)\right\}_{i=1}^{n}$ of n clinical records; each machine receives M blocks:

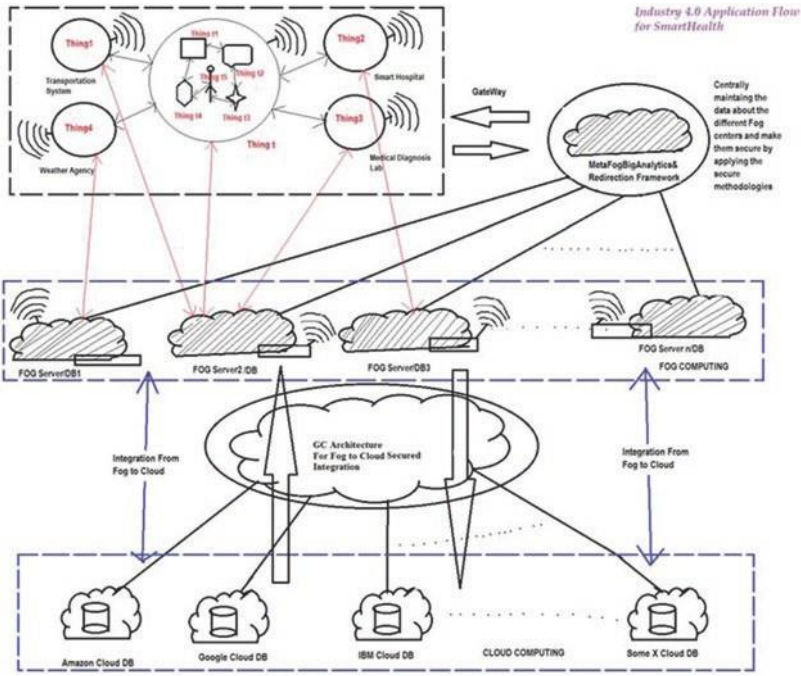

Fig. 10.13 MetaFog-Redirection (MF-R) architecture with Grouping & Choosing (GC) architecture

where $M = \frac{|V|}{m}$. Assume that m is the number of systems in the cloud. In MapReduce framework, M is automatically assigned by mappers (Kang et al. 2014). After splitting into *M* block sizes i^{th} reducer get V_i value. Stochastic gradient descent (SGD) method is used in the reducer function to get the efficient weights and reduce the error function. In the above algorithm, MAP and REDUCE method runs T iterations. Each iteration generates weights $k^{q^{(t+1)}}$ where $k \epsilon [1; 2; ..; M]$. The final weight θ^{t+1} is calculated based on the average of $k^{q^{(t+1)}}$. The average of weight θ^{t+1} can be defined by $q^{t+1} = \frac{1}{M} \sum_{k=1}^{M} k^{q^{(t+1)}}$.

5.1 Mahout Implementation of SGD for Logistic Regression

Apache Mahout is a library of scalable machine learning algorithms. Apache Mahout is implemented on top of Apache Hadoop and using the MapReduce paradigm. Machine learning is a type of artificial intelligence focused on enabling machines to learn without being explicitly programmed, and it is commonly used to

Table 10.3 Cleveland Heart Disease Database (CHDD) for training the logistic regression

SNo	Attribute	Range	Description
1	Age	Continuous	Age in years
2	Sex	0–1	Sex 0 = female; sex 1 = male
3	Cp	1–4	Chest pain type (Cp 1 = typical angina; Cp 2 = atypical angina; Cp 3 =non-anginal pain; Cp 4 = asymptomatic)
4	Trestbps	Continuous	Resting blood pressure (in mm Hg)
5	Chol	Continuous	Serum cholesterol in mg/dl
6	Fbs	0–1	(Fasting blood sugar >120 mg/dl) (Fbs 1 = true; Fbs 0 = false)
7	Restecg	0–2	Resting electrocardiographic results (Restecg 0 = normal; Restecg 1 = having ST-T wave abnormality; Restecg 2 = showing probable or define left ventricular hypertrophy by Estes' criteria)
8	Thalach	Continuous	Maximum heart rate achieved
9	Exang	0–1	Exercise induced angina (Exang 1 = yes; Exang 0 = no)
10	Old peak	Continuous	ST depression induced by exercise relative to rest
11	Slope	1–3	The slope of the peak exercise ST segment (Slope 1 = upsloping; Slope 2 = flat; Slope 3= downsloping)
12	Ca	0–3	Number of major vessels (0–3) colored by fluoroscopy
13	Thal	3–7	(Thal 3 = normal; Thal 6 = fixed defect; Thal 7 = reversible defect)
14	Num	0–1	Diagnosis classes (Num 0 = healthy; Num 1 = patient who is subject to possible heart disease)

improve future performance based on previous outcomes. Big data is stored on the HDFS; Apache Mahout is used to execute machine learning algorithms that extract meaningful patterns from datasets. The abovementioned MapReduce-based logistic regression can be done with the help of Apache Mahout. Mahout implementation of logistic regression using SGD supports the following command lines:

