Chapter 6
Interfacing Physical and Cyber Worlds: A Big Data Perspective

Zartasha Baloch, Faisal Karim Shaikh, and Mukhtiar A. Unar

Abstract With the increase in utilization and pervasiveness of smart gadgets, there is a rise in new application domains. For that reason, computational technologies are progressing very rapidly, and computations are becoming an essential part of our life. Cyber-physical systems (CPSs) are a new evolution in computing that are integrated with the real world along with the physical devices to provide control in real-time environments. CPS generally takes input through sensors and controls the physical system through cyber systems using actuators. Such systems are really complex and challenging as they control real environments. This necessitates a proper interfacing of physical and cyber domains. To this end, the data generated by physical devices is getting bigger and bigger that is collectively acknowledged as big data. The real challenge in interfacing cyber and physical domains is the efficient management of big data. Accordingly, this chapter discusses big data sources and the relevant computing paradigms. It also classifies and discusses the main phases of data management for interfacing CPS, viz., data acquisition, data preprocessing, storage, query processing, data analysis, and actuation.

Keywords Big Data • Cyber-physical systems • Cloud computing • Data analytics • Decision support systems • Data management • Big data sources

6.1 Introduction

The computing paradigm has evolved in line with the development of the latest and newer technologies. With these advancements, there is a perception that 1 day computing will become the fifth utility (after water, gas, electricity, and telephone) which will be essential for everyday needs of the society [1]. Cyber-physical

Z. Baloch (✉) • M.A. Unar
IICT, Mehran University of Engineering and Technology, Jamshoro, Pakistan
e-mail: zartasha.baloch@faculty.muet.edu.pk

F.K. Shaikh
IICT, Mehran University of Engineering and Technology, Jamshoro, Pakistan

TCMCORE, STU, University of Umm Al-Qura, Mecca, Saudi Arabia

Z. Mahmood (ed.), *Data Science and Big Data Computing*,
DOI 10.1007/978-3-319-31861-5_6

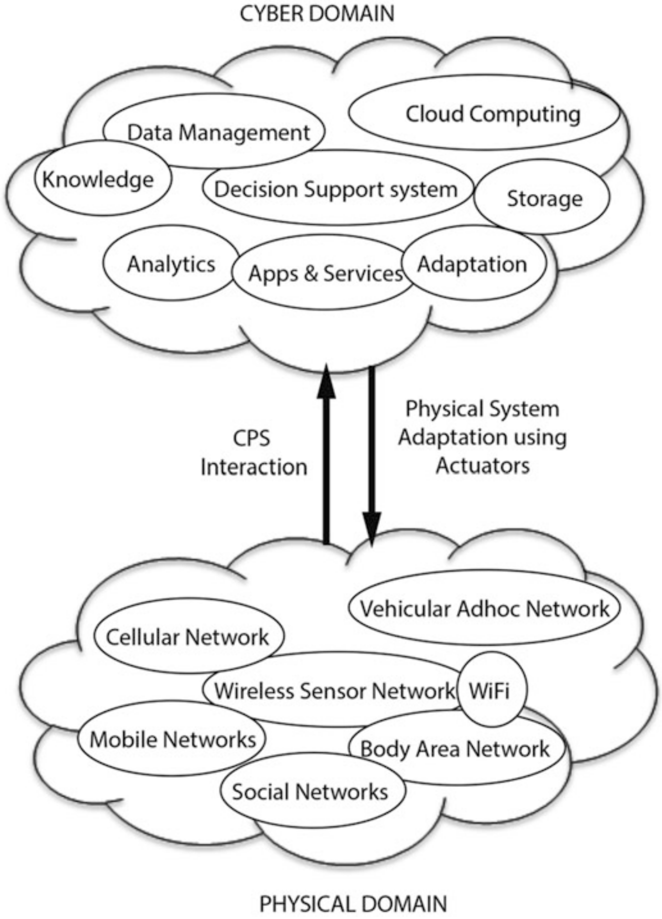

Fig. 6.1 Generic cyber-physical system

systems (CPSs) may support a new wave of computing by actively engaging it with the real world in real time [2]. Cyber-physical system is a *new generation of systems with integrated computational and physical capabilities that can interact with humans through many new modalities* [3]. It is a bridge between the cyber world and the physical world [4], where the physical world is simply the real world and the cyber world comprised of computing paradigms.

A cyber-physical system is the integration of the physical world with the cyber world to monitor and control physical entities by using feedback loops. It is an emerging technology which provides computing and communication facilities to the real-world systems and adds intelligence to the physical entities (see Fig. 6.1). CPS uses digital capabilities of computing to control analog physical systems.

In cyber-physical systems, multiple static or mobile sensors and actuators may be used that are integrated with intelligent decision support systems [5, 6]. The sensors are constrained due to low energy, low computational power, and less storage capacity. Also a sensor does not have enough storage capacity to accommodate huge datasets. Cloud computing is a solution to some of these issues related to sensors. The combination of sensors and cloud is known as sensor cloud [7]. The sensor cloud infrastructure is a vital part of CPS, where the cloud performs computing (cyber) activities and sensor supports physical activities [7].

Kim et al. [8] proposed a generic framework for design, modeling, and simulation of CPS. The paper highlights many important features that need to be part of that framework. The features include heterogeneous application support, physical modeling environments that support mathematical expressions, scalability that helps to increase in number of sensors deployed, support to connect with existing simulation tools, and software reusability, and all the proprietary solutions and open standards should be integrated into generic framework [8].

Due to the increase in the use of smart computing devices, a huge amount of data is generated across the physical world. The term big data is used for those huge datasets and is defined as *a massive volume of both structured and unstructured data that is so large that it is difficult to process using traditional database and software techniques* [9]. There are many sources of big data at the physical world; there may be wireless sensor networks, social networks, wireless body area networks, mobile networks and vehicular ad hoc networks, etc. The data are physically managed through some data management frameworks, and then that captured data is sent to the cyber world for analytics.

The cyber world may include big data, big data management, cloud storage, data analytics, and decision support systems. The continuous data growth poses many challenges. The major issues are storing that data and extracting valuable information from such a large amount of data. The data is not limited, but it is increasing exponentially so there is a major issue to store this data efficiently and in a cost-effective manner. The cloud storage provides a cost-effective way to facilitate the users with ease of computing, storing, and networking resources. As the big data is in large quantity and all of that data is not important, there is a need to extract valuable data through data analytics. Big data analytics is the process of capturing, arranging, and analyzing huge sets of data to identify patterns and valuable information [10]. The analyzed data will be sent back to the physical world. Big data is a buzz word today, so it provides wide space for research in this field. This chapter presents the review of various technical aspects of big data for cyber-physical systems.

The remaining chapter is organized as follows. Section 6.2 discusses various sources of big data. Section 6.3 briefly describes data management at cyberspace that includes cloud computing and decision support systems. Interfacing cyber and physical worlds is discussed in Sect. 6.4. Section 6.5 identifies the main challenges of cyber physical systems in terms of big data, and Sect. 6.6 concludes the chapter.

