IET Cyber-Physical Systems: Theory & Applications

Review Article

Remote health care cyber-physical system: quality of service (QoS) challenges and opportunities

ISSN 2398-3396
Received on 20th October 2016
Revised on 4th November 2016
Accepted on 8th November 2016
doi: 10.1049/iet-cps.2016.0023
www.ietdl.org

Tejal Shah[1], Ali Yavari[2], Karan Mitra[3], Saguna Saguna[3], Prem Prakash Jayaraman[4], Fethi Rabhi[1], Rajiv Ranjan[5] ✉

[1]School of Computer Science and Engineering, The University of New South Wales, Sydney, Australia
[2]School of Science, RMIT University, Melbourne, Australia
[3]Department of Computer Science, Electrical and Space Engineering, Lulea University of Technology, Skellefteå, Sweden
[4]School of Software and Electrical Engineering, Swinburne University of Technology, Australia
[5]School of Computing, Newcastle University, Newcastle, UK
✉ *E-mail: rranjans@gmail.com*

Abstract: There is a growing emphasis to find alternative non-traditional ways to manage patients to ease the burden on health care services largely fuelled by a growing demand from sections of population that is ageing. In-home remote patient monitoring applications harnessing technological advancements in the area of Internet of things (IoT), semantic web, data analytics, and cloud computing have emerged as viable alternatives. However, such applications generate large amounts of real-time data in terms of volume, velocity, and variety thus making it a big data problem. Hence, the challenge is how to combine and analyse such data with historical patient data to obtain meaningful diagnoses suggestions within acceptable time frames (considering quality of service (QoS)). Despite the evolution of big data processing technologies (e.g. Hadoop) and scalable infrastructure (e.g. clouds), there remains a significant gap in the areas of heterogeneous data collection, real-time patient monitoring, and automated decision support (semantic reasoning) based on well-defined QoS constraints. In this study, the authors review the state-of-the-art in enabling QoS for remote health care applications. In particular, they investigate the QoS challenges required to meet the analysis and inferencing needs of such applications and to overcome the limitations of existing big data processing tools.

1 Introduction

Our population is expanding exponentially due to increased life expectancy coupled with lowered mortality rates. Consequently, there is a large section of population that is ageing, and as per predictions this ageing population is only going to increase further [1]. For instance, in several countries the percentage of people over 65 years old exceeds the world average of 8% such as 15% in Australia, 15% in the USA, and 17% average in Europe. This percentage will have increased by 2050 when 25.7% of the OECD (countries in Organization for Economic Cooperation and Development) population is expected to be over 65 years of age of which 10% of the population is predicted to be over 80 [2]. With age come a variety of health care problems that often require continuous and long-term medical care that can put severe strain on health care resources and increase costs. Therefore, there is a growing emphasis on finding alternative non-traditional ways such as home care to manage patients so as to ease the burden on health care services and control costs. Research also shows that a large number of elderly people do prefer home care with remote monitoring [3].

Remote patient monitoring system is one such technology that is being widely adopted wherein those patients whose conditions allow them to remain at home are encouraged to do so while being monitored remotely. This can apply to a large section of elderly population such as: (i) the terminally ill – in-home care gives them the chance to spend the last days of their life in the comfort of their home with familiar surroundings instead of at a hospital, (ii) the chronically ill – those with chronic conditions such as diabetes mellitus, cancer, hypertension, chronic obstructive pulmonary disease, and others require long-term care and regular monitoring of their vital parameters. Internet of things (IoT) body sensors collect these parameters at pre-defined regular intervals and transmit the data to the health care manager for assessment and to alert them in case of any emergencies, thus reducing the number of visits a patient may otherwise have had to make to the health care centre. This has dual benefits of decreasing patient inconvenience and freeing up the time and space of the health care centre for other patients, and (iii) the memory impaired – patients with memory loss may otherwise be healthy enough to continue living independently but may need assistance to remind them of certain important tasks such as taking their medications on time. In-home care gadgets can be configured to aid the patients with such tasks while also notifying their health care providers in case of any deviations.

Remote health care monitoring applications require the use of several body sensors to regularly measure the health parameters as well as environmental sensors to monitor the ambient parameters and transmit contextual information to the patient's health care network (which may consist of health care workers, emergency services, and health care centres). The sensors collect patients' clinical data and the state of their surroundings, which are then transmitted to the appropriate health care centres and providers.

Despite the evolution of big data processing technologies (such as Hadoop and Apache Storm) and scalable infrastructure (such as virtualised clouds), there remains a significant gap as regards to heterogeneous data collection, real-time analysis, and automated decision support (semantic reasoning) based on defined quality of service (QoS) constraints. Given the increase in volume, velocity, and variety of sensor data health care sensors, special techniques and technologies for analysis and inferencing are required. These challenges are significantly pronounced within health care where data is being generated exponentially from biomedical research, remote body sensors, and electronic patient records among others. These limitations (i.e. being unable to satisfy QoS) can negatively affect patient care especially with respect to in time alerts and diagnostic suggestions. Making clinical decisions based on incomplete information can be error prone. There are several aspects that influence the immediate and long-term decisions including patients' current medical status, comorbidities, current medications, past clinical history, and contextual factors such as patient location, behaviour etc. However, manually analysing all the information together and predicting the outcome of their interactions can be extremely difficult and subjective leading to missed and/or overlooked information. Besides, the process is dependent on the patient information being made available to the appropriate professionals as and when required. The process becomes more cumbersome and error prone when the number of patients keeps increasing. Hence, automating the tasks of collecting, sharing, and analysing patient information is gaining increasing importance. Moreover, the use of semantic technologies such as ontologies is widely accepted to add intelligence to the process of analysis, thereby improving the process of decision making.

Current approaches to define QoS and the corresponding service level agreements (SLAs) to support health care services are in infancy and are not addressed in the state-of-the-art research. Given the inherent complexities introduced by various dimensions of such systems: namely, the edge/physical layer (devices), infrastructure layer (networking, processing, and storage), and analytics layers (analytic reasoning and inferring algorithms to draw insights using from data), defining QoS for is a hard problem.

In this paper, we take the pioneering steps in: (i) presenting the workflow of remote health-monitoring applications; (ii) identifying the challenges in analysing the data originating from health care IoT sensors; and (iii) identifying the challenges in identifying and guaranteeing QoS metrics required to meet the analysis and inferencing needs of the health care applications. Our vision of a QoS driven remote health-monitoring system will have multiple SLAs to suit the needs of the application. For example, we will look beyond the current cloud SLAs by exploring metrics such as event detection accuracy, time to detect, and respond with alerts that can directly be mapped to service cost. Our vision aims to closely follow the most popular pay for usage cloud model. For example, in the future, users of such remote health care applications say paying '$100 per month for the service will expect: (i) events such as heart attack, falls are detected (from IoT devices) within x milliseconds and (ii) alerts are automatically sent to the doctors, caregivers, and emergency ambulance teams within y minutes of event detection'. Furthermore, we envision that the system will be intelligent enough to be able to correlate and analyse data obtained from patient's medical records and background medical knowledge base. Such hybrid and timely information correlation will equip the health care professionals with the right information at the right time and personalised to the patient in order to provide timely and appropriate medical care.

