Received February 10, 2018, accepted March 21, 2018, date of current version May 2, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2821445

Big Data Challenges and Data Aggregation Strategies in Wireless Sensor Networks

SABRINA BOUBICHE[1], DJALLEL EDDINE BOUBICHE[1], AZEDDINE BILAMI[1], AND HOMERO TORAL-CRUZ[2]

[1]Lastic Laboratory, Department of Computer Science, University of Batna 2, Batna 05000, Algeria
[2]Department of Sciences and Engineering, University of Quintana Roo, Chetumal 77019, Mexico

Corresponding author: Djallel Eddine Boubiche (de.boubiche@univ-batna2.dz)

ABSTRACT The emergence of new data handling technologies and analytics enabled the organization of big data in processes as an innovative aspect in wireless sensor networks (WSNs). Big data paradigm, combined with WSN technology, involves new challenges that are necessary to resolve in parallel. Data aggregation is a rapidly emerging research area. It represents one of the processing challenges of big sensor networks. This paper introduces the big data paradigm, its main dimensions that represent one of the most challenging concepts, and its principle analytic tools which are more and more introduced in the WSNs technology. The paper also presents the big data challenges that must be overcome to efficiently manipulate the voluminous data, and proposes a new classification of these challenges based on the necessities and the challenges of WSNs. As the big data aggregation challenge represents the center of our interest, this paper surveys its proposed strategies in WSNs.

INDEX TERMS Big data, data aggregation, wireless sensor networks.

I. INTRODUCTION

The technological advancement in several research areas, including wireless communications, led the researchers to focus on the wireless sensor networks field [1], [2] which represents an innovating technology occupying a crucial place in the data processing, combining wireless communication, detection functions and embedded technology. This emerging technology is gaining ground and is becoming increasingly ubiquitous in all the aspects of the environmental monitoring and processing. The main feature of these systems is the possibility of their deployment in remote and hostile locations, providing users with flexible organization options and facilitating the access to data.

A wireless sensor network (WSN) is formed by collections of sensor nodes widely deployed in generally inaccessible areas and forming data propagation networks. Their main role is to observe a process, collect data and transmit them to a base station for handling. The sensors have introduced the idea of sensor networks based on the collaborative effort between large sets of sensors.

Sensor networks have developed rapidly in recent years and their deployment represents an advantage for new applications. The large utilization of WSN applications and the diversity of the involved fields contributed to increase the volume of data collected and processed. Indeed, when the WSN networks grow and gain in volume and the deployment space, the data collected and processed grow exponentially requiring thus efficient processing, and making consequently traditional data processing methods difficult to use.

Big data technology [3], [4] can represent an effective solution for collecting, analyzing, storing and transmitting data in voluminous wireless sensor networks. Indeed, since the applications of WSN are increasing massively, the sensors deployed are responsible of producing the data in large volume making wireless sensor networks key contributors of big data.

The big data paradigm in wireless sensor networks [5] is young and emerging. It was initially adapted to wired networks, but is gaining momentum in wireless sensor networks, increasing thus the need for new technologies and architectures to handle data. The term big data is used to characterize large data sets that can be complex and therefore difficult to manage by conventional data processing methods. In a wireless sensor network, these huge masses of data are generated every minute and must be collected by the sensor nodes before being transmitted to the base station.

The explosion of big data in WSNs is becoming challenging to handle according to the wireless sensor networks and the big data requirements. Indeed, the big data challenges are combined with the wireless sensor networks to form a new classification based on the requirements of the two technologies.

The goal of this paper is to introduce the big data paradigm in wireless sensor networks, address its main concepts and analytic tools and survey the proposed strategies for their integration in wireless sensor networks. Also, this paper bases on the data requirements in wireless sensor networks to propose a new classification of big data challenges. The data aggregation is one of the main challenges of big sensor data. Thus, its proposed strategies will be addressed in this paper.

The rest of the paper is organized as follows: Section 2 presents the big data concept, its dimensions and analytics tools integrated into wireless sensor networks. Section 3 addresses the challenges of big data. In this section, we also propose a new classification of big data based on the requirements of wireless sensor networks with respect to big data. In section 4, we survey, in details, the different strategies proposed for aggregating big data in wireless sensor networks. Finally, we conclude the paper in section 5.

II. BIG DATA CONCEPT, DIMENSIONS AND ANALYTIC TOOLS IN WIRELESS SENSOR NETWORKS

Big data [3], [4] is a novel technology that represents large sets of data that can be complex and difficult to handle using traditional data processing tools. Compared to traditional datasets, big data defines masses of unstructured data requiring more real-time handling. The big data paradigm can also be defined as the association between large and voluminous data collection and dedicated algorithms allowing exploitations that may largely exceed the classical application of data analysis processes and methodologies.

Initially, the big data model was commonly described using the "3V" framework [6] (volume, velocity and veracity). Recently, the rule of "5Vs" dimensions is used for big data. They represent the vital key elements regarding the characteristics of big data systems. Works also integrated other Vs (6Vs, 7Vs and 9Vs) [7] as key aspects of big data. Fig1 shows the main big data dimensions:

- Volume: The large amount of data requiring storage, processing and organization.
- Velocity: Corresponds to the data generation, processing and transmission speed.
- Variety: Describes the different types of data collected from a variety of sources, processed and stored in different formats.
- Veracity: Concerns the noise problem, the different anomalies in the large amount of data and the degree of significance of the stored data compared to the analyzed problem.
- Value: Describes the quality of the huge amount of data and the explicit or implicit relationships between data.