6.2 Data Generation by Physical Systems: Big Data Sources

The first step in big data scenario is data generation. There are many sources of big data which are generating highly diverse and complex datasets. These sources include wireless sensor networks, mobile ad hoc networks, social networks, vehicular networks, RFIDs, web servers, online transactions, etc.

Big data can be structured, unstructured, and semistructured. The data, which are well organized and are based on some data model, are referred to as structured data. On the other hand, the unstructured data does not follow any data model. The semistructured data is the combination of structured and unstructured. It is a type of structured data, but somehow it lacks the data model structure and uses markers or tags to mark specific data elements. For example, emails contain unstructured data, but it has some fields like date, time, sender, recipient, etc. which are considered to be as structured data. Generally, big data is considered as unstructured.

There are three main characteristics of big data: volume, variety, and velocity [11]. The *volume* characteristic is defined as the amount of data, *variety* as different formats of data/data sources, and *velocity* is the speed at which the data is growing [12]. The data is not just large in volume, but there is variety of complex datasets. The real challenge is to handle that diversity and variety. We can categorize the data growth as business application data, personal data, and machine data [13]. The data generated by business applications is moderate in volume, variety, and velocity. This type of data is highly structured data. It includes online transactions. The personal data includes web logs, documents, emails, social media, etc. It is highly unstructured data, and it is moderate in variety but high in volume and velocity. The data growth is two times more than business application data. The third category is machine data which include sensors, machine logs data, audio and video recordings, bio-informatics, etc. This type of data is highly structured, and it is high in volume, variety, and velocity [14]. The growth is three times more than business application data.

In this section, we discuss a number of common data sources such as wireless sensor networks, social networks, body area networks, and vehicular ad hoc networks.

6.2.1 Wireless Sensor Networks

In the past few years, the applications of wireless sensor networks (WSNs) have been increasing rapidly, such as monitoring, event detection, surveillance, etc. Wireless sensor network is a wireless network of many small devices which are capable of sensing, computation, and communication. A sensor network consists of multiple sensor nodes, which are small and lightweight. The sensor nodes are generally dispersed in a sensor field as shown in Fig. 6.2. Every sensor node contains a transducer, microcomputer, transceiver, and a power source [15, 16]. When a sensor node senses a physical phenomenon, an electrical signal is

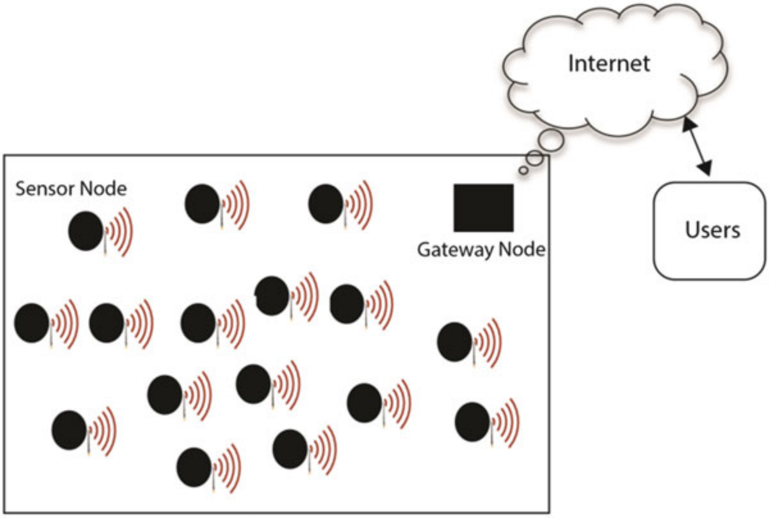

Fig. 6.2 Generic wireless sensor network model

generated by the transducer, which is processed and stored by microcomputer. The collected data will then be sent to the sink/gateway which sends it back to the end user via Internet/satellite or any other means of communication [17].

The traditional technologies for data processing, storing, and reporting provide limited support for analyzing WSN data. These technologies have become prohibitively expensive, while dealing with sensor-generated big data [18]. Even then, they cannot handle the processing requirements for real-time processes such as fire detection, natural disasters, and traffic control [18]. Thus, the research is directed toward new technologies for processing big data. There are many attempts in combining big data and WSN. Jardak et al. [19] proposed a data model for structuring the stored data by allowing a wide range of analysis by using Bigtable, Hadoop, and MapReduce algorithms. A Hadoop-based cloud storage solution for WSN data is presented by Fan et al. [20]. Similarly, Ahmed et al. [21] proposed an infrastructure for integrating cloud computing with WSNs.

6.2.2 Social Networks

A social network is a Web-based service that connects people with each other. The user can make their profile and share information, experiences, ideas, etc. The most popular social networking sites are Twitter [22], Facebook [23], Pinterest [24], Google+ [25], Instagram [26], and many more. These social sites are very popular among youngsters. Data is being shared on these networks, which include billions of photos, videos, and other information. Most of the data is unstructured with high volume and velocity.

The social networks can have enormous benefits for the society as it can help in disasters. Through the social networks, the information about disasters can quickly be disseminated among the people. The most prominent and widely used social media networks like Twitter and Facebook are playing an important role in the propagation of information which could be of different genres. The widespread use of hashtag trends can help with easy access of the latest trends going on. In case of any disaster or catastrophic crisis, the faster spread of information through these sites could be an epidemic in saving lives and providing assistance for the further course of action. One of the crisis situation examples could be the current deadliest earthquake which struck Nepal in April 25, 2015 [27, 28], leaving behind thousands of people dead and other severe casualties. It was within minutes that this news broke through the whole social networks and spread throughout the whole world. Immediate actions were taken to help the people affected by this devastating tragedy. This was due to social networking sites which showed the world how severe the situation was, and because of it various rescue and relief aids were instantly sent from around the world to Nepal. Social networks have become a binding force in the world where within seconds information could be propagated from one corner of the world to the other. The only disadvantage is that we cannot verify the credibility of the information being generated on the social networking sites. Furthermore, social media analysis can be helpful for the organizations to redesign their policies to address the public issues [29]. Social networks are leading toward a new generation of crowd sourcing applications [30], which will help in analyzing in-depth physical environments.