The rest of this paper is organised as follows: in Section 2, we describe the cloud of things (CoT) followed by Section 3 where we explain how semantic reasoning makes a remote health care application more intelligent. Furthermore, we discuss the need for QoS in these health care applications. In Section 4, we present in detail the research directions for enforcing the QoS metrics in future remote health care monitoring applications. Finally, we conclude this paper in Section 5.

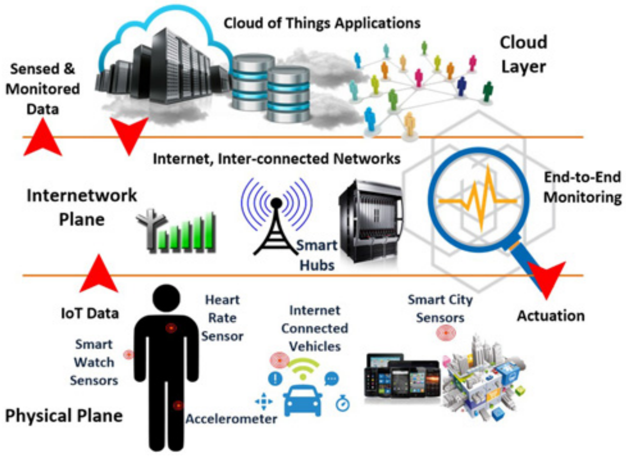

Fig. 1 *Conceptual architecture of CoT*

2 CoT: background

The remote patient monitoring system is supported primarily by the recent advancements in two technologies: namely, IoT and cloud computing. These technologies are already becoming part of our daily lives and are attracting significant interest from both industry and academia. The term IoT collectively describes technologies and research disciplines that enable the Internet to reach out into the world of physical objects. Technologies such as radio frequency identification (RFID), short range wireless communications, real-time localisation, and sensor networks have become increasingly pervasive, thus making the IoT a reality. According to the recent Gartner report [4], it is estimated that IoT will grow to 26 billion units by 2020, excluding PCs, tablets, and smartphones (40 billion things including tablets and smartphones). The revenue as a result of this growth is estimated to be ~$1.9 trillion. The IoT will fuel a paradigm shift of a 'truly connected world', in which everyday objects become inter-connected and smart with the ability to communicate many different types of information with one another. Cloud computing on the other hand allows IT-related resources [e.g. central processing unit (CPU), applications, network, and storage] to be provided as virtualised services to the customers under a usage-based payment model. Using cloud computing, customers (e.g. SMEs, governments, and universities) can leverage these virtualised services on the fly; ensuring that they do not have worry about the infrastructure details such as where these resources are hosted or how they are managed.

The CoT is our vision 'Of the collection of smart IoT sensors, IoT gateways (e.g. raspberry pi 3, UDOO board, esp8266 etc.), software defined networking devices solutions (e.g. Cisco IOx, HP OpenFlow, and Middlebox Technologies) at the network edge fully connected to and integrated with the traditional cloud data centre(s) for data storage, processing, analytics, and visualisation'. We expect that the CoT paradigm will support the development of novel remote patient monitoring applications that are composed of IoT devices and the high volume, high velocity, and high variety data produced by these IoT devices processed using big data technologies deployed over on public/private cloud environments. The CoT paradigm allows the use of sensors and devices to sense, collect, store, process, and analyse data related to different physical and environmental aspects of a patient in order to provide different health care services. However, such systems generate significantly large amounts of data and in order to deliver value to the patients, care givers, and administrators. In other words, remote health care applications need to embrace the big data explosion.

As a matter of fact, the health care sector is considered the fastest growing segment in the big data universe. Fig. 1 depicts a conceptual overview of the CoT ecosystem that includes context of health care and other application domains such as smart cities. The physical layer comprises things such as smart sensing devices, human sensors, connected cars, and smartphones that sense and actuate the physical world. On the other hand, the cloud layer is responsible for modelling and representation of the physical entities as programmable virtual entities. The cloud layer also includes the application layer that is composed of IoT services, application, and business processes that make use of the virtual entities and their virtual representation to control/monitor/detect state changes in the physical world. For example, consider the query 'Provide the indoor temperature in Room 1.23' or 'Set light level in Room 2.57 –15'. To support such queries, the interactions and associations between the physical layer, the virtual entity, and the IoT application need to be modelled. For example, the associations will contain the information that Sensor 456 provides the indoor temperature of Room 1.23 and the measurement unit is Celsius. In our vision of CoT, the virtual entities have functions to interact with the IoT device subsystem as well as functionality for discovering and looking up IoT devices that can provide information about virtual entities. Furthermore, it contains all the functionality needed for managing associations, as well as dynamically finding new associations and monitoring their validity, e.g. due to the mobility of IoT devices that can impact the state of the virtual entities.

3 Real-time reasoning in remote health care: the need for QoS

In remote-monitoring applications, the use of semantic web technologies is generally limited to leveraging the sensor data with contextual information about the sensors (types, location etc.) and patient's environment (location, time of day etc.). While the incorporation of contextual information makes the system more intelligent and enables improved decision making based on a variety of relevant factors, it does not take into account the patient's health status in its entirety, which can have damaging consequences especially pronounced in the elderly who are most likely to have multiple morbidities at any given time.

Fortin *et al.* [5] in their study of primary care practise in Québec, Canada concluded that the prevalence of multiple morbidities increases with age with more than 95% of patients aged 65 years and above having several comorbidities. Brett *et al.* [6] arrived at a similar conclusion in Australia of rise in both multimorbidities and their severity with increasing age. The increasing complexity of illnesses in turn requires the data to be analysed for several decision-making purposes including but not limited to: (i) rule out conflicting recommendations (such as contraindicated medications); (ii) identify risks (such as potentially serious complications that may arise due to the interactions of two or more illnesses); (iii) suggest preventive actions (initiate a prevention plan for probable future risks); and (iv) detect missed and/or incorrect diagnoses among others. To obtain comprehensive knowledge for informed decision making, the sensor data must also be leveraged with information about patients' past and present medical conditions. This is required to see the complete picture of the patients' clinical condition and present health status. Relying only on the sensor data can be dangerous as vital information maybe missed or overlooked. With the help of the following scenario, we explain the importance of leveraging sensor data with patient's historical data and reasoning over it to generate timely alerts.