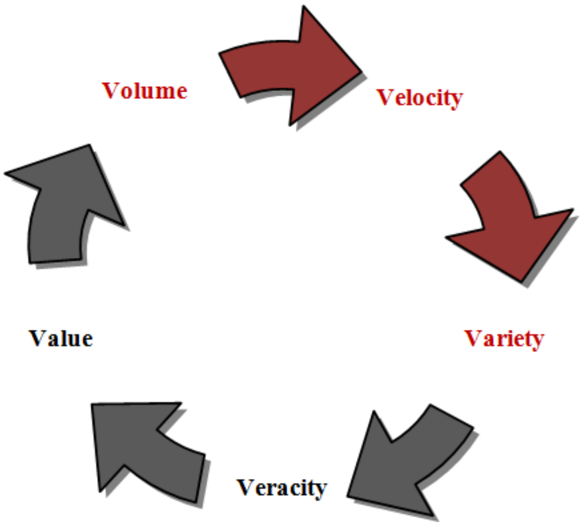

FIGURE 1. Big Data 5Vs dimensions.

The big data paradigm is based on analytical techniques which are mainly based on MapReduce and Hadoop concepts.

Hadoop [8] is an emerging open source data processing framework extensively used in the data exhaustive applications like big data Analysis. It offers flexible and fault tolerant parallel and distributed processing environment. Hadoop uses simple programming models for big data distributed handling. It is based on four main processing modules: the Hadoop Common composed of a set of utilities and serialization libraries supporting the Hadoop modules, the Hadoop Distributed File System (HDFS) which is the Hadoop storage layer that stores large volumes of unstructured data, the Hadoop YARN (Yet Another Resource Negotiator) responsible of the management of the cluster resources, the planning of the tasks and the monitoring of the processing operations of individual cluster nodes, and the Hadoop MapReduce, which implements the MapReduce algorithm.

MapReduce [9] is a data framework which represents the most common framework for big data analytics. It offers quantifiability, fault-tolerance, easy programming, and adaptability. MapReduce uses a data processing algorithm [10] composed of two directives: Map and Reduce principally used to process voluminous data in distributed environments by iterating on a set of input data, creating key/value pairs per register, grouping and saving the results, and reducing each group.

The Map directive creates several small chunks of data from the input data stored in the HDFS, and received line by line. Data are then transmitted to the Reduce directive to produce a new set of output stored in the HDFS. The Map and Reduce directives are sent by Hadoop to the appropriate servers in the cluster. The system verifies the data task completion, and copies data around the cluster between the nodes. Then, data are collected, aggregated and sent back to the Hadoop server.

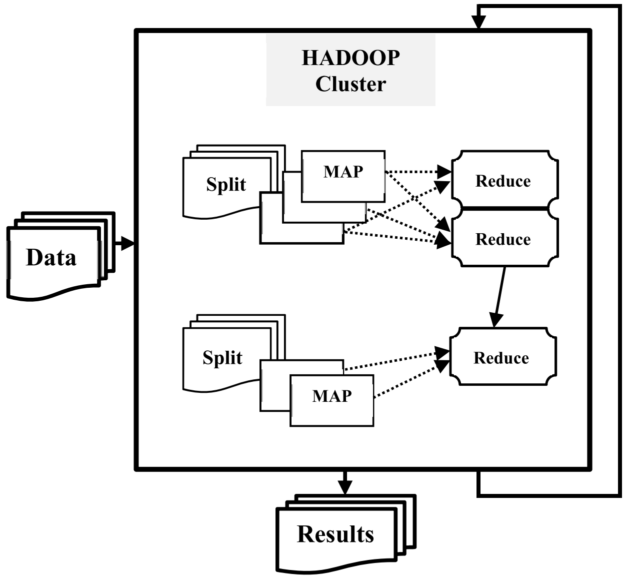

FIGURE 2. Hadoop and MapReduce architecture.

Fig2 shows the Hadoop and MapReduce architecture. Some works have been proposed, where big data analytics tools are integrated into wireless sensor networks [11]:

Farrah *et al.* [12] aim to analyze tool for data collected in wireless sensor networks. For this, they proposed a data warehouse protocol based on Hadoop virtual cluster and a Hadoop data warehousing framework, namely Hive [13] based on queries written on a SQL-like language called Hive Query Language (HiveQL), and converted to MapReduce jobs.

Garcia Rio and InceraDiguez [14] integrated big data tools in sensors pollution monitoring data collecting, storage and analytics. For this, they proposed a model based on two modules: a data acquisition module (DAM) for data gathering, pre-processing and transmission, and a data processing module (DPM) for real time detections based on stream processing and Hadoop and MapReduce algorithms for the related analytics.

In [15], in order to process voluminous data while saving energy in a distributed sensor network, the authors proposed an aggregation technique based on Hadoop framework with single/multi clustered architectures. The authors used an independent and energy efficient, light-weight database oriented data aggregation system, namely PLANetary [16] to find optimal routes through the sensor network.

III. BIG DATA CHALLENGES IN WIRELESS SENSOR NETWORKS

The notable growth and emergence of different network technologies and the explosion in their utilization led to an impressive augmentation in big data generation and handling. Thus, and due to this increasing and the huge volume of big data, the development of big data application meets obstacles and challenges that must be overcome to efficiently manipulate the impressive volume of data deployed.

Various works have proposed classifications of big data [3], [6], [17]. In their proposed classifications, authors mainly focus on big data dimensions, management and handling. As our point of interest is for big data in wireless sensor networks, we assume that the classical big data challenges can be associated with the wireless sensor networks to cover the requirements of both of these technologies.

In this paper, we propose a new classification of big data challenges in wireless sensor networks as shown in Fig 3.