6.2.3 Vehicular Ad Hoc Networks

The research on integrating communication technologies with vehicles has begun since long ago. The communication between vehicles by using ad hoc networks is known as vehicular ad hoc networks (VANETs) [31, 32] as shown in Fig. 6.3. It is a subcategory of intelligent transport systems and mobile ad hoc networks. The vehicles can share necessary information with each other (referred to as vehicle

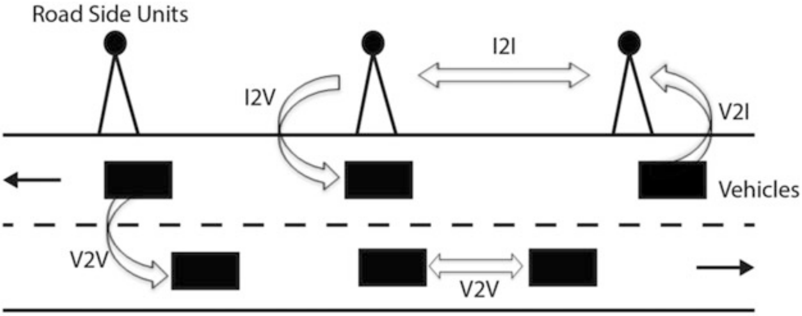

Fig. 6.3 Generic vehicular ad hoc network
]

to vehicle or V2V), such as traffic information (traffic jam or accident), emergency warnings, weather information, road condition warnings, etc. Furthermore, the data can also be shared with the data center to the passing vehicles (referred as infrastructure to vehicle or I2V and vice versa). It is not only important to pass along the latest information but also to remove the outdated data [33, 34]. Since there are many vehicles passing on the roads, they may consume the total bandwidth in data dissemination. Therefore, it is important to efficiently transmit the data by using limited bandwidth [35].

6.2.4 Wireless Body Area Networks

The recent development in wireless networks and microelectronics has resulted in wireless body area networks (WBANs). A WBAN may consist of miniature lightweight sensor nodes with low power and is used for healthcare application to monitor physiological status of the human body, such as blood pressure, blood sugar, ECG, pulse rate, etc. [36].

Figure 6.4 shows a generic architecture for WBAN-based health monitoring system where sensor nodes senses medical data and sends it to the base station (BS).

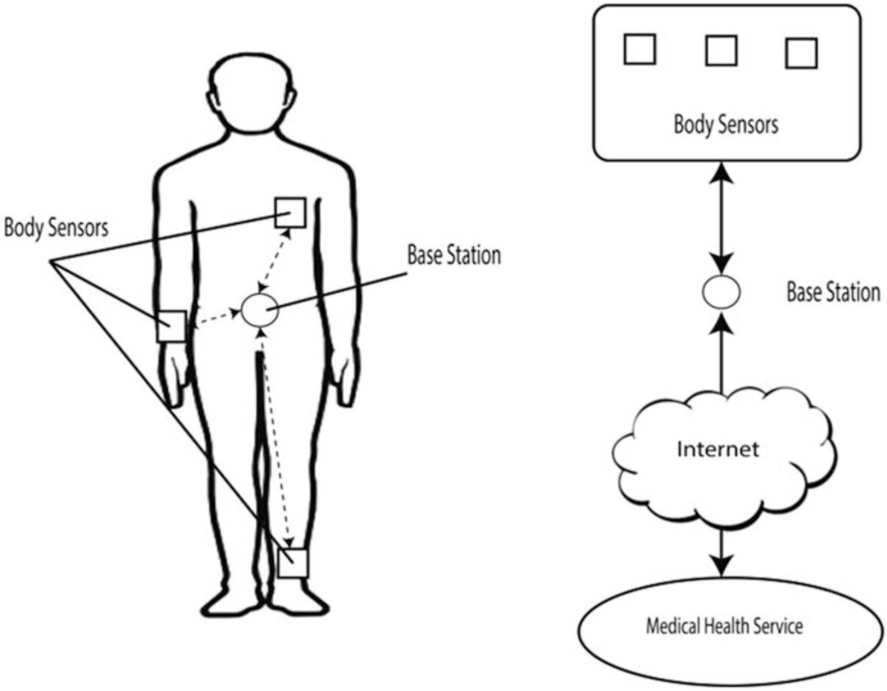

Fig. 6.4 Generic WBAN scenario

Table 6.1 Comparison of big data sources

Big data sources	Volume	Velocity	Variety	Structured/unstructured
WSN	High	High	Low	Structured
Social networks	Moderate to high	High	Low to moderate	Unstructured
VANET	Moderate to high	Moderate	Low to moderate	Structured
WBAN	High	High	High	Unstructured

The BS then transmits the data to the doctor for real-time diagnosis, to a database for keeping medical records or to a particular device that generates emergency alerts via medical health service (MHS) center [37, 38].

The common factor in all the emerging data sources is that they all are continuously generating data. The data is high in volume, with lots of variety and with dynamic velocities [39]. In general, the characteristics of big data for these data sources are summarized in Table 6.1. The data from these four sources are mostly high in volume and velocity. As the WSN, VANET, and WBAN data are sensor-generated data, so they have a vast data variety depending on the nature of sensors used and are mostly highly structured. The social network data is moderate to high in volume but does not have much data variety. This type of data is mostly unstructured, as it mostly contains images and audio and video streams.

6.3 Data in Cyber Systems: Big Data Management

Once the raw data is collected from the physical world, it is passed to the cyber world for further processing. In this section, we will mainly consider how the data is managed in the cyber world and what the prerequisites are. This section also highlights the role of existing cloud computing paradigm and decision support systems that solves the problem of storage and provides other computational facilities.

6.3.1 Cloud Computing Paradigms

For the past few years, cloud computing and big data have been the two key fields which gained significant attention of the researchers [40, 41]. A few years back, large datasets assumed to be of a few terabytes, but nowadays this concept has been changed, and individual applications are producing more than that: new units being used as terabytes and petabytes. With this continuous growth of data, it is difficult for an organization to handle the data which is too big, too versatile, and too fast because the traditional storage methods are not designed for such a huge data. One solution is cloud storage. Cloud computing is the emerging technology that provides users to perform complex computations without maintaining expensive hardware and software. Although cloud computing is currently used by almost all the leading companies, still there is no universally agreed definition [42]. Gartner [43] defines cloud computing as a computing style which provides scalable and elastic IT-enabled capabilities as a service by using Internet technologies. It provides many computational services to the users such as infrastructure as a service (IaaS) [44], platform as a service (PaaS) [44], and software as a service (SaaS) [44]. The IaaS offers storage and processing infrastructure as a service. The user is provided virtualized infrastructure without worrying about hardware resources [45]. The PaaS provides a platform for the software developers to write and upload their application code [45]. The SaaS is the most common layer of cloud computing. It provides software application as a service to users on pay-as-per-use basis [45]. Cloud computing has many advantages over other computational services which includes parallel processing, security, scalable data storage, and resource virtualization [46]. It also reduces maintenance of infrastructure. Cloud computing supports virtualization; through virtualization software, a simple computer can behave like a supercomputer at an affordable cost [46].