Consider a patient named Jill, aged 71, who has the following chronic medical conditions: diabetes mellitus type 1, hypertension, and asthma. She is being monitored at home via various body sensors that regularly measure her blood glucose levels, blood pressure, medications, respiratory rate etc. One particular day, Jill gets an acute asthma attack also known as status asthmaticus. Being a life threatening condition, an alert is immediately sent by her body sensors to emergency medical services, her health care provider, and the local health care service centre. The emergency workers arrive and prepare to give her the appropriate medications. Corticosteroids are one of the mainstays for reversing status asthmaticus [7]; however, corticosteroids can put the patients with diabetes mellitus type 1 at a risk of developing diabetic ketoacidosis [8] and high blood pressure in hypertensives [9] both of which are again life threatening conditions if not controlled. In the absence of any more information about Jill's medical conditions apart from the alert sent by the sensors about the asthma attack, the emergency workers would inject her with high doses of corticosteroids, thereby putting her at further risks and complicating the situation even more. On the other hand, if the sensor data is leveraged by additional information from Jill's medical record such as that of her comorbidities, and the data are analysed together with background knowledge of the domain, then the potential risks: namely, diabetic ketoacidosis and elevated blood pressure can be derived and timely alerts sent to the emergency workers. In this way, important relevant information that can affect the patient's condition is neither overlooked nor missed. Consequently, the risks of injecting corticosteroids can be known beforehand and the appropriate precautionary measures can be set up to either eliminate the risk or manage it efficiently. This scenario highlights the importance of two vital processes in remote health monitoring of patients: (i) timely availability of complete patient health record at point of care and (ii) automated real-time reasoning over patient's conditions based on the background domain knowledge to obtain new knowledge.

3.1 Semantic technology for remote health care

Semantic web technologies such as ontologies and semantic reasoners are commonly used for reasoning over complex information as presented in the above scenario to obtain new actionable knowledge that could have been otherwise missed or overlooked [10, 11]. In medical situations, overlooking or missing any such vital information can have disastrous consequences. Therefore, we propose that in addition to obtaining contextual information, semantic web technologies should also be used in remote-monitoring applications for complex reasoning over patients' past and present medical conditions to derive a more comprehensive understanding about the patients' overall health status. Since the information obtained from the sensors is not viewed in isolation, better insights can be obtained for decision making and the number of false positive alarms can be minimised.

3.1.1 Ontology: Ontology is a formal, declarative model of a domain that captures the domain knowledge in the form of concepts and relationships between them. Web Ontology Language or OWL [12, 13] is an expressive ontology representation language based on description logic and is a Web Standard. The use of ontologies ensures retention of meaning in the shared information but OWL has several restrictions imposed on it to ensure decidability and computation completeness. Therefore, to increase the expressiveness and to model action sequences, semantic rules are often added to the ontology [14]. The ontology that models the background knowledge of a domain along with rules forms the knowledge base that is used with instance data to obtain new knowledge as output. A formal knowledge base thus makes a system intelligent by:

(i) Providing a reusable and shareable domain description.
(ii) Designing context-aware systems.
(iii) Enabling meaningful data sharing and integration.
(iv) Deriving new actionable knowledge to aid in more informed decision making.

In case of remote monitoring in health care, the hard challenge is to model an ontology that can comprehensively capture the domain knowledge and can cross-link across multiple real-time and historical data feeds for complex event processing. The processing involves reasoning so as to obtain both contextual and actionable knowledge. Since technology keeps changing rapidly, the ontology should also be adaptable so that existing concepts can be modified to reflect the current knowledge and extendable so that newer concepts can be added without undermining the existing model.

3.1.2 QoS agnostic reasoners: The ability to perform automated reasoning by performing reasoning algorithms on OWL ontologies increased the interest to use this form of knowledge representation and subsequently increased the number and size of ontologies. The emergence of big data and IoT on the one hand and growing demands for intelligent and smart applications on the other hand have brought significant challenges in high performance, efficient, real-time, and scalable reasoning. It is understood that by leveraging cloud-based technologies one can implement scalable reasoning techniques that meet real-time QoS constraints. As we discuss next, there exist several approaches that attempt to scale reasoning techniques using MapReduce programming model; however, they are still very limited especially in context of meeting real-time QoS constraints.

It is no surprise that traditional reasoning techniques have largely ignored the performance and scalability aspects, hence recent studies have shown that traditional reasoning techniques are incapable of reasoning over massive data and large ontologies [15, 16]. The primary causes of this limitation are: (i) complexity of the ontologies; (ii) centralised structure of the reasoners; and (iii) limited incremental reasoning capability. In general, the primary aim of measurement and evaluation of the complexity in computer science is to quantitatively understand the challenges of implementation, development, or maintenance of an approach or model. Complexity is an important challenge in most of the computer science research areas dealing with knowledge representation and data integration. Ontology complexity has been the subject of considerable research in recent years to improve understanding, development, maintenance, and integration of ontologies. For example, Zhang *et al.* [17] attempted to adapt software complexity metrics proposed by Weyuker [18] to measure and quantitate complexity of ontologies in both ontology and class levels. However, quantitation of the ontology quality including usefulness, perfection, accuracy, and possibility of integration to the other commonly used ontologies are not measurable with proposed metrics in the recent research works. Distributed cloud-based reasoning approaches have the potential to improve the performance (i.e. meet QoS constraints) of reasoning on large ontologies and dataset.

Recently, several researchers attempted to propose and customise current reasoning techniques for virtualised cloud infrastructure. Urbani *et al.* [19] proposed a distributed and scalable reasoning technique based on MapReduce [20] and deployed it on top of Hadoop [21] and a compute cluster of up to 64 commodity machines. Schlicht and Stuckenschmidt [22] proposed MapResolve as a scalable and distributed reasoning solution for description logic ontologies on MapReduce framework [20]. The authors investigated ontology reasoning on MapReduce and proposed a solution to avoid repeated inferences as the main problem in using MapReduce-based reasoning approaches. Although there are several attempts to make distributed reasoning on MapReduce, there is no generic and scalable solution for distributed reasoning on OWL ontologies. Urbani *et al.* [23] proposed WebPIE (web scale inference engine) as a scalable parallel inference engine based on MapReduce for the large-scale dataset. The proposed approach is more efficient and supports the incremental reasoning that performs the reasoning only on the portion of the data that is changed since the previous reasoning. Grau *et al.* [24] in a similar approach proposed an incremental reasoning by taking the advantage of the similarities between different versions of an ontology. The approach is independent from reasoning calculus and can be used with any reasoner. Moreover, the author claimed that the incremental classification proposed in this paper is nearly real time for almost experimented ontologies. Liu *et al.* [25] proposed both incremental and distributed reasoning methods for large-scale ontologies and based on MapReduce framework. The approach implemented and evaluated using a cluster of eight nodes and on top of Hadoop that shows high-performance reasoning and runtime searching, especially for incremental knowledge base. Although there are performance improvements in recent incremental reasoning approaches, there is not a complete evaluation of the accuracy and adaptability of these approaches in large-scale datasets and OWL ontologies.

3.2 Workflow of semantic reasoning in CoT paradigm

As the amount of sensor data grows in massive amounts, the information that can be obtained from it increases proportionally. The data however requires timely analysis in order for the information to be available for use. A combination of cloud and semantic technologies for analysing the IoT sensor data can help in various important ways such as: (i) analysis of large amounts of data in a real (near-real) time to reveal implicit information hidden inside the vast volumes of data, (ii) personalised care plan for patients, (iii) increasingly intelligent and better informed decision support, (iv) timely availability of patient information to the health care practitioners or providers, and (v) automating the process of analysis and decision making. Fig. 2 gives an overview of how cloud and semantic technologies operate in case of remote in-home monitoring of patients.