Our classification of big data challenges is based on four key axes that represent the main pillars considered in the wireless sensor networks:

Network clustering is the first pillar in WSNs technology. It represents the main step in the hierarchy of classification that is intensely related to all the other steps. Clustering determines the organization and the deployment of the nodes in the network, their positioning to other nodes and to the base station (BS). It also determines the paths and the order in which the data will be transmitted, the way they are transmitted, and the strategies used in their transmission. Clustering also defines the communication strategies between the nodes of the network.

Big data processing in wireless sensor networks is a critical challenge that needs efficient strategies in order to collect, analyze, store and aggregate the large volumes of data. Big data gathering is a challenging processing task. Indeed, even if the data received by each node in the network appear insignificant, the data generated by the entire network generate an important ration of big data. Thus, the large data volume gathering becomes critical, which requires the use of adapted techniques to deal with this challenge.

The data collected by the sensors require analysis and storage. Analysis methods are needed to handle the increasing volumes of data simultaneously. They also need to be improved, to reduce the response time and save more energy to extend the network's lifetime.

Data aggregation in big data wireless sensor networks is an important paradigm, which represents an efficient solution to deal with the voluminous data by combining the similar data, eliminating consequently the redundant data problem and reducing thereby the resource consumption.

The data clustered and processed need to be secured. Security plays a fundamental role in big data wireless sensor networks, and represents one of their main challenges. Different security mechanisms adapted to big data in WSNs can be used to protect the data at all the levels of the network. However, the security challenge in big sensor data has not been addressed despite its critical importance and strategies focus only on the other challenges.

Clustering is strongly interrelated to the other challenges in the classification. Indeed, the organization of the networks is the basis on which the data processing strategies are defined. The Nodes clustering can represent an effective way to tackle with the energy consumption challenge by grouping the nodes efficiently so that the communication distances and the amount of transmitted messages can be reduced,

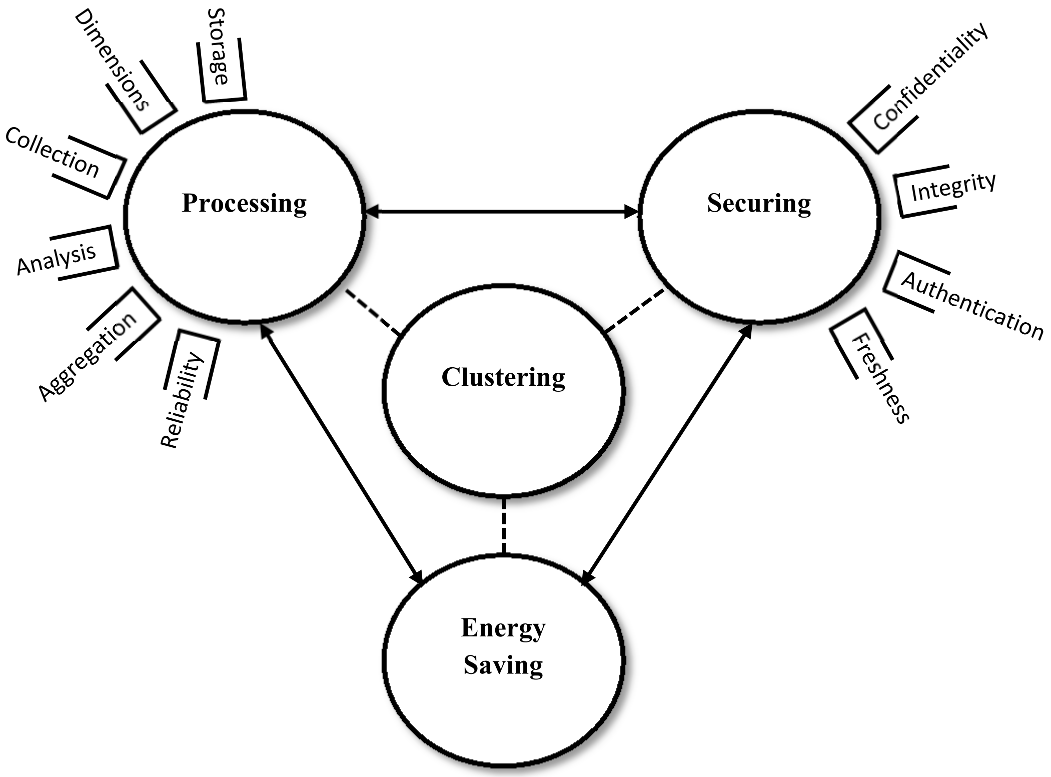

FIGURE 3. Proposed classification for big data challenges in wireless sensor networks.

reducing consequently their energy consumption and the energy consumption of the entire network. In addition, and for optimal data processing, energy efficient mechanisms should be deployed. Also, and due to the voluminous data and the clustering and processing constraints, it is essential to deploy mechanisms to secure the network and protect the big data from different attacks, during the network deployment and the data processing, while saving energy.

In the following, we present the different proposed strategies based on the big data challenges in wireless sensor networks:

A. DATA ENERGY EFFICIENT CLUSTERING

AbdulRazak *et al.* [18] and Kunal and Manasa [19] propose clustering approaches using mobile sink based Energy-efficient big data gathering as an effective solution in densely distributed sensor networks. Also, and to reduce the energy consumption, the authors propose optimal derivations of the clusters.

The proposed clustering algorithms are mainly based on modified Expectation-Maximization (EM) technique that represents a classical clustering algorithm which uses the Gaussian mixture distribution of the nodes in the network. Then, optimal numbers of clusters are calculated based on an objective function defined as the sum of the required energy of data and the data request message transmissions.

In [20] a Distributed K-means Clustering Algorithm adapted to large wireless sensor networks is proposed. The authors based their proposed work on distributed clustering performed at each sensor that collaborates with its neighboring sensors to reduce the communication overhead among the sensors. The proposed algorithm uses an attribute-weight-entropy regularization technique to achieve the ideal assignment of attribute weights to exact the essential features.