There is a big challenge for researchers to design an appropriate platform for cloud computing that handles it and performs data analytics. There are many cloud service providers, whenever an enterprise tend to migrate from IT system to the cloud, the decision can be difficult, as the enterprise want to evaluate cost, benefits, and risks of using cloud computing [47]. Hashem et al. [47] presented two decision support tools for migration to the public IaaS cloud. These tools help an enterprise to make cloud migration decisions. The first tool is a cost modeling tool [47], which can be used in modeling the requirements of enterprise data, applications, and infrastructure along with the usage patterns of computational resource. This tool can also be used to compare the cost of cloud services from different cloud service providers with different deployment options and usage. The second tool, which is a spreadsheet, shows the usage benefits of IaaS cloud, and it also provides beginning point for risk assessment [47].

The data management applications in the cloud include two main data management components: transactional data management and analytical data management [48]. Transactional data management is the database related with transactions like banking and online reservations [49]. Shared nothing is simply a distributed architecture in which each node consists of a processor, main memory, and disk, and the nodes communicate with each other via interconnecting network [50]. For implementation of transactional data management, the usage of shared nothing architecture may result in the complex distributed locking and commit protocols [48]. There is an extensive risk of storing transactional data on untrusted host because of the sensitive information which includes credit card numbers and pin codes. Analytical data management is for the applications that query a data store for business intelligence and decision making purposes [48]. Its scale is larger than transactional database management. The analytical data management systems are best suited for execution in a cloud environment. Generally, for implementation of analytical data management, the use of shared nothing architecture is well suited. The atomicity, consistency, and isolation are easy to obtain as compared to transactional data.

6.3.2 Service-Oriented Decision Support Systems

Decision support system (DSS) is a computer application that analyzes business data and presents it in a way so that users can make business decisions more easily [51]. It may use artificial intelligence for analyzing data. DSS finds certain patterns in data which helps humans to take decisions. For example, DSS helps doctors to diagnose the disease on the basis of symptoms.

There are three service models for service-oriented DSS [52], namely, data as a service (DaaS), information as a service (IaaS), and analytics as a service (AaaS). These are discussed in the following subsections.

6.3.2.1 Data as a Service

The service-oriented architecture provides access to the data from anywhere, independent of the platform. The data as a service provides the business applications, a facility to access the data wherever it resides [53]. With the provision of DaaS, the data quality can be maintained at central place, i.e., at cloud. According to Demirkan et al. [52], the data cleansing and data enriching can be done by two solutions, namely, master data management (MDM) and customer data integration (CDI), where customer data can be placed anywhere and can be accessed as a service through any application that has service provision.

6.3.2.2 Information as a Service

Sometimes the information repositories at organizations are not efficiently designed to transmit information to the required destinations; this is due to the increased complexity of processes and architectures. Demirkan et al. [52] defines information as a service as an idea to make information quickly available to users, processes, and applications in an organization. This shares real-time information with emerging applications and hides complexities. It increases availability with virtualization. It also provides master data management (MDM), content management services, and business intelligence services [52].

6.3.2.3 Analytics as a Service

Analytics as a Service can be defined as the combination of cloud computing and big data analytics [53]. It enables data scientists to access datasets that are centrally managed by cloud providers. The business analysts can make decisions effectively and delivers successful outcomes. AaaS is a cloud-based analytical platform, where several data analytical tools are available, which can be configured by end users to process and analyze large amounts of heterogeneous data.

6.4 Interfacing Cyber World with Physical World

The error-free interaction between cyber and physical worlds is not an easy task because the physical world is mostly unpredictable. The most critical element is that human lives are dependent on the system. Thus, availability, connectivity, predictability, and repeatability are very much important for the cyber-physical interface [54].

In general, CPS can be divided into three parts: physical world, cyber world, and interfacing physical and cyber worlds. The sensors sense physical characteristics, then the sensed data is passed to cyber world to perform computations, and finally, response is generated through some actuators (see Fig. 6.1).

The physical components may include power sources, energy storage, and physical transducers that perform energy conversions in physical domains [55]. The cyber components may include data stores, computation, and I/O interfaces. The interface contains both physical and cyber components and adds a few more components to connect them. Rajhans et al. [56] presented two connector types for modeling the interface between cyber and physical worlds. The connectors are physical-to-cyber (P2C) and cyber-to-physical (C2P) connectors. Simple sensors can be used for physical-to-cyber connector type and actuators can be used for cyber-to-physical connector type. For the complex interfaces in CPS, physical-to-cyber transducer and cyber-to-physical transducer may also be used, which have ports to cyber and physical components on each side [56].

Applications of CPS apparently have potential to overshadow the twentieth-century IT advancements. CPS applications include many components that cooperate through an unpredictable physical environment. In this regard, reliability and security are major issues to be resolved. The CPS applications include transportation, defense, energy and industrial automation, health and biomedical, agriculture, and other critical infrastructures [57].

Cyber-physical cloud computing is the integration of CPS and cloud computing. The CPCC architectural framework is “a system environment that can rapidly build, modify, and provision cyber physical systems composed of a set of cloud computing based sensor, processing, control, and data services” [58]. The customers can access available resources through Internet independent of location and devices.

CPCC automatically manages all the resources. CPCC can benefit many systems such as traffic management, intelligent power grids, disaster management systems, healthcare, etc. [58].

Despite the fact that data management is the focal point of interest for many researchers, there is still lack of an agreed-upon definition for data management. It includes many phases. The data management phases are defined by many researchers; these steps depend on the nature of data to be managed. Some phases can be added or removed accordingly. We will discuss some of them in this section. In TDWI report [39], the data management is defined as data collection, storage, processing, and delivery, and it considers data management as a broad practice that includes many data disciplines such as data quality, data integration, data warehousing, event processing, database administration and content management, etc. According to Mokashi et al. [59], data management includes data collection, data storage, and query processing. Padgavankar et al. [60] consider data management as a four-step process, i.e., data generation, big data acquisition, big data storage, and data analysis. Furthermore, Sathe et al. [61] use four tasks for WSN data management, i.e., data acquisition, data cleaning, query processing, and data compression.

Accordingly, we classify the main phases of data management for interfacing CPS as data acquisition, data preprocessing, storage, query processing, data analysis, and actuation.

6.4.1 Data Acquisition

Data acquisition is simply data gathering or data collection. As the physical world is generating huge datasets, the data acquisition process determines which data should be collected along with minimal energy consumption. This is challenging task, because of the data uncertainty due to natural errors, noise, and missed readings in sensor data [62]. It is responsible for efficiently collecting samples from sensors in CPS. In sensor data acquisition, the main objective is to achieve energy efficiency because sensors are battery powered and located mostly at unreachable locations. Sathe et al. [61] presented model-based data acquisition techniques that are designed to handle challenges such as minimal energy consumption and communication cost.

6.4.2 Data Preprocessing

In the data acquisition phase, the raw data gets collected. The acquired sensor datasets may sometimes contain erroneous or redundant data, which will definitely occupy more storage space and will affect data analysis [60]. Therefore, before data storage, preprocessing may be applied which may include data cleaning, data fusion, and data compression [60].