Fig. 2 shows that patient monitoring sensors send the data to a stream processing engine, which is a system that can preprocess data closer to the source of data. It processes the data to detect any abnormality. The stream with no abnormal event is sent to a remote server for storage and can be used for historical analyses or clinical analyses at a future time. These datasets, which did not raise alarms can be managed by non-relational, distributed databases such as Apache HBase [26]. There are several advantages of using HBase including its native support for Apache Hadoop data processing engine and Apache Hive data querying engine, Apache HBase, Apache Hadoop, and Apache Hive have emerged as a recent alternative to traditional data warehousing tools such as DB2 and Teradata. HBase database has emerged as the more flexible and scalable alternative to traditional data warehouse tools such as Teradata and DB2 due to its native support for big data processing (Hadoop) and querying (Hive) engines, which will allow scalable processing of historical datasets if and when required by the semantic reasoning engine. In case of abnormality detection (e.g. high blood pressure, high glucose level, and sudden fall), the stream processing engine signals the health data management orchestrator (HDMO) to check whether the detected abnormality needs more actions. The above abnormality detection datasets are also stored within Apache HBase along with non-abnormality data for future cross-referencing and anomaly verification. The HDMO is a specialised software program which runs in the cloud and has all the required information such as where are the ontologies deployed, where are historical record stored and how to query and reason over these historical records. We believe that the scalability of abnormality detection algorithm will be handled by HDMO in response to changing data volume and data velocity. This would require workload characteristic and QoS modelling across the software components (see Fig. 2) involved in the processing of data associated with the detection algorithm.

After receiving input from stream processing engine, HDMO uses the patient identification information (available from incoming sensor feed) to query the historical records stored in databases. It also connects to a TripleStore and fetches the relevant ontological rule engine that is necessary to reason across the historical and real-time data from sensors. TripleStores are specialised databases for storage and management of ontologies with reasoning and query processing capabilities [27]. For the purpose of scalability for the incoming sensor feeds, the HDMO will fetch the ontology from the TripleStore and replicate it across multiple virtual machines that already have the necessary ontology libraries pre-installed. This way ontological reasoning engine can be run on demand in a distributed fashion and the patient record as well updated with the new information in real time.

The sensor data with the abnormal parameters is passed into the virtualised TripleStore, which also fetches the patient's electronic health record (EHR) from an external database. The EHR contains patient's history as well as information about current conditions. The semantic reasoner in the TripleStore analyses the sensor data together with the data from the EHR against the background domain knowledge to obtain new knowledge that could be a possible new diagnosis, an intervention suggestion, a warning against certain actions, a recommendation for a preventive measure etc. The aim is to integrate all relevant information for decision making: patient's environment and present condition (derived using semantic tagging of sensor data), background domain knowledge, and patient's past conditions and current comorbidities (obtained from patient's EHR). The output is sent to the EHR, which is updated with the new knowledge. The reasoner once again reads the EHR to derive further knowledge and the process continues this until no further derivations are possible. Once satisfied, the EHR pushes the new

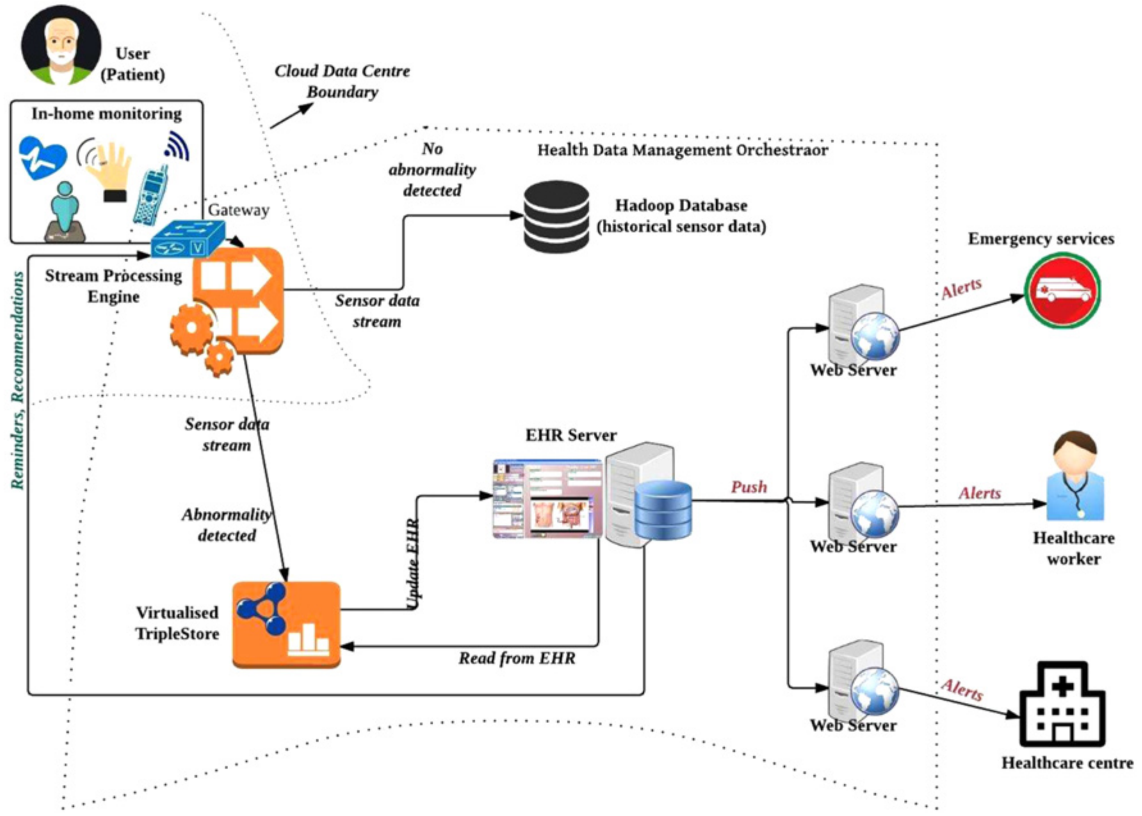

Fig. 2 *Overview of remote health care application using the CoT paradigm*

updates as alerts and suggestions to the authorised and relevant recipients. To ensure scalability, the EHR should be implemented as a managed relational database service (e.g. MySQL cluster CGE), which can be scaled in response to query rates and other QoS requirements. Our assumption that EHR system is implemented as a managed relational database service is reasonable considering interoperability across legacy health care systems were designed and implemented using relational database programming model (e.g. SQL). Although the entire process is automated, the final decisions always rest with the health care professionals.