- Training the model
- Mahout org.apache.mahout.classifier.sgd.TrainLogistic – passes 1 – rate 1 – lambda
- 0.5 –input heart.csv – features 21 – output heart.model – target Num – categories 2 –
- Predictors Thalach Trestbps Fbs – types n.
- Testing the model

  Mahout org.apache.mahout.classifier.sgd.RunLogistic – input heart.csv – model. heart.model – auc – scores – confusion

5.2 Model Development with Cleveland Heart Disease Database (CHDD)

Logistic regression is trained using the prior clinical records and sensor data of the patients. The prediction model can use current sensor data (blood pressure, blood sugar level, and heart rate) of the patient to predict the heart disease status. In this

Table 10.4 Contingency table of Cleveland Heart Disease Database (CHDD)

Health measurement	Variable name in Cleveland Heart Disease Database (CHDD)	Threshold level in wearable sensor	Disease	
			Yes	No
Heart rate	Thalach	Low (<60)	0	0
		Medium (60–100)	7	1
		High (>100)	163	132
Blood pressure	Trestbps	Low (<60)	0	0
		Medium (60–140)	100	137
		High (>140)	39	27
Blood sugar	Fbs	Low (<70)	0	0
		Medium (70–120)	117	141
		High (>120)	22	23

analysis the prediction model uses the current sensor data obtained from body sensor devices through cloud and big data technologies. Cleveland Heart Disease Database (CHDD) is the de facto database for heart disease research (Manogaran 2017). This database is used to train the proposed prediction model, and it contains 76 attributes; but all published experiments refer to using a subset of 14 of them. The Cleveland database is widely used by the machine learning (ML) researchers till date. Table 10.3 depicts the number of attributes, description, and its range. Contingency table is depicted in Table 10.4.

6 Conclusion

This chapter proposes Internet of Things (IoT) architecture to store and process scalable sensor data (big data) for healthcare applications. Proposed architecture is consists of two main sub-architecture, namely, MetaFog-Redirection (MF-R) and Grouping & Choosing (GC) architecture. MapReduce-based logistic regression is implemented with the help of Apache Mahout. Logistic regression is trained using the prior clinical records from the Cleveland Heart Disease Database (CHDD) and sensor data of the patients. The prediction model can use current sensor data (blood pressure, blood sugar level, and heart rate) of the patient to predict the heart disease status. In this analysis, the prediction model uses the current sensor data obtained from body sensor devices through the cloud and big data technologies.

References

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs, 33*(7), 1123–1131.

Blum, J. M., & Magill, E. (2010). The design and evaluation of personalised ambient mental health monitors. In 7th Annual IEEE Consumer Communications and Networking Conference (pp. 1–5). Institute of Electrical and Electronics Engineers (IEEE).

Chakravorty, R. (2006). A programmable service architecture for mobile medical care. In Fourth Annual IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOMW'06) (pp. 5-pp). IEEE.

Chawla, N. V., & Davis, D. A. (2013). Bringing big data to personalized healthcare: A patient-centered framework. *Journal of General Internal Medicine, 28*(3), 660–665.

Freifeld, C. C., Mandl, K. D., Reis, B. Y., & Brownstein, J. S. (2008). HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association, 15*(2), 150–157.

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.

Grefenstette, J. J., Brown, S. T., Rosenfeld, R., DePasse, J., Stone, N. T., Cooley, P. C., ..., Guclu, H. (2013). FRED (A Framework for Reconstructing Epidemic Dynamics): An open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC Public Health, 13*(1), 1.

Jee, K., & Kim, G. H. (2013). Potentiality of big data in the medical sector: Focus on how to reshape the healthcare system. *Healthcare Informatics Research, 19*(2), 79–85.