B. DATA GATHERING

Doreswamy and Kunal [21], in order to address the continuous big data gathering and the energy efficiency of densely distributed WSNs, proposed an effective method based on the mobility of sink nodes for the selection of clusters. For this, a Dynamically Growing Cellular Self Organizing Map (DGC-SOM) data clustering algorithm based on SOM [22]. [23] is generated to select the cluster head. The proposed algorithm DGC-SOM allows the introducing of new nodes based on a threshold of energy loss. DGC-SOM gives better performance in terms of throughput, packet ratio, packet delivery, transmission and Quality of Service (QoS).

In [24] and [25] authors, and to reduce the delay generated by the mobile sink in the wireless network, proposed data gathering solutions based on M-mobile collectors which cross fixed length paths, reducing thus the complexity generated by their mobility. The data gathering methods are preceded by an improved expectation maximization clustering technique. Then, and to minimize the energy consumption of the network, optimal number of clusters are calculated using an objective function based on the sum of the energy consumption in one cycle of M- Collector patrol.

Wu *et al.* [26] aim to maintain the low structural distortion of the collected data, and reduce the number of active nodes in the network. For this, they proposed a Structure Fidelity Data Collection (SFDC) framework having as main objective to maintain the data fidelity in term of structural similarity. The proposed framework is based on a structural distortion approach that uses image fidelity metric to assess the quality of the distorted image.

Arivoli and Chitra [27] proposed a big data gathering protocol named Hybrid Dynamic Energy Routing Protocol (HDERP). The proposed protocol is characterized by its reliability, which ensures a hop-by-hop data transmission path, a higher life period and effective energy consumption, reducing thus the end-to- end delay.

C. DATA ANALYSIS

Ang *et al.* [28] used analytic approaches to calculate the energy consumption of the nodes and the optimal number of clusters for two mobile data collection models: MULE (multihop model) and SENMA (single hop model) experimented with different very large-scale network scenarios in a short lapse of time. Also, and to minimize the energy consumption in large scale WSNs, authors proposed multi cluster models for determining the optimal number of clusters.

Saneja and Rani [29] aim to address the scalability and the correlation limitations of big data in wireless sensor networks for the detection of faulty sensors. For this, the authors proposed an outlier scalable to big data detection approach based on correlation and dynamic SMO (Sequential Minimal Optimization) regression. Based on Hadoop MapReduce framework, the proposed approach aims to find out the strongly correlated attributes and to efficiently detect the anomalous nodes, reducing then the number of false alarms.

Ejaz *et al.* [30] surveyed the recent proposed frameworks related to big data analytics for IoT (Internet of Things). The works principally aim to overcome the challenges of analyzing large amount of data. The authors also explored the big IoT-generated data processing and analytics platforms, and studied the IoT big data and analytics requirements. Based on important parameters, the authors taxonomized the IoT big data and analytics solutions.

D. ENERGY EFFICIENCY

Liu *et al.* [31], and in order to resolve the multi-agent itinerary problem, proposed an energy efficient protocol for spanning tree node construction as a basis of a routing itinerary planning scheme. Their proposed solution is firstly based on a multi-agent-based distributed WSN (DWSN) model and an energy consumption model. Then, the routing itinerary algorithm named DMAIP (Dfocus-Modulation-type Active Image Processing) is built. The routing algorithm is also extended to reduce long-distance transmission.

Singh and Kumar [32] proposed a new energy efficient big data aggregation protocol based on Real-time Big Data Gathering algorithm (RTBDG) to recursively split the sensor network into various parts symmetrically about a root node. The authors also based their work on HEED (Hybrid Energy-Efficient Distributed clustering) distributive clustering algorithm which considers a hybrid of energy and communication cost when selecting CHs [33]. Kaur and Rjput [34] also based their work on RTBDG.

E. DATA STORAGE

Xu *et al.* [35] developed large data storage and retrieval algorithm for wireless sensor networks, based on non-uniform distribution of the nodes, which uses a simple routing protocol. The objective of the algorithm is to estimate the real distribution and the addresses of the sensor nodes while considering the redundancy of the data among the neighboring nodes.

IV. BIG DATA AGGREGATION PROPOSED STRATEGIES IN WIRELESS SENSOR NETWORKS

Data aggregation is one of the key challenges in big data wireless sensor networks. Data aggregation allows combining data from different sources to eliminate the redundancy, and reduce consequently the consumption of resources available in the network. Data aggregation is a subset of data fusion that involves the use of techniques that combine and gather data from multiple sources to make more effective and potentially more accurate inferences, reparations and associations.

Strategies are proposed to deal with this challenge. They are mainly based on the correlation between the data aggregation, the clustering and the energy consumption challenges of big sensor data. In the following, we survey the big data aggregation strategies proposed for wireless sensor networks:

A. DISTRIBUTED COMPRESSIVE DATA AGGREGATION IN LARGE-SCALE WIRELESS SENSOR NETWORKS

The authors of the paper [36] proposed a distributed algorithm based on local minimization to dynamically build a routing path to reduce data traffic for aggregation based on compression sampling.

The primary goal is to minimize the general traffic in the hybrid data aggregation process with low overhead costs. The authors assume that the aggregation of the data is performed in cycles by correctly programming the network and that no transmission error occurs during the application of the CS source coding scheme. Therefore, the routing path in the data collection process will form an aggregation tree rooted at the BS. In addition, a node is selected as an aggregator when the size of the data it links is greater than a given M.

The proposed compressive data aggregation algorithm is divided into two steps:

- Construct the data aggregation tree in a distributive way,
- Reduce the data traffic by adjusting the routing path locally.