The main steps of a CoT-based remote health-monitoring application can be summed up as follows: (i) a home-based gateway sends the sensor data to a stream processing cluster running inside cloud; (ii) the sensor data is processed by the stream query engine and divided into two – one with an event and one without. The eventful data is semantically enriched using tags and reprocessed to determine the patient's status; (iii) this data is then sent to a virtual machine containing a copy of the domain knowledge base that analyses it further to derive contextual information (information about patient's location, time of day etc.). It also pulls patient's EHR for analysing the sensor data in combination with patient's historical data as well as the current comorbidities if any. The output is generated in the form of new knowledge, which is then written into the EHR, (iv) the output in the form of alerts is finally sent over web services to the appropriate health care targets such as the patient's doctor, emergency services, local health care centre etc. The internal working of the cloud system (with the semantic reasoning, knowledge base, and historical data) is presented in Fig. 3.

4 QoS issues for cloud-hosted remote health care applications: research directions

In this section, we will present the research directions in enforcing QoS metrics for future remote health care monitoring applications. We start with first identifying QoS management challenges in current CoT system, and finally present the related research issues with respect to design and development of real-time patient monitoring applications.

The current mechanisms for QoS provisioning and supporting SLAs [28–31] in IoT and clouds have major limitations. In the light of CoT, these mechanisms will have to be radically reconsidered or even reinvented to meet the challenges posed by upcoming remote health care CoT applications. Figs. 1 and 2 capture the complexities of CoT applications from the physical device (data collection) to virtual layer (storage and processing) to the application layer (delivery). However, QoS guarantee for the remote health care CoT is expectedly challenging, and an emerging discipline. This is due to the shortage of standardised, end-to-end approaches for QoS assurance (between the end user, IoT devices, and the cloud), the complexity of the integration of different layers (see Figs. 1 and 2), and the presence of a plethora of QoS constraints and parameters at each layer. We expect that the traditional way of QoS assurance will not be sufficient. For instance, we will soon be looking at satisfying requirement such as 'detect/notify events within 5 minutes of occurrence' rather than the traditional model to 'guarantee 99.99% CPU availability'. In this section, we articulate the research directions pertaining to the QoS for CoT applications from the perspectives of physical/edge layer (IoT) and cloud layer. Finally, we summarise how these research challenges are intertwined with remote health care applications.

4.1 IoT edge/physcial layer

QoS semantics: The IoT physical/edge layer (see Fig. 1) comprises heterogeneous sets of devices ranging from wireless sensor networks, body area networks, and other virtual sensors (e.g. weather and flood services). With technological advancements and evolution of new remote health care applications, this space is going to get further crowded with billions of devices each with different capabilities and QoS, parameters, and constraints. The heterogeneity

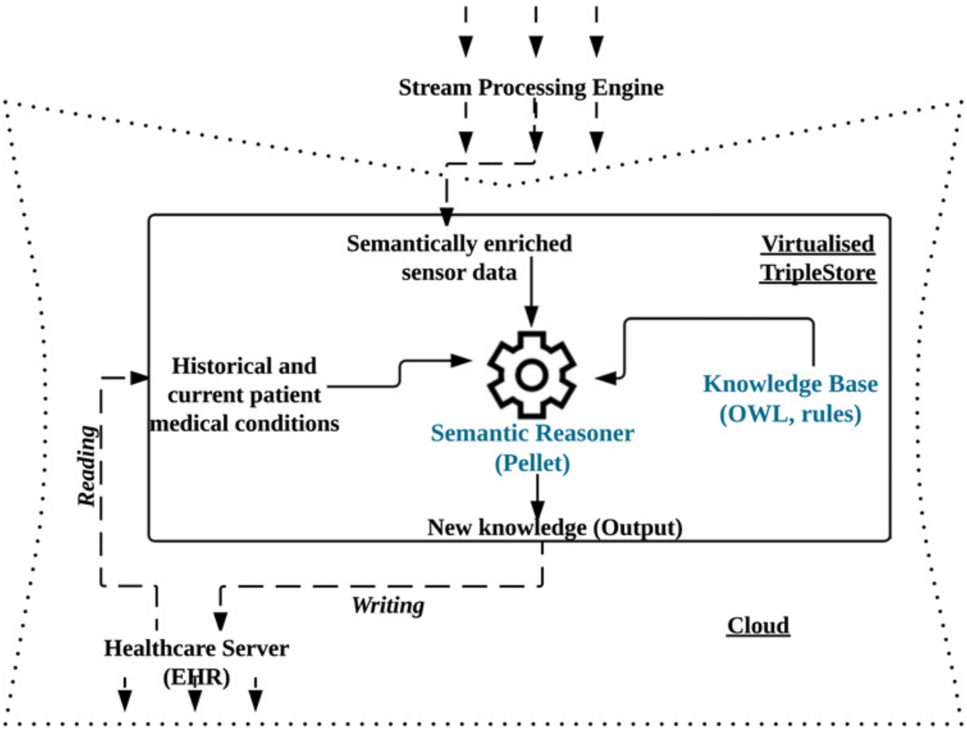
Stream Processing Engine
Semantically enriched sensor data
Virtualised TripleStore
Historical and current patient medical conditions
Knowledge Base (OWL, rules)
Semantic Reasoner (Pellet)
Reading
New knowledge (Output)
Writing
Healthcare Server (EHR)
Cloud

Fig. 3 *Internal working of a TripleStore in remote health care application scenarios*

among the devices in this layer requires adaptive QoS constraints that can deal with all kinds of traffic demands. For example, emergency message from a body sensor that has detected a heart attack is more important than a regular 'all is well' monitoring message. Such emergency messages will need to be identified and the entire ecosystem needs to support delivery of emergency messages to the end user within acceptable limits. The QoS of the physical layer is related directly to the quality and the timeliness of the IoT data used for decision support and automation. This includes (but is not limited to): *sampling rate, transmission speed, data quality, coverage, mobility, location, and cost*.

The *sampling rate* determines the rate at which a sensor measures an observed phenomenon (e.g. 5 Hz). Different remote health applications require different sampling rates based on their criticality. The *transmission speed* is network dependent and refers to the actual rate at which data is sent to the cloud layer from the physical layer. This is influenced by network topology and the device connectivity [e.g. Bluetooth, wireless fidelity (WiFi), and fourth generation (4G)]. *Data quality* is a complex metric [32] that needs to be carefully considered for future adoption and sustainability of CoT applications in health care and other domains. However, similar to variation in quality of real-world things (e.g. quality of cotton), the need for high-quality data satisfied by metrics such as high accuracy (recently calibrated) and minimal error is very application dependent (health care is a typical example where data accuracy and quality could play a vital role in saving life). The *coverage* at the device layers identifies the extent to which the sensor data covers the observer phenomenon geographically.

Mobility is another key QoS metric as the devices or the people wearing the devices at the physical layer are inherently mobile. Mobility enables greater coverage but also introduces further challenges that impact other metrics such as *location*, quality, and transmission speed. Finally, *cost* is a summative metric that relies on the previously identified metrics. For example, the cost of IoT data that is 99% accurate could be much higher when compared with data that is 70% accurate. This sort of cost could be used to differentiate the level of care service provided to the end user. More critical cases could desire a more accurate system while basic monitoring cases could work with lesser accuracy. It is clear from the above discussion that QoS will enable CoT systems to cater to the needs of different applications hence reaching a wider audience. Moreover, our vision of CoT of the future is an open world where multiple devices and analytics applications (owned and operated by independent providers) can be fused together to create innovative CoT applications/services.