Kang, D., Lim, W., Shin, K., Sael, L., Kang, U. (2014). Data/feature distributed stochastic coordinate descent for logistic regression. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (pp. 1269–1278). ACM.

Kumar, P., & Lee, H. J. (2011). Security issues in healthcare applications using wireless medical sensor networks: A survey. *Sensors, 12*(1), 55–91.

Lopez, D., & Gunasekaran, M. (2015). Assessment of vaccination strategies using fuzzy multi-criteria decision making. In *Proceedings of the Fifth International Conference on Fuzzy and Neuro Computing (FANCCO-2015)* (pp. 195–208). Springer International Publishing.

Lopez, D., & Manogaran, G. (2016). Big data architecture for climate change and disease dynamics. In G. S. Tomar, N. S. Chaudhari, R. S. Bhadoria, & G. C. Deka (Eds.), *The human element of big data: Issues, analytics, and performance*. Boca Raton: CRC Press, Taylor & Francis.

Lopez, D., & Sekaran, G. (2016). Climate change and disease dynamics-a big data perspective. *International Journal of Infectious Diseases, 45*, 23–24.

Lopez, D., Gunasekaran, M., Murugan, B. S., Kaur, H., Abbas, K. M. (2014). Spatial big data analytics of influenza epidemic in Vellore, India. In Big Data (Big Data), 2014 IEEE International Conference on (pp. 19–24). IEEE.

Lorincz, K., Malan, D. J., Fulford-Jones, T. R., Nawoj, A., Clavel, A., Shnayder, V., et al. (2004). Sensor networks for emergency response: Challenges and opportunities. *IEEE Pervasive Computing, 3*(4), 16–23.

Malan, D., Fulford-Jones, T., Welsh, M., Moulton, S. (2004). Codeblue: An ad hoc sensor network infrastructure for emergency medical care. In International workshop on wearable and implantable body sensor networks, *5*.12–15

Manogaran, G., Lopez, D. (2017a). Health data analytics using scalable logistic regression with stochastic gradient descent, *International Journal of Advanced Intelligence Paradigms*, *8*(2).

Manogaran, G., & Lopez, D. (2017b). Disease surveillance system for big climate data processing and dengue transmission. *International Journal of Ambient Computing and Intelligence (IJACI)*, *8*(2), 88–105.

Manogaran, G., & Lopez, D. (2017c). Spatial cumulative sum algorithm with big data analytics for climate change detection. *Computers and Electrical Engineering*. http://dx.doi.org/10.1016/j.compeleceng.2017.04.006.

Manogaran, G. C. T., Lopez, D., Vijayakumar, V., Abbas, K. M., & Sundarsekar, R. (2017a). Big data knowledge system in healthcare. In C. Bhatt, N. Dey, & A. Ashour (Eds.), *Internet of things and big data technologies in next generation healthcare, studies in big data series*. Switzerland: Springer International Publishing.

Manogaran, G., Thota, C., & Sundarsekar, R. (2017b). Big data security intelligence for healthcare industry 4.0. In L. Thames & D. Schaefer (Eds.), *Cybersecurity for industry 4.0*. Switzerland: Springer International Publishing.

Ng, J. W., Lo, B. P., Wells, O., Sloman, M., Peters, N., Darzi, A., ... ,Yang, G. Z. (2004, September). Ubiquitous monitoring environment for wearable and implantable sensors (UbiMon). In International Conference on Ubiquitous Computing (Ubicomp).

Salathe, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., et al. (2012). Digital epidemiology. *PLoS Computational Biology*, *8*(7), e100216.

Schlitt, J. T., Lewis, B., & Eubank, S. (2015). ChatterGrabber: A lightweight easy to use social media surveillance toolkit. *Online Journal of Public Health Informatics*, *7*(1), 52–53.

Thota, C., Manogaran, G., Sundarsekar, R., & Vijayakumar V. (2016). Big data security framework for distributed cloud data centers. In M. Moore (Ed.), *Cybersecurity Breaches and issues surrounding online threat protection*. IGI Global

Viceconti, M., Hunter, P., & Hose, R. (2015). Big data, big knowledge: Big data for personalized healthcare. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1209–1215.

Wood, A., Virone, G., Doan, T., Cao, Q., Selavo, L., Wu, Y., ... ,Stankovic, J. (2006). Alarm-net: Wireless sensor networks for assisted-living and residential monitoring. University of Virginia Computer Science Department Technical Report, 2.