  The data aggregation tree construction is divided into two phases:
- The first phase of the construction consists in calculating the shortest distance between all the nodes and the BS. For this, a set of unvisited nodes is created and in which all the nodes except the BS are marked as unvisited. Then, for each given node and once all of its neighbors are considered, it is marked as visited, removed from all unvisited nodes and its tentative distance is recorded as the shortest distance. The process is repeated until the set of the unvisited nodes is empty or the smallest tentative distance between all the nodes in the unvisited set is equal to the infinity.

The second phase of the construction consists of finding the edges of the tree of the shortest path:

- After the shortest path distance is found for all the nodes, a parent P_x is assigned to each vertex x different from the BS. Once the parent of all the different nodes of the BS is determined, shortest path tree is composed of the edges between all the nodes and their parents.

Once the aggregation tree is built, a distributed local minimization algorithm (LM) is used to calculate whether switching to a different neighbor can reduce data traffic. each node performs the following steps:

- Each node x collects the size of the received data and the identity of the parents of its neighbors with two hops by exchanging INFO messages with its neighbors.
- For each non-child neighbor with a distance equal to or less than the BS, x measures the local data traffic change if the node x changes its parent to Ni. The new local traffic is calculated assuming node x is redirected to Ni. In the case where Ni and x have the same parent, parental traffic is counted only once. If the new traffic is less than the original traffic, x records the reduced traffic size and the ID of the neighbor.
- After all the neighbor measurements are complete, if the node x is not locked, it selects the neighbor and minimizes the local traffic and sends a LOCK message to prevent it from updating its parent. Otherwise, node x defers its action until it is unlocked by an UPDATE message to continue. Once the LOCK message is recognized by y (i.e., y is successfully locked), x updates its parent to y, and then broadcasts an UPDATE message to inform its neighbors so that they can execute this algorithm with the updated information. The UPDATE message also unlocks y so that it can continue its own calculation and update. If x does not receive an acknowledgment from y after a period of time, it waits for a random time delay and sends a LOCK message again.

The proposed approach is experienced and the simulation results demonstrate that the tree structure has a significant impact on the efficiency of compressive data aggregation. Also, the results show that the proposed solution generates much lower overhead cost than the near optimal solution, making it more suitable for WSN in practical applications.

B. SENSOR DATA AGGREGATION IN A MULTI-LAYER BIG DATA FRAMEWORK

The work proposed in [37] aims to aggregate data in WSNs based on big data. For this purpose, the authors propose a multi-layer big data aggregation infrastructure as well as a priority-based Dynamic Data Aggregation Protocol (PDDA) implemented on the sensor nodes responsible for the data collection.

The authors proposed a three-layer data aggregation infrastructure, where data aggregations are performed in Internet-connected base stations (BSs) and large data servers. Next, the paper presents a Dynamic Data Aggregation Scheme (PDDA) based on priorities for sensor networks because the sensors collect a large amount of redundant data. The proposed PDDA scheme is a hybrid approach that uses clustered and tree-based approaches based on application types. The cluster approach is used to aggregate real-time emergency data that reduces the end-to-end data transmission delay since these data have the highest priority and must be transmitted with a minimum data transmission delay. The tree approach is used for non-real-time applications. Cluster-based and tree-based topologies select certain nodes as active ones that provide all network coverage. Thus, the proposed PDDA approach achieves energy efficiency and reduces data processing time and overhead at the big data server level.

The proposed data aggregation infrastructure has three layers:

(i) Data aggregation at the level of the sensors - layer 1

(ii) Data aggregation at the base station (BS) - layer 2

(iii) Data aggregation at the big data Server or NoSQL Server - Layer 3 server.

The proposed PDDA system provides data aggregation priority based on the type of data captured. For example, critical applications in real time, in case of emergency, will have more priority than non-real-time applications.

Sensors of layer 1 transmit data to the upper layers through the base station or the gateway nodes. However, to achieve efficient aggregation of data, the sensor networks used for the different types of applications are designed to have a different network topology. For example, for critical or real-time applications, clustering-based aggregation is used when sensors transmit data to the base station through their cluster head (CH). If the CH is far from the base station, the CH will consume more power, but it will eventually transmit data through a minimum number of hops and, as a result, should reduce data latency. In the proposed approach, a number of nodes are selected as active nodes that cover the entire network area [38]. Then, the clusters are formed and an active node is selected as CH for each cluster. The active nodes of the cluster members detect and transmit data to the CHs while the CHs filter or reject the redundant critical data and transmit them to the gateway node so that it can transmit them to the central database or to the control station with a minimum delay [39].

On the other hand, for non-real-time applications, achieving energy efficiency is more important. As a result, the sensors form a tree topology and transmit data via the shortest path to the gateway node or the BS. Initially, the nodes will be identified as located at different levels of the network depending on the number of hops for the gateway node. Then, the shortest path of the gateway node to the active nodes will be created using the method presented in [40]. Active nodes at the lowest level will detect the event of interest and transmit to the active nodes at the higher level. Parent nodes in this tree always perform data aggregation using different aggregation functions such as MAX, MIN, MEAN, MEDIAN, SUM, and resend to active nodes at the higher level until the data reach the gateway. This approach should result in well-distributed power dissipation on all active nodes and also a lower power consumption of the network, even if the number of hops from the sensor node to the BS is higher compared to the counterpart based on the clustering of this proposed approach. This is due to the energy consumption in a sensor node that is directly proportional to the square of the distance that a packet of data travels from one node to another [41]. However, this approach may have a longer data transmission delay because it traverses several levels and spends time at each node of this hierarchy for data processing. Thus, the proposed PDDA approach offers a compromise between energy efficiency and data transmission delay.