Some of these metrics identified maybe trivial for a single device per se. However, the notion of CoT is for millions of such devices to collaborate and achieve a goal that is otherwise impossible to do it alone. Consider the example of a temperature sensor measuring temperature in Fahrenheit. When this information is shared with a hospital system that accepts data in Celsius, it will be inconsistent and misleading. The question here is then: Who should agree on the formats? Is it the job of the human or can the devices auto-negotiate the semantics? There is need to develop novel methods to map and represent the QoS into a language that can be used by the whole CoT ecosystem.

4.2 CoT: cloud layer

4.2.1 Distributed QoS-aware CoT data processing: In a CoT paradigm, the devices will be connected to the clouds for data storage, processing, analytics, and visualisation (see Figs. 1 and 2). Cloud computing offers virtually infinite set of resources [CPU, memory, storage, and input/output (I/O)] at reasonable costs based on a multi-tenant and pay-as-you-model. Clouds enable applications to be hosted in a highly distributed manner across multiple datacentres around the globe. At the same time, clouds ensure that the applications are highly available and scalable. These characteristics have made clouds to be widely adopted by private and public sector industries for reliable and efficient service delivery.

With the emergence of IoT, it is expected that 40 yottabytes of data will be generated by the end of 2020 [33]; therefore, requiring large-scale data (big data) processing at multiple cloud datacentres. This is a challenging problem as most of the cloud data processing and analytics tools such as MapReduce are optimised to run in a single datacentre [33]. Furthermore, there exist no comprehensive mechanisms where large datasets, in particular health related datasets (historical and real-time), spanning multiple datacentres can be efficiently processed. This can mainly be attributed to the bottlenecks in network communication medium, and cloud and network infrastructure that can be congested and/or may not provide sufficient throughput to transfer large datasets in a timely manner for QoS-aware data processing and analytics. Therefore, there is a need to address the challenge of distributed QoS-aware data processing in CoT ecosystems.

4.2.2 Cross-layer cloud and network QoS monitoring: A typical cloud system comprises of three layers: infrastructure-as-a-service (IaaS), platform-as-a-service (PaaS), and software-as-a-service (SaaS) [34]. Each layer has a specific set of metrics. For example, for SaaS: application's throughput and CPU utilisation; for PaaS: system up time, number of requests served; and for IaaS: memory and CPU utilisation. Cross-layer cloud QoS monitoring is a challenging task [35]. This is based on the assertion that there are no standard formats and APIs for cloud QoS monitoring. Currently, each cloud provider such as Amazon EC2, Google Compute Engine, and Microsoft Azure provide their own set of APIs. On top of that only Amazon provides a limited set of QoS metrics for cloud monitoring. Companies such a CloudHarmony provide cloud-monitoring information for multiple cloud providers but do not provide metrics for comparison and analysis of cloud providers. Furthermore, most of research in this domain does not aim to collect a large pool of data for numerous metrics belonging to each cloud layer. Alhamazani *et al.* [35] has recently tried to address this challenge and built a service for cross-layer cloud QoS data collection. However, they do not provide methods for QoS analysis of various clouds and recommending methods for application hosting.

End-to-end QoS also involves network QoS in the form of propagation delay or network latency and throughput. One of the bottlenecks to ensure high network QoS is the network latency. In that, if an application requests some data from the CoT application hosted on the cloud, this data may involve some time to arrive back to the application. This is due to the distance (in hops) between the application and the cloud datacentre (assuming no cloud processing delay). For instance, the latency will be shorter (in tens of milliseconds) if the application running on device present in Europe (connected via Ethernet) requests data from European datacentre instead of South Asian datacentre where the latency can be in the order of hundreds of milliseconds. These factors necessitate large-scale cross-layer cloud QoS monitoring and network QoS monitoring for QoS-aware CoT ecosystem.

Guaranteeing performance SLAs (which are expressed as constraints on QoS metrics) for remote health care CoT applications requires clear understanding of important performance metrics across cloud-hosted big data processing frameworks (e.g. Apache Kafka, Apache Storm, Apache Hadoop etc.) and hardware resources (CPU, storage, and network). The problem is further complicated due to the fact that the QoS metrics at CoT device layer, CoT application layer, cloud-based big data processing frameworks layer, and cloud-based hardware resource layer are not necessarily the same [36, 37]. For example, the key QoS metrics are: (i) event detection delay and decision-making delay at CoT application level; (ii) throughput and latency in distributed messaging queuing systems (e.g. Apache Kafka); (iii) response time in batch processing systems (e.g. Apache Hadoop); (iv) response time for processing top-k queries in transactional systems (e.g. Apache Hive); (v) read/write latency and throughput for the file system of big data cluster; (vi) delay introduced by ontology reasoning in the TripleStore to identify significant events from real-time and historical data; and (vii) utilisation and energy efficiency for CPU resources. Therefore, it is not yet clear how (i) these QoS metrics could be defined and formulated coherently across layers and (ii) the various QoS metrics could be combined to give a holistic view of the data flows across multiple IoT sensors, big data software frameworks, semantic processors, and hardware resources. To ensure application-level performance SLAs/QoS there is also a need to monitor workload metrics (data volume, data velocity, data variety, and sources, and types and mix of search queries) across big data processing frameworks such that appropriate workload characterisation models could be developed. The hard challenge is how to collect and integrate monitoring data from all the big data processing frameworks and hardware resources for administrators to easily track and understand application-level SLAs/QoS without the need to understand the complexity of the whole platform.

4.2.3 QoS-aware cloud and network selection and orchestration: Over the past decade, we have witnessed a significant increase in the usage of smartphones, sensors, and applications. These technologies will become an integral part of the CoT ecosystem, thereby bootstrapping novel applications in areas such as health care and emergency management. It is evident that for these classes of applications, data would be sent (sensor data from heart rate monitor) and retrieved (processed response) while the users (e.g. emergency personal and patients) are on the move. Mobility has its inherent challenges, for example, a mobile device may connect to different access networks, for example, 3G and WiFi, where each network offers different latencies for upstream data (from the device) to downstream data from the clouds. This creates several challenges: first, there is a need to select a best network as mobile devices can connect to heterogeneous access networks leading to stochastic network QoS; and second, QoS-aware cloud selection as the CoT applications and data will be hosted in multiple datacentres. Therefore, we consider end-to-end QoS provisioning in CoT as a joint optimisation problem where both network and cloud QoS should be optimised together.