6.4.2.1 Data Cleaning

It is a process of finding incomplete, inaccurate, and unreasonable data and then correcting the errors to improve data quality [63]. In the data cleaning process, the errors will be removed from raw sensor data. For incomplete datasets, regression or interpolation models can be used to reconstruct missing data. Alonso et al. [64] proposed an extensible receptor stream processing (ESP) framework for online data cleaning of the acquired sensor data streams.

6.4.2.2 Data Fusion

As the name shows, data fusion combines data from various data sources. Sensor fusion is a technique to merge data from many sources to provide accurate and comprehensive information [65]. It is a technique to address sensor impairments. Some other terms are also used in the literature that are related to data fusion such as decision fusion, multisensor fusion, and information fusion.

Data fusion is a technique of combining data from various sensors and information from related databases to attain accuracy and more specific inferences than by using a single sensor [66]. The data fusion techniques can be categorized into three categories, that is, data association, state estimation, and decision fusion [67].

6.4.2.3 Data Compression

As the sensors continuously generate huge datasets, and sometimes the collected data contains redundant data, this redundancy is common in environmental monitoring. The data compression techniques can help to reduce the redundancy which helps in reducing storage space [60]. Data transmission is more energy consuming than computation; thus, reduced data size, before data transmission, will minimize the overall energy consumption [68]. Different data compression schemes have been discussed by Kimura et al. [68]. Sathe et al. [61] also discussed many data compression techniques such as linear approximation, model approximation, and orthogonal transformation. Marcelloni et al. [69] introduces a new lossless compression algorithm that is suitable for reduced computational and storage resources of a WSN node. Many other techniques are proposed in the literature, some of them are derived from signal processing [70] and some has used correlations between sensor data to compress the data streams [71–73].

6.4.3 Data Storage

Due to the limitations of sensors, it is important to store the sensor data efficiently elsewhere [74], to improve the data retrieval and analytics processes. As the sensors generate huge datasets, the question arise: do all the generated data is required to be stored? For different applications, the answer could be different. For example, in real-time applications, mostly, the recent data is important [75], so there is no need to keep all the data for long periods. In such cases, live data streaming will be an appropriate approach. In some applications, the historical data need to be stored for future analysis, in those cases; the historical storage approach will be appropriate [76]. For a variety of applications, both the storage approaches can be combined to make an efficient storage system, but this could be very challenging [75].

The second important issue is determine where to store that data. Many researchers have done work in this direction. Three data storage methods for WSN have been discussed by Xing et al. [74]; these are local storage, external storage, and data centric storage (DCS). In local storage, only short-lived data is stored in the sensor node. In external storage, data is stored at an external point for further processing. While in data-centric storage, data is stored along with the name or location. In DCS, the related data is classified and named according to its meaning. The data with the same name will be stored in the same sensor node. For a particular name, the user queries will be sent directly to that particular node which holds that named data [59].

For live querying data, many data management techniques have been proposed, whereas for querying historical data, only a few data management solutions have been developed. Those techniques are discussed by Diao et al. [76].

With the technology advancements, the storage devices are becoming more energy efficient and cheaper in price. Thus the sensor networks are transforming from communication-centric to storage-centric perspective which provides a network that efficiently stores data from sensors [76, 77]. The data can be batched or accessed later. Energy efficiency can be improved by batching sensed data. In storage-centric sensor network, the applications must be delay tolerant because the data is not transmitted immediately. For the applications where immediate response is needed, delay cannot be tolerated, in such cases communication-centric approach is appropriate [77].

6.4.4 Query Processing

Another important component of data management is data retrieval or query processing. Many important model-based query processing techniques, which aim to process queries by retrieving minimum amount of data, are presented by Sathe et al. [61]. Apart from that, these techniques also handle missing data and create an abstraction layer on sensor network by using these models [78, 79]. Some of the techniques are based on hidden Markov model (HMM) [80] or dynamic probabilistic model which is for spatiotemporal evolution of the data from sensors [79].

For querying the real-time applications of CPS, different researchers have developed many tools that can be named as information flow processing (IFP) systems [81]. Information flow processing is an application domain where users are required to collect data from various data sources to process it within due time [81]. After processing data, the collected data is generally discarded, except some critical applications where historical analysis is important. This is done using two popular models that are data stream processing [82] and complex event processing models [83]. The data stream processing model is processing data from various sources to produce output data streams. The data stream management system (DSMS) is also based on database management systems (DBMS) with a few differences such as DBMSs deals with data that is not updated constantly, whereas DSMSs are specially designed to deal with data that is updated continuously [81]. Apart from few differences, there are more similarities in between both of them.

The recent developments in DSMSs are reviewed by Golab et al. [84]. The complex event processing model considers information flow items as notification of events of the physical world, which will be filtered and combined to visualize what is happening in the form of high-level events [81]. The approach mainly focuses to detect patterns of low-level events that will eventually be combined to represent the high-level events that will be notified to the parties that are interested. An architecture for real-time analysis and processing of complex-event streams of sensor networks, which is based on semantically rich event models, is presented by Dunkel et al. [85].

6.4.5 Data Analysis

The data analysis is the most significant phase of data management for CPS interfacing. As CPS data is larger in magnitude, the real challenge is to extract the insight value from it which is valuable. The purpose of data analysis is to sift valuable information. It helps organizations to better cope with the needs of their customers and to make better decisions [86].

The traditional analytical methods, based on statistics and computer science, may still be used for big data analysis, such as cluster analysis, factor analysis, correlation analysis, regression analysis, real-time analysis, offline analysis, memory level analysis, business intelligence (BI) analysis, and massive analysis [60]. Some advanced analysis techniques are also required to handle the complexity of real-world heterogeneous datasets. Big data analytics is the set of modern techniques that are designed to operate on heterogeneous data with large magnitudes [87]. The intelligent quantitative methods, such as artificial intelligence, robotics, artificial neural networks, or machine learning, can be used to explore and to identify hidden patterns and their relationships [87].

In a typical CPS for environment monitoring, most of the collected data is considered as regular, but some of them may be irregular; such data is known as atypical data [88]. Atypical data is extremely crucial as it identifies a change in environmental condition; therefore, such data need to be analyzed. Different approaches have been discussed in the literature [88–90] to analyze atypical data in the CPS. Tang et al. [91] proposed a method named as Tru-alarm, which finds out trustworthy alarms for the cyber-physical systems. It uses data analysis to eliminate noisy data that can cause false emergency alarms.

6.4.6 Actuation

Actuation is the most crucial element in CPS because it controls the environment. In many CPS applications, sensing data is not just sufficient, but a response is also required to show how the system reacts in a particular situation [92]. For example, in fire alarm systems, the actuators may be deployed to shower water on the fire. Another example could be of an agricultural environment where crops can be monitored and pesticides can be sprinkled by an actuation process if needed.

Data actuation is the process in which the processed data is sent to actuators to perform some action. It transfers data back to the physical systems. Thouin et al. [93] discussed different actuation strategies to acquire desired actions to be performed on physical devices. A dynamic actuation strategy is group of decision rules to find the actuation nature which will be executed throughout the course of operations in a wireless sensor and actuator networks [93].