Normally, sensor nodes are deployed for a specific application and form their network topology based on the type of the application. However, these sensors can be reused in other applications and change their topology if the application changes. Sensors are aware of the application change by checking the data packet they are sensing and transmitting because the data packets contain the application types, which helps layers 2 and 3 to process and store data in the appropriate places.

The simulation shows that the proposed PDDA scheme dissipates less energy compared to traditional cluster and cluster data aggregation approaches. As a result, the network lifetime of the proposed scheme should be longer than that of the cluster and tree approaches. The results demonstrate that PDDA approach data transmission is inferior to that of clustering-based tree and clustering approaches.

The proposed PDDA data aggregation approach selects only a few active nodes that cover the entire network, which reduces the total power consumption of the network. Involving fewer active nodes in processing and data transmission also reduces end-to-end data transmission time.

C. LIFTING WAVELET COMPRESSION BASED DATA AGGREGATION IN BIG DATA WIRELESS SENSOR NETWORKS

In the work proposed in [42], the authors aim to the energy-efficient elimination and the compression of redundant data with the objective of recovering the original data. To balance the aggregation load of a large-scale WSN, the authors propose a new energy-efficient dynamic clustering algorithm using spatial correlation, which provides a compressive and distributed data aggregation in each cluster. The authors used a fast and distributed data compression approach based on a lifting wavelet to reduce the amount of raw data. In addition, the approach offers a high recovery of raw data.

The authors propose a data compression algorithm based on a distributed high-speed wavelet to compress the data captured for large-scale WSNs, which can effectively reduce the amount of transmitted data and recover the original data with great accuracy. The originality of the proposed approach lies in the following points:

(i) The network can be dynamically grouped by exploiting spatial correlation and user requirements rather than complete aggregation in the network.

(ii) Spatial and temporal redundancy can be reduced by the proposed data compression algorithm.

(iii) The proposed approach achieves a good balance between the accuracy of data retrieval and the energy consumption. Spatial correlation is used to determine the cluster members. Once the data detected in a cluster have a strong correlation, a cluster member can represent the nodes in its neighboring area with its own data and the cluster head compresses and sends only the data of its members to the base station rather than all the received data. In addition, the approach allows some cluster members to exit a cluster if their abnormal data are detected. Thus, the energy will be saved by the withdrawals of some nodes and the size of the data will be reduced by deleting the redundancy.

The main contributions of the proposed approach are as follows:

- The authors propose a dynamic clustering algorithm based on the spatial correlation of data to eliminate redundancy, which can reduce the size of the redundant data and thus effectively extend the life of the network.
- The benefit of the dynamic clustering algorithm is analyzed in terms of complexity of time and space. The optimal number of clusters is derived based on related energy consumption.
- The authors rely on the use of a fast, distributed, wavelet compression technique to aggregate data at each cluster head and send the data to the base station. Prior to the transmission, the wavelet coefficients are further compressed and encoded to reduce the amount of coefficients, which can ensure the accuracy of data retrieval by occupying only a small amount of storage space.

1) CLUSTERING MODEL BASED ON SPATIAL CORRELATION (CDSC)

The network is modeled as an undirected graph G = (V, E), where V is the set of sensor nodes and E is the set of edges composed of all the links of the WSN. Advantages nodes in terms of energy and geographic locations are suitable for acting as cluster heads. Cluster heads not only collect data, but also transmit them to the base station, making them the most dynamic nodes in the network. Note that some nodes in a region may represent all the metrics received from their neighboring nodes because of the strong spatial correlations between the data. All the sensor nodes are divided into two categories: candidate members and normal nodes based on their residual energy levels before the clustering. The critical component of the proposed clustering model is the selection of cluster heads and cluster members.

The clustering algorithm proposed for each sub region is divided into several stages:

- All nodes are divided into two groups: CandSet which represents the candidate nodes to become cluster heads, and OrdSet which represents the remaining nodes. The division is done using the following formula:

$$p = \frac{E_{cur}}{E_{total}} \tag{1}$$

Where: p represents the state of the remaining energy, Ecur represents the state of the residual energy of a given node, and Etotal indicates the total energy capacity.

- A node from the Set group will be selected as the cluster head, provided that it is closest to the regional center. In each sub region, the sensor node i in CandSet finds the shortest geographical distance from the central sub-region. Then the node with the shortest remote broadcasts becomes cluster head. Then the remaining CandSet nodes become members of the cluster.
- The last step is to find the normal nodes next to a cluster. For that, for each node, the following formulas (13) and (3) are checked to verify if they satisfy all the constraints:

$$x_{min} \geq x_{actual} \times R_3 \tag{2}$$

$$x_{max} \leq x_{actual} + x_{actual} \times (1 - R_3) \tag{3}$$

Where: x_actual represents the sensory data of an existing member of the cluster at the current time. x_min and x_max represent the minimum and maximum sensory values of the adjacent normal nodes of a cluster member. Rs is the adjustable similarity ration in the BS.

In this case, the node is represented by the corresponding cluster member and does not join this cluster. Otherwise, the node will be added as a new member. The remaining normal nodes will be removed from this cluster.

2) WAVELET DISTRIBUTED LIFTING COMPRESSION

The data are compressed based on the proposed data correlation clustering algorithm (CDCC) which not only provides a fast computation, but also a substantial backup of the memory space.

The data received by the CH are stored as a twodimensional matrix. The matrix can be broken down into four sub-bands by in-line or column lift wavelets. The four sub bands are: a low frequency sub band and three high frequency sub bands. The most useful information is concentrated in the 8 low frequency sub-band. As a result, some frequency coefficients that contain less information can be eliminated.