A large body of work exists in the area of cloud selection. For example, Garg *et al.* in [38] present analytic hierarchy process (AHP)-based method for cloud ranking and selection. However, there is a dearth of literature that considers joint cloud and network selection. Mitra *et al.* in [39] proposed a system for cloud and network selection while users are on the move in heterogeneous access networks. However, these methods are also limited. For example, when a set of applications running on multiple devices selects a particular cloud for application data processing, all the requests may automatically get transferred to that cloud datacentre, creating burden on servers running on that datacentre, leading to overprovisioning on one datacentre and under-provisioning on another datacentre. Furthermore, this may also trigger VM migrations and data replication across multiple datacentres leading to an inefficient CoT ecosystem. This necessitates development of novel cloud orchestrators that are QoS and mobility-aware to for efficient QoS provisioning in CoT.

4.3 QoS for real-time patient monitoring

One of the critical aspects for the successful deployment of the real-time patient monitoring applications is the end-to-end QoS that directly affects the timeliness of the care given to the patient [40]. In that, it requires the data generated from the health care sensors to be collected, transmitted (via the network), processed, analysed, and used/acted on in a timely manner. For data processing, analysis and usage, we consider the cloud platform where the remote health-monitoring application is deployed as-a-service. However, end-to-end QoS assurance for such applications is a challenging task. It is due to the presence of several factors that may affect the end-to-end QoS of health care applications. For instance, the Internet path between the home gateway and the cloud (where the application is deployed), and the path between the cloud and the health care centre/worker and the emergency services may exhibit stochastic behaviour due to network congestion, random, and burst packet losses and network jitter.

Clouds may also exhibit stochastic behaviour due to their multi-tenant model and may significantly affect applications performance in terms of CPU, memory, and disk I/O operations [41]. To make things worse, different combinations of virtual machines, datacentre locations, and price may affect different components of the application stack (considering multi-tiered model). For instance, web server application-level QoS might by suitable; however, for the same application, MapReduce operations for data analytics may suffer due to variation in CPU and disk I/O operations. Therefore, in this case, the overall QoS may suffer. This necessitates monitoring of the whole application lifecycle, that is, from data generation, to data processing, and finally data consumption. By this we consider network and cloud monitoring. A cloud system is divided into three layers: IaaS, PaaS, and SaaS [34]. Each of these layers represents complex set of resources and involves several metrics. For example, application throughput at SaaS layer and CPU utilisation at PaaS layer. A real-time health care application is a multi-tiered application that spans across all cloud layers, and therefore requires cross-layer cloud monitoring. Its performance depends on the understanding and monitoring of QoS parameters across all cloud layers such that big data platforms such as Apache Kafka, Apache Storm, and Apache Hadoop, and hardware resources (CPU, storage, and network) can efficiently be processed. From the state of the art, we assert that cross-layer cloud monitoring is a challenging task and there is a dearth of research done in this domain.

Real-time health care application should consider the following questions: (i) how could the aforementioned QoS metrics be defined across network and cloud sides and (ii) how various QoS metrics should be combined to give a holistic view of the data flows across sensors, big data software frameworks, and cloud resources. To ensure QoS for such health care applications, there is also need to monitor the workload metrics (data volume, data velocity, data variety, and sources, and types and mix of search queries) across big data processing frameworks and various ontology reasoning systems such that appropriate workload characterisation models could be developed. The hard challenge is how to collect and integrate monitoring data from all the big data processing frameworks and hardware resources for administrators to easily track and understand application-level SLAs/QoS without the need to understand the complexity of the whole platform.

As can be observed, there is a need to consider relevant parameters from both cloud (QoS_c) and network (QoS_n) perspective to determine the overall QoS of the health-monitoring application. The complexity lies in the fact that there can be X number of parameters from networking side and Y number of parameters from cloud side. Therefore, making sense of $O \times M$ is quite a challenging task. To determine the QoS for a real-time health care application, we define a QoS metric as

$$\text{QoS} = f\{\text{QoS}_\text{c}, \text{QoS}_\text{n}, \text{location}, \text{battery}, \ldots, N\} \qquad (1)$$

where N represents the total number of QoS parameters (from both cloud and network perspective). As can be seen from the equation, determining the end-to-end QoS is a multidimensional and a complex problem that needs to be carefully addressed. To solve this metric and assuming the QoS parameters are carefully selected, multi-attribute decision-making algorithms [42] such as simple additive weighting technique for order of preference by similarity to ideal solution or AHP can be used. Garg *et al.* [38] present an AHP-based cloud ranking and selection model. This model can be beneficial for evaluating right cloud resources for multi-tiered health care applications. However, their model does not consider big data application hosting. Mitra *et al.* [39] proposed M2C2, a system that can support mobile patients for requiring real-time health care. Their system considers cloud and network selection as a joint optimisation problem and supports the selection of suitable cloud and suitable network. Most importantly, M2C2 considers end-to-end QoS monitoring, i.e. it monitors both cloud and network resources. However, both these approaches are not validated in real-time health care domain.

For the real-time health care application, mobility also poses serious concerns. For instance, a patient wearing a plethora of sensors, for example, the heart rate monitor and respiratory rate monitor can be mobile. These sensors may connect to his/her smart phone that acts as a gateway. Therefore, these sensors and gateways can be mobile and are expected to connect to heterogeneous access networks such as WiFi and 3G. Each of these networks offer different coverage and network characteristics such as throughput and delay. This necessitates the monitoring of different network types, and at the same time selection of right virtual machine type for each application tier [39].

5 Conclusions

CoT can prove to be a disruptive technology enabling novel applications in domains of health care, smart cities, smart manufacturing etc. Since CoT includes multiple computing paradigms such as IoT, cloud, and big data analytics systems, it is extremely challenging to design and develop applications using CoT while ensuring they meet the QoS criteria.

Considering remote health care application as a use case, we highlighted the end-to-end QoS and resource management issues that will arise in the future CoT ecosystem. We briefly discussed the state of the art to understand the research gaps. In the future, we envision a QoS-aware remote health care application that incorporates both the past and present medical conditions of the patients and uses a combination of IoT sensing, cloud, big data processing, and semantic web technologies to help health practitioners in making timely decisions. However, as we highlighted in this paper, this would require considerable research and development effort across multiple disciplines of computer science in collaboration with health care experts. Other important research exists in modelling QoS of software components while considering end-to-end data-privacy and data-anonymisation as discussed in our papers [43, 44].

6 Acknowledgment

This paper is an extended version of our conference paper titled ‘Orchestrating Quality of Service in the Cloud of Things Ecosystem’ published with 2015 IEEE International Symposium on Nano-electronic and Information Systems, December 2015.