6.5 Future Challenges and Opportunities

The CPS is a multidisciplinary technology, which involves communication and networks, embedded systems, and semantic technologies. To take the maximum benefit from CPS and to handle big data that flows in between the cyber and physical worlds, there are many challenges to be addressed. A few challenges of CPS in big data perspective are given below.

Volume As discussed earlier, CPS data is enormous and keeps growing continuously, processing that huge dataset can lead to many challenges. Data abstraction, i.e., summarizing the data and making it human comprehensible, is one of the biggest challenge for big data generated across CPS. Another challenge could be efficient use of distributed processing to scale the CPS computations. The simple computations can become complex when scaling from terabytes to petabytes. Even sequential scans to petabytes data takes too much time. The indexing techniques are also very challenging while scaling to huge volume.

Variety There is a huge variety of datasets with different data formats, which need to be integrated together. As the data is collected from distinct sources, the structure of data can be complex and data processing can also be very complex. Thus, efficient techniques are needed to cope with the increasing variety of data.

Velocity With these fast growing datasets, it is challenging to focus on the data trends and the correlations between data. There is a great need of robust and real-time techniques to cope with velocity of the data generation and processing.

Veracity Sometimes, the sensors in a CPS generate erroneous data, or some data is missed due to erroneous communication. Therefore, it is challenging to find trustworthiness of the data.

Value The main challenge of interfacing CPS is to transform the collected raw data into useful information in order to facilitate the decision making process. The efficient transformation techniques are needed to provide the accurate value of the information.

Query Load Generally, the query loads vary and are unpredictable. Due to lack of flexibility, it is complex to handle these variations. Conti et al. [94] proposed a new term data vitalization to sense the query load variations. There is still lot of work needed to optimize the query load in accordance with the needed information and available resources.

Quality of Service The methodologies to precisely capture and communicate information and the quality needs of an application should be researched. Due to increase in scalability and complexity of data, the computational techniques and their results are very complex to reproduce. Thus, the relationship between data from information producing systems and the operational systems need to be studied such that application's quality of service requirements are fulfilled efficiently.

Knowledge Association The constant sensor data streams are required to be processed by CPS. These streams need to be efficiently associated with the existing knowledge [6, 58]. For the complex and uncertain data, the temporal and spatial correlations must be used with data mining tools to retrieve valuable knowledge [6]. There is very little work in this direction, and more research is needed for efficient knowledge association across CPS.

Open CPS Architecture A new open architecture is required, which can be customized in different situations by different application scenarios. The physical components are mostly unreliable; tools are needed to build a reliable CPS that should be resilient to tolerate malicious attacks on the data [57]. For the complex design of CPS, new modeling and analytical tools are essential to be utilized [95].

6.6 Conclusion

In the era of advanced computing, there is an emergent and rapid technological enhancements in the fields of embedded systems, human computer interaction, cloud computing, data analysis, cyber-physical systems, and many other computing aspects. Cyber-physical systems, a new wave of computing, have enabled many applications that were not practical before. The data from cyber-physical systems is enormous and growing constantly which poses many challenges in this field. This chapter discusses the state-of-the-art of the cyber-physical systems from big data perspective. The data generation sources, cyberspace paradigms, and interfacing them with the physical and cyber world have been discussed. From data generation to its storage, different phases of data management for interfacing the two worlds have also been elaborated. The main issues are efficient storage and processing of cyber-physical systems big data. The cyber-physical system cloud computing infrastructure has also been discussed which provides the framework to interface with the computing devices. Furthermore, the research issues related to big data in cyber-physical systems have been highlighted. The cyber-physical systems are in its way of development; therefore, significant issues and challenges must be addressed by researchers for long-term success.

References

1. Buyya R, Yeo CS, Venugopal S, Broberg J, Brandic I (2009) Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility. Futur Gener Comput Syst 25(6):599–616
2. Wolf W (2009) Cyber-physical systems. Computer 3:88–89
3. Baheti R, Gill H (2011) Cyber-physical systems. Impact Control Technol 12:161–166
4. Rajkumar R, Lee I, Sha L, Stankovic J (2010) Cyber-physical systems: the next computing revolution. In: Proceedings of the 47th design automation conference. ACM, pp 731–736
5. Shaikh FK, Zeadally S (2015) Mobile sensors in cyber-physical systems. Book Chapter in cyber physical system design with sensor networking technologies, IET, 2015 (to appear)
6. Wu FJ, Kao YF, Tseng YC (2011) From wireless sensor networks towards cyber physical systems. Pervasive Mob Comput 7(4):397–413
7. Haque AS, Aziz SM, Rahman M (2014) Review of cyber-physical system in healthcare. Int J Distrib Sens Netw 2014:1–20
8. Kim JE, Mosse D (2008) Generic framework for design, modeling and simulation of cyber physical systems. ACM SIGBED Rev 5:1
9. Bloomberg J (2013) The big data long tail. http://www.devx.com/blog/the-big-data-long-tail.html. Accessed 17 Jan 2015
10. Kambatla K, Kollias G, Kumar V, Grama A (2014) Trends in big data analytics. J Parallel Distr Com 74(7):2561–2573
11. Madden S (2012) From databases to big data. IEEE Internet Comput 3:4–6
12. Rouse M (2015) 3Vs (volume, velocity & variety). http://whatis.techtarget.com/definition/3Vs. Accessed Apr 2015
13. Hitachi Data Systems (2015) Capitalize on big data. http://www.hds.com/assets/pdf/hitachi-webtech-educational-series-capitalize-on-big-data.pdf. Accessed 20 Mar 2015