In the following, the main procedures of the proposed wavelet data compression method:

First, the original data perform a first-level wavelet transformation. The wavelet transformation process is divided into three steps:

Fractionation Process: For each row of the data matrix, the data at a given time interval are divided into an even and an odd sequence.

Prediction Process: A prediction operator is executed on the signals at time intervals to predict the even data signal.

Update Process: Consists of running a new update operator, which updates the original peer signal to the new peer signal.

During the process of transforming into line wavelet, the original data are replaced by a low frequency coefficient and a high frequency coefficient. The higher is the time correlation of the data the lower is the value of the high coefficient. Similarly, in the process of wavelet transformation of columns, greater spatial correlation implies a lower high frequency coefficient. When the first level wavelet transform is complete for all rows and columns, the original data are converted to a low frequency portion and three high frequency portions.

The proposed algorithm is compared to other approaches and the simulation results show that the CDCC algorithm is superior in terms of energy saving. Although the compression of the lifting wavelet is constrained by its compression ratio, its recovery accuracy may be greater than 98% if the parameters are adjusted appropriately. Indeed, experimental results demonstrate that the data correlation clustering (CDSC)-based consolidation method proposed for data aggregation outperforms other methods to extend the network lifetime and reduce the amount of transmitted data. The proposed dynamic clustering algorithm and the wavelet-based compressive data aggregation technique can achieve better performance, for example greater recovery accuracy data and considerable energy savings.

D. DATA AGGREGATION WITH PRINCIPAL COMPONENT ANALYSIS IN BIG DATA WIRELESS SENSOR NETWORKS

In this paper [43], a Principal Component Based Data Aggregation (PCA) algorithm is proposed to efficiently transmit large data detected with low latency while eliminating redundancy of data in cluster head nodes to minimize the complexity of the transmitted data.

The authors of the paper proposed a distributed clustering algorithm based on similarity to place the nodes with a strong similarity in the same cluster.

1) ENERGY SYSTEM MODEL

A large number of sensors in wireless sensor networks will produce large amounts of data. These detected data are collected in big data [14]. When the base station wants to consume the least energy to accommodate all the data, the distribution of big data must be understood first. There are two ways to get the data distribution. The first way is that sensor node processes and retrieves the associated data and then transmits the processed data to the base station. The other way is that when transmitting big data, all the nodes transmit their raw data to the BS. BS processes this data and makes the best choice for the network. The proposed approach adopts the second option.

In the proposed model, the energy consumed to transmit i bytes of data set in each node is given by the following formula:

$$E_T\left(l,d\right)=\begin{cases} l*E_{elec}+l*E_{fs}*d^2, & if\ d<d_0 \\ l*E_{elec}+l*E_{amp}*d^4 \end{cases} \tag{4}$$

Where: d is the distance between the transmitter and the receiver and d0 is the communication radius of the node.

Eelec is the energy consumption of the transmitter and the receiver.

Efs ∗ d2 is the power consumption of the amplifier in the communication range.

Eamp ∗ d4 is the power consumption of the amplifier beyond the range of communication.

The network is divided into clusters based on the similarity between the nodes. The energy consumed by a ch by aggregating 1 byte of data from its members is given by the following formula:

$$E_p=k*l*E_{pr} \tag{5}$$

Where: Ep represents the energy consumption of the data aggregation process in the cluster; k is the number of nodes in the cluster.

The energy consumed throughout the cluster is given by the following formula:

$$E_{Cluster}=E_T\left(l,d\right)+E_R\left(l,d\right)+E_P \tag{6}$$

Where Er (1, d) represents the energy consumption for receiving the data bytes for each node and which is given by the following formula:

$$E_R\left(l,d\right)=l*E_{elec}\left(4\right) \tag{7}$$

2) CLUSTERING ALGORITHM BASED ON DATA SIMILARITY

The authors proposed a clustering algorithm for data aggregation based on the similarity of data. The latter is defined according to two different aspects, namely the magnitude of similarity of the data and the correlation of the data.

Based on the similarity values, a clustering algorithm suitable for data aggregation based on the PCA algorithm is proposed. The latter ensures the partitioning of the sensor nodes in the clusters with a high level of similarity, and makes it possible to select an appropriate node to become CH.

The proposed algorithm is detailed as follows:

- Each sensor node calculates the similarity with its neighboring nodes. If two nodes u and v satisfy a given similarity threshold ξ, they will set up a *uv* arc. Thus, all the nodes will form a graph g. The nodes will then be sorted according to their degree in the set of the nodes *s* and the node with the widest degree will be chosen like CH. To reduce power consumption in each cluster, the number of nodes is reduced to k-1.

The CH node selects k-1 nodes with high similarity from its neighboring nodes and removes them from the set s of the nodes. the procedure is repeated until the widest degree of nodes in the set s is less than k-1.

In the case where the number of remaining nodes in s is small, the similarity calculation procedure is not triggered, but the number of nodes in the cluster is reduced so that the remaining node becomes CH.

- Once the clusters have become stable, the members rotate the CHs.
- If the bs finds that the inter cluster data have a difference greater than a given threshold or if half of the nodes do not satisfy the similarity values, it decides to reactivate the clustering algorithm.

3) DATA AGGREGATION BASED ON PCA (PRINCIPAL COMPONENT ANALYSIS)

The PCA data compression algorithm is a highly recommended algorithm for sensors with limited capabilities. It allows reducing the degree of dimension of the data sets while maintaining the characteristics of the greatest contribution to the variance.

The proposed algorithm is detailed as follows:

- The network is divided into several clusters based on the proposed clustering method.
- The CH collects the data of its members, and puts them in an observation matrix x.
- From the observation matrix x, the covariance matrix c is calculated.
- The eigenvalues and the corresponding eigenvectors of the matrix C are calculated;
- The eigenvalues are ranked to obtain the greatest value;
- The eigenvectors corresponding to the largest eigenvalue are selected to form the transformation matrix P;
- The projection matrix is calculated from the transformation matrix P;
- The calculated projection matrix is sent to the BS.