7 References

1 Bureau, P.R.: ‘World population data sheet’, 2015
2 E.C. OECD Health Policy Studies: ‘United kingdom – a good life in old age? monitoring and improving quality in long-term care’, 2013
3 Bengtsson, T.: ‘Population ageing – a threat to the welfare state?: the case of Sweden’ (Springer Science & Business Media, Berlin, 2010)
4 Gartner, U.M.: ‘Gartner’s 2014 hype cycle for emerging technologies maps the journey to digital business’, 2014
5 Fortin, M., Bravo, G., Hudon, C., *et al.*: ‘Prevalence of multimorbidity among adults seen in family practice’, *Ann. Fam. Med.*, 2005, **3**, (3), p. 223
6 Brett, T., Arnold-Reed, D.E., Popescu, A., *et al.*: ‘Multimorbidity in patients attending 2 Australian primary care practices’, *Ann. Fam. Med.*, 2013, **11**, (6), p. 535
7 Young, D.J., Salzman, G.A.: ‘Status Asthmaticus in adult patients’, *Hosp. Phys.*, 2006, **42**, (11), p. 13
8 Rehman, A., Setter, S.M., Vue, M.H.: ‘Drug-induced glucose alterations part 2: drug-induced hyperglycemia’, *Diab. Spectr.*, 2011, **24**, (4), p. 234
9 McKay, L.I., Cidlowski, J.A.: ‘Corticosteroids in the treatment of neoplasms, (BC Decker)’, 2003
10 Valls, A., Gibert, K., Sánchez, D., *et al.*: ‘Using ontologies for structuring organizational knowledge in home care assistance’, *Int. J. Med. Inf*, 2010, **79**, (5), p. 370
11 Dang, J., Hedayati, A., Hampel, K., *et al.*: ‘An ontological knowledge framework for adaptive medical workflow’, *J. Biomed. Inf.*, 2008, **41**, (5), p. 829
12 W.W.W.W. Consortium*et al.*: ‘OWL web ontology language current status-w3c’
13 Golbreich, C., Wallace, E.K.: ‘W3C recommendation’, 2012
14 Shah, T., Rabhi, F., Ray, P.: ‘Investigating an ontology-based approach for Big Data analysis of inter-dependent medical and oral health conditions’, *Cluster Comput.*, 2015, **18**, (1), p. 351
15 Kang, Y.B., Pan, J.Z., Krishnaswamy, S., *et al.*: Aaai, 2014, pp. 80–86
16 Kang, Y.B., Li, Y.F., Krishnaswamy, S.: Int. Semantic Web Conf., 2012, pp. 198–214
17 Zhang, H., Li, Y.F., Tan, H.B.K.: ‘Measuring design complexity of semantic web ontologies’, *J. Syst. Softw.*, 2010, **83**, (5), p. 803
18 Weyuker, E.J.: ‘Evaluating software complexity measures’, *IEEE Trans. Softw. Eng.*, 1988, **14**, (9), p. 1357
19 Urbani, J., Kotoulas, S., Oren, E., *et al.*: Int. Semantic Web Conf., 2009, pp. 634–649

20 Dean, J., Ghemawat, S.: *Commun. ACM*, 2008, **51**, (1), p. 107

21 Shvachko, K., Kuang, H., Radia, S., *et al.*: 'The hadoop distributed file system'. 2010 IEEE 26th Symp. on Mass Storage Systems and Technologies (MSST), 2010, pp. 1–10

22 Schlicht, A., Stuckenschmidt, H.: 'MapResolve'. Int. Conf. on Web Reasoning and Rule Systems, 2011, pp. 294–299

23 Urbani, J., Kotoulas, S., Maassen, J., *et al.*: 'WebPIE: A web-scale parallel inference engine using MapReduce', *Web Semant., Sci. Serv. Agents World Wide Web*, 2012, **10**, p. 59

24 Grau, B.C., Halaschek-Wiener, C., Kazakov, Y.: 'History matters: Incremental ontology reasoning using modules'. The Semantic Web, 2007, pp. 183–196

25 Liu, B., Huang, K., Li, J., *et al.*: 'An incremental and distributed inference method for large-scale ontologies based on mapreduce paradigm', *IEEE Trans. Cybern.*, 2015, **45**, (1), p. 53

26 George, L.: 'HBase: the definitive guide' (O'Reilly Media, Inc., Sebastopol, 2011)

27 Sequeda, D.J.: 'Introduction to: triplestores'

28 Duan, R., Chen, X., Xing, T.: 'A QoS architecture for IOT, Internet of Things (iThings/CPSCom)'. 2011 Int. Conf. on and Fourth Int. Conf. on Cyber, Physical and Social Computing, 2011, pp. 717–720

29 Awan, I., Younas, M.: 'Towards QoS in internet of things for delay sensitive information'. Int. Conf. on Mobile Web and Information Systems, 2013, pp. 86–94

30 Jin, J., Gubbi, J., Luo, T., *et al.*: Network architecture and QoS issues in the internet of things for a smart city, 2012 Int. Symp. on Communications and Information Technologies (ISCIT), 2012, pp. 956–961

31 Nef, M.A., Perlepes, L., Karagiorgou, S., *et al.*: 'Enabling qos in the internet of things'. Proc. of the Fifth Int. Conf. on Communications, Theory, Reliability, and Quality of Service (CTRQ 2012), 2012, pp. 33–38

32 Davenport, T., Redman, T.: 'Build data quality into the internet of things', *Wall Str. J.*, 2015

33 Wang, L., Ranjan, R.: 'Processing distributed internet of things data in clouds', *IEEE Cloud Comput.*, 2015, **2**, (1), p. 76

34 Armbrust, M., Fox, A., Griffith, R., *et al.*: 'A view of cloud computing', *Commun. ACM*, 2010, **53**, (4), p. 50

35 Alhamazani, K., Ranjan, R., Jayaraman, P.P., *et al.*: 2015

36 Ranjan, R.: 'Streaming big data processing in datacenter clouds', *IEEE Cloud Comput.*, 2014, **1**, (1), p. 78

37 Ranjan, R., Kolodziej, J., Wang, L., *et al.*: 'Cross-layer cloud resource configuration selection in the big data era', *IEEE Cloud Comput.*, 2015, **2**, (3), p. 16

38 Garg, S.K., Versteeg, S., Buyya, R.: 'A framework for ranking of cloud computing services', *Future Gener. Comput. Syst.*, 2013, **29**, (4), p. 1012

39 Mitra, K., Saguna, S., Åhlund, C.: 'M 2 C 2: A mobility management system for mobile cloud computing'. 2015 IEEE Wireless Communications and Networking Conf. (WCNC), 2015, pp. 1608–1613

40 Khoi, N.M., Saguna, S., Mitra, K., *et al.*: 'IReHMo: An efficient IoT-based remote health monitoring system for smart regions'. 2015 17th Int. Conf. on E-health Networking, Application & Services (HealthCom), 2015, pp. 563–568

41 Leitner, P., Cito, J.: arXiv preprint arXiv:1411.2429, 2014

42 Hwang, C.L., Masud, A.S.M.: 'Multiple objective decision making – methods and applications: a state-of-the-art survey' (Springer Science & Business Media, 2012), vol. 164

43 Puthal, D., Nepal, S., Ranjan, R., *et al.*: *IEEE Cloud Comput.*, 2016, **3**, (3), p. 64

44 Nepal, S., Ranjan, R., Choo, K.K.R.: 'Trustworthy processing of healthcare big data in hybrid clouds', *IEEE Cloud Comput.*, 2015, **2**, (2), p. 78