14. Hurwitz J, Nugent A, Halper F, Kaufman M (2015) Structured data in a big data environment. www.dummies.com/howto/ content/structured-data-in-a-big-data-environment.html. Accessed 2 Apr 2015
15. Shaikh FK, Zeadally S, Siddiqui F (2013) Energy efficient routing in wireless sensor networks. In: Next-generation wireless technologies. Springer, London, pp 131–157
16. Rouse M (2006) Wireless sensor networks. http://searchdatacenter.techtarget.com/definition/sensor-network. Accessed 20 Feb 2015
17. Akyildiz IF, Vuran MC (2010) Wireless sensor networks, 4th edn. Wiley, New York
18. Rios LG, Diguez JEAI (2014) Big data infrastructure for analyzing data generated by wireless sensor networks. In: IEEE international congress on big data (BigData Congress), 2014. IEEE, pp 816–823
19. Jardak C, Riihijärvi J, Oldewurtel F, Mähönen P (2010) Parallel processing of data from very large-scale wireless sensor networks. In: Proceedings of the 19th ACM international symposium on high performance distributed computing. ACM, pp 787–794
20. Fan T, Zhang X, Gao F (2013) Cloud storage solution for WSN in internet innovation union. Int J Database Theory Appl 6(3):49–58
21. Ahmed K, Gregory M (2011) Integrating wireless sensor networks with cloud computing. In: Seventh international conference on Mobile Ad-hoc and Sensor Networks (MSN), 2011. IEEE, pp 364–366
22. Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media?. In: Proceedings of the 19th international conference on world wide web. ACM, pp 591–600
23. Ellison NB, Steinfield C, Lampe C (2007) The benefits of Facebook "friends:" social capital and college students' use of online social network sites. J Comput-Mediat Commun 12 (4):1143–1168
24. Gilbert E, Bakhshi S, Chang S, Terveen L (2013) I need to try this?: a statistical overview of pinterest. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp 2427–2436
25. Shervington M (2015) What is google Plus? A complete user guide. http://www.martinshervington.com/what-is-google-plus/. Accessed 20 Apr 2015
26. Hochman N, Schwartz R (2012) Visualizing instagram: tracing cultural visual rhythms. In: Proceedings of the workshop on Social Media Visualization (SocMedVis) in conjunction with the sixth international AAAI conference on Weblogs and Social Media (ICWSM–12), pp 6–9
27. Watson I, Mullen J, Smith-Spark L (2015) CNN. Nepal earthquake: death toll passes 4,800 as rescuers face challenges. http://edition.cnn.com/2015/04/28/asia/nepal-earthquake/. Accessed on 05 May 2015
28. Ravilious K (2015) Nepal quake 'followed historic pattern'. http://www.bbc.com/news/science-environment-32472310. Accessed on 28 Apr 2015
29. Garg Y, Chatterjee N (2014) Sentiment analysis of Twitter feeds. In: Big data analytics. Springer International Publishing Switzerland, pp 33–52
30. Felemban E, Sheikh AA, Shaikh FK (2014) MMaPFlow: a crowd-sourcing based approach for mapping mass pedestrian flow. In: Proceedings of the 11th international conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MOBIQUITOUS '14)
31. Yousefi S, Mousavi MS, Fathy M (2006) Vehicular ad hoc networks (VANETs): challenges and perspectives. In: 6th international conference on ITS telecommunications proceedings, 2006. IEEE, pp 761–766
32. Ali F, Shaikh FK, Ansari AQ, Mahoto NA, Felemban E (2015) Comparative analysis of VANET routing protocols- on placement of road side units. Int J Wirel Pers Commun, Springer, pp 1–14, 2015. doi:10.1007/s11277-015-2745-z
33. Zhang Y, Zhao J, Cao G (2010) Roadcast: a popularity aware content sharing scheme in vanets. ACM SIGMOBILE Mobile Comput Commun Rev 13(4):1–14
34. Sutariya D, Pradhan SN (2010) Data dissemination techniques in vehicular ad hoc network. Int J Comput Appl 8(10):35–39

35. Dubey BB, Chauhan N, Kumar P (2010) A survey on data dissemination techniques used in VANETs. Int J Comput Appl 10(7):5–10
36. Talpur A, Baloch N, Bohra N, Shaikh FK, Felemban E (2014) Analyzing the impact of body postures and power on communication in WBAN. Procedia Comput Sci 32:894–899
37. Khelil A, Shaikh FK, Sheikh AA, Felemban E, Bojan H (2014) DigiAID: a wearable health platform for automated self-tagging in emergency cases, In: 4th international conference on wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI, pp 296,299
38. Aziz Z, Qureshi UM, Shaikh FK, Bohra N, Khelil A, Felemban E (2015) Revisiting routing in wireless body area networks. In: Emerging communication technologies based on wireless sensor networks: current research and future applications. CRC Press (to appear)
39. TDWI Best Practices Report (2015) Managing big data. http://tdwi.org/research/2013/10/tdwi-best-practices-report-managing-big-data.aspx?tc=page0. Accessed 01 Mar 2015
40. Dinh HT, Lee C, Niyato D, Wang P (2013) A survey of mobile cloud computing: architecture, applications, and approaches. Wirel Commun Mob Comput 13(18):1587–1611
41. Agrawal D, Das S, El Abbadi A (2010) Big data and cloud computing: new wine or just new bottles? Proc VLDB Endowment 3(1–2):1647–1648
42. Elazhary H (2014) Cloud computing for big data. MAGNT Res Rep 2(4):135–144
43. Gartner IT glossary (2013) Cloud computing. http://www.gartner.com/it-glossary/cloud-computing. Accessed 01 Apr 2015
44. Rodero-Merino L, Vaquero LM, Gil V, Galán F, Fontán J, Montero RS, Llorente IM (2010) From infrastructure delivery to service management in clouds. Futur Gener Comput Syst 26 (8):1226–1240
45. Patidar S, Rane D, Jain P (2012) A survey paper on cloud computing. In: Second international conference on Advanced Computing & Communication Technologies (ACCT), 2012. IEEE, pp 394–398
46. Khajeh-Hosseini A, Sommerville I, Bogaerts J, Teregowda P (2011) Decision support tools for cloud migration in the enterprise. In: IEEE international conference on Cloud Computing (CLOUD), 2011. IEEE, pp 541–548
47. Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU (2015) The rise of "big data" on cloud computing: review and open research issues. Inf Syst 47:98–115
48. Abadi DJ (2009) Data management in the cloud: limitations and opportunities. IEEE Data Eng Bull 32(1):3–12
49. Das S, Agrawal D, El Abbadi A (2009) Elastras: an elastic transactional data store in the cloud. USENIX HotCloud 2:7
50. Valduriez P (2009) Shared-memory architecture. In: Encyclopedia of database systems. Springer US, New York, pp 2638–2638
51. Jill Dyche (2015) Data as a service explained and defined. http://searchdatamanagement.techtarget.com/answer/Data-as-a-service-explained-and-defined Accessed on 20 Mar 2015
52. Demirkan H, Delen D (2013) Leveraging the capabilities of service-oriented decision support systems: putting analytics and big data in cloud. Decis Support Syst 55(1):412–421
53. Mathiprakasam M (2015) The road to analytics as a service. http://www.forbes.com/sites/oracle/2014/09/26/the-road-to-analytics-as-a-service/. Accessed on 20 Mar 2015
54. Poovendran R (2010) Cyber–physical systems: close encounters between two parallel worlds [point of view]. Proc IEEE 98(8):1363–1366
55. Shaikh FK, Zeadally S, Exposito E (2015) Enabling technologies for green internet of things. IEEE Syst J 99:1–12
56. Rajhans A, Cheng SW, Schmerl B, Garlan D, Krogh BH, Agbi C, Bhave A (2009) An architectural approach to the design and analysis of cyber-physical systems. Electronic Communications of the EASST, 21:1–10
57. CPS Steering Group (2008) Cyber-physical systems executive summary. CPS Summit
58. Simmon E, Kim KS, Subrahmanian E, Lee R, de Vaulx F, Murakami Y, Zettsu K, Sriram RD (2013) A vision of cyber-physical cloud computing for smart networked systems. NIST, Gaithersburg

59. Mokashi M, Alvi AS (2013) Data management in wireless sensor network: a survey. Int J Adv Res Comput Commun Eng 2:1380–1383
60. Padgavankar MH, Gupta SR (2014) Big data storage and challenges. Int J Comput Sci Inf Technol 5:2
61. Sathe S, Papaioannou TG, Jeung H, Aberer K (2013) A survey of model-based sensor data acquisition and management. In: Managing and mining sensor data. Springer US, New York, pp 9–50
62. Aggarwal CC (2013) Managing and mining sensor data. Springer Science & Business Media, New York
63. Chapman AD (2005) Principles and methods of data cleaning. GBIF, Copenhagen
64. Jeffery SR, Alonso G, Franklin MJ, Hong W, Widom J (2006) A pipelined framework for online cleaning of sensor data streams. IEEE, p 140
65. Elmenreich W (2002) Sensor fusion in time-triggered systems, Ph.D. thesis, Faculty of Informatics at the Vienna University of Technology, Austria. http://www.vmars.tuwien.ac.at/~wilfried/papers/elmenreich_Dissertation_sensorFusionInTimeTriggeredSystems.pdf
66. Hall David L, Llinas J (1997) An introduction to multisensor data fusion. Proc IEEE 85 (1):6–23
67. Castanedo F (2013) A review of data fusion techniques. Sci World J 2013:1–19
68. Kimura N, Latifi S (2005) A survey on data compression in wireless sensor networks. In: International conference on Information Technology: Coding and Computing (ITCC), 2005, vol. 2. IEEE, pp 8–13
69. Marcelloni F, Vecchio M (2008) A simple algorithm for data compression in wireless sensor networks. Commun Lett IEEE 12(6):411–413
70. Agrawal R, Faloutsos C, Swami A (1993) Efficient similarity search in sequence databases. Springer, Berlin/Heidelberg, pp 69–84
71. Gandhi S, Nath S, Suri S, Liu J (2009) Gamps: compressing multi sensor data by grouping and amplitude scaling. In: Proceedings of the 2009 ACM SIGMOD international conference on management of data. ACM, pp 771–784
72. Wang L, Deshpande A (2008) Predictive modeling-based data collection in wireless sensor networks. In: Wireless sensor networks. Springer, Berlin/Heidelberg, pp 34–51
73. Arion A, Jeung H, Aberer K (2011) Efficiently maintaining distributed model-based views on real-time data streams. In: Global Telecommunications Conference (GLOBECOM 2011). IEEE, pp 1–6
74. Xing K, Cheng X, Li J (2005) Location-centric storage for sensor networks. In: IEEE international conference on mobile adhoc and sensor systems conference. IEEE, p 10
75. Petit L, Nafaa A, Jurdak R (2009) Historical data storage for large scale sensor networks. In: Proceedings of the 5th French-speaking conference on mobility and ubiquity computing. ACM, pp 45–52
76. Diao Y, Ganesan D, Mathur G, Shenoy PJ (2007) Rethinking data management for storage-centric sensor networks. In: CIDR, vol. 7, pp 22–31
77. Dutta P, Culler DE, Shenker S (2007) Procrastination might lead to a longer and more useful life. In: The sixth workshop on Hot Topics in Networks (HotNets-VI) pp 1–7
78. Deshpande A, Madden S (2006) MauveDB: supporting model-based user views in database systems. In: Proceedings of the 2006 ACM SIGMOD international conference on management of data. ACM, pp 73–84
79. Kanagal B, Deshpande A (2008) Online filtering, smoothing and probabilistic modeling of streaming data. In: IEEE 24th international conference on Data Engineering, ICDE 2008. IEEE, pp 1160–1169
80. Bhattacharya A, Meka A, Singh AK (2007) Mist: distributed indexing and querying in sensor networks using statistical models. In: Proceedings of the 33rd international conference on very large data bases. VLDB Endowment, pp 854–865
81. Cugola G, Margara A (2012) Processing flows of information: from data stream to complex event processing. ACM Comput Surv (CSUR) 44(3):15

82. Babcock B, Babu S, Datar M, Motwani R, Widom J (2002) Models and issues in data stream systems. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems. ACM, pp 1–16
83. Luckham D (2002) The power of events, vol 204. Addison-Wesley, Reading
84. Golab L, Özsu MT (2003) Issues in data stream management. ACM Sigmod Rec 32(2):5–14
85. Dunkel J (2009) On complex event processing for sensor networks. In: International symposium on autonomous decentralized systems, 2009. ISADS'09. IEEE, pp 1–6
86. Miller S (2013) Big data analytics. Podcasts at Singapore Management University, Available at: http://ink.library.smu.edu.sg/podcasts/8
87. Big Data in the Cloud Converging Technologies-Intel (2014) http://www.intel.com/content/www/us/en/big-data/big-data-cloud-technologies-brief.html. Accessed on Apr 2015
88. Tang LA, Yu X, Kim S, Han J, Peng WC, Sun Y, Gonzalez H, Seith S (2012) Multidimensional analysis of atypical events in cyber-physical data. In: IEEE 28th international conference on Data Engineering (ICDE), 2012. IEEE, pp 1025–1036
89. Tang LA, Yu X, Kim S, Han J, Peng WC, Sun Y, Leung A, La Porta T (2012) Multidimensional sensor data analysis in cyber-physical system: an atypical cube approach. Int J Distrib Sens Netw 2012:1–19
90. Yu X, Tang LA, Han J (2009) Filtering and refinement: a two-stage approach for efficient and effective anomaly detection. In: Ninth IEEE international conference on Data Mining, 2009. ICDM'09. IEEE, pp 617–626
91. Tang LA, Yu X, Kim S, Han J, Hung CC, Peng WC (2010) Tru-alarm: trustworthiness analysis of sensor networks in cyber-physical systems. In: IEEE 10th international conference on Data Mining (ICDM), 2010. IEEE, pp 1079–1084
92. Xia F, Kong X, Xu Z (2011) Cyber-physical control over wireless sensor and actuator networks with packet loss. In: Wireless networking based control. Springer, New York, pp 85–102
93. Thouin F, Thommes R, Coates MJ (2006) Optimal actuation strategies for sensor/actuator networks. In: 3rd annual international conference on mobile and ubiquitous systems: networking & services, 2006. IEEE, pp 1–8
94. Conti M, Das SK, Bisdikian C, Kumar M, Ni LM, Passarella A, Roussos G, Tröster G, Tsudik G, Zambonelli F (2012) Looking ahead in pervasive computing: challenges and opportunities in the era of cyber–physical convergence. Pervasive Mob Comput 8(1):2–21
95. Guturu P, Bhargava B (2011) Cyber-physical systems: a confluence of cutting edge technological streams. International conference on advances in computing and communication