The algorithm proposed in this approach is tested and the experimental results have demonstrated its effectiveness in reducing the amount of data transmitted and the energy consumption in the network.

4) SCALABLE PRIVACY-PRESERVING BIG DATA AGGREGATION MECHANISM

The authors of the paper [44] proposed a scalable, confidential data aggregation scheme (SCA-PBDA) for big data to meet the requirements for privacy and transmission efficiency.

The proposed scheme is based on a gradient topology structure that allows the sensor nodes to be equally divided into clusters formed of a cluster head CH, an equal number of members (CMS) and auxiliary cluster heads (aCHS). The network clustering allows the privacy-preserving configuration and the inter-cluster data aggregation techniques to perform the inter-cluster data aggregation.

The proposed work is principally based on two points:

- The energy consumption of the node is used to determine the topology of the network based on a gradient-based equal network clustering method. The clustering method allows the identical nodes to support the uniform privacy-preserving configuration and inter cluster data aggregation.
- Privacy preserving data configuration and scalable intra and inter cluster data aggregation are ensured through a scalable privacy-preserving data aggregation method.

5) CLUSTERING METHOD

The proposed clustering method is based on a gradient establishment. For this, a gradient establishing message GE is broadcasted by the sink, which gradient field value is 0, to the nodes through which they can determine their distance from the BS. For this, the nodes and after they obtain their gradient value 1 according to the first received GE, will add the value to the hop count field of GE and update it. Then, the sensors broadcast the updated GE after a given delay. The procedure of broadcasting and updating the GE is repeated until all the sensor nodes have obtained their gradient value

The CH nodes of each cluster are elected from the sensor nodes and the aCHs will be allocated according to the energy consumption of CHs, and which is negatively correlated with their gradient values. The following formula represents the total energy consumption of a chi with a gradient value i:

$$EC_{CHi} = EC_{Receiving} + EC_{Aggregating} + EC_{Transmitting} \quad (8)$$

Where $\mathrm{EC}_{\mathrm{Receiving}}$ represents the energy consumption of data received from the CMS and from other CHS.

$\mathrm{EC}_{\mathrm{Aggregating}}$ is the energy consumption resulting from data aggregation.

$\mathrm{EC}_{\mathrm{Transmitting}}$ represents the energy consumption resulting from the aggregated data transmission.

To prevent the partitioning of the network and the communication interrupts caused by the energy depletion of relay CHs closer to the sink, their energy consumption is shared by adaptively allocating aCH for the clusters. The inter-cluster privacy-preserving data aggregation is established through an identical cluster composed of CH and CMs.

The sink proceeds in the initialization phase by broadcasting the GE message to establish the network gradient. Then, sensor nodes with gradient value i elect themselves as CHs with a calculated probability. Other sensor nodes send request for joining the cluster as CMs or aCHs.

6) INTRA-CLUSTER PRIVACY-PRESERVING DATA AGGREGATION

Before the data aggregation is performed, the privacypreserving positions are determined. Firstly, the sink determines the global true value position set (GTPS). This last is used to tag the true sensor data values. Each node generates a data index set *I* for the sensor data to indicate that privacypreserving sensor data is composed of true sensor data values and the camouflage filling values which guarantee the privacy of the true sensor data values. The more I value is larger, large is the space created to fill more camouflage values of the sensor nodes. However, this will create additional communication overhead.

For each sensor node, the sink attributes Node Private Position Set (NPPS) and Node True Position Set (NTPS). In 10 the proposed scheme, the position of the true sensor data value of each node is tagged using NTPS. In addition, the positions of the nodes are tagged by the sink to place the true sensor data value and the restricted camouflage values. To facilitate the inter-cluster privacy-preserving data aggregation, privacy-preserving camouflage filling method is deployed in every cluster of the network. True sensor data value and restricted and unrestricted camouflage values are placed in the appropriate positions.

The transmission of the protected data to the sink is performed in a multi-hop way, where their transmission between CMs and CHs is performed in single-hop way.

When privacy-preserving data are received by the CH from its CMs, it performs the MAX data aggregation as shown in the following formula:

$$Data_{Aggregated} = \bigcup_{i=1,\ldots,l\ j=1,\ldots,CS} max\left(d_{ij}\right) \quad (9)$$

Where $\mathrm{Data}_{\mathrm{Aggregated}}$ represents the aggregated privacy-preserved data and d_{ij} represents the data value of the node's j ith position.

7) INER-CLUSTER PRIVACY-PRESERVING DATA AGGREGATION

As the cluster composition and the privacy-preserving camouflage filling method are identical, the same data aggregation technique can be used.

In the inter-cluster privacy-preserving aggregation, the aggregated privacy-preserving data are re-aggregated at the relay CHs to obtain global aggregated results containing the maximum sensor data value of all sensors on the link:

$$Data_{reAggregated} = \bigcup_{i=1,\ldots,l\ j=1,\ldots,RP} max\left(d_{ij}\right) \quad (10)$$

Where $\text{Data}_{\text{reAggregated}}$ represents the re-aggregated privacy-preserving data and n_{RP} represents the number of received privacy-preserving aggregation data packets.

It is noted that when the inter-cluster privacy-preserving data aggregation is performed, the CHs perform only the simple re-aggregation on their own and received intra-cluster privacy-preserving data without having knowledge of their detailed content. Thus, big sensor data privacy is guaranteed.

8) RECOVERY OF THE AGGREGATED RESULT

The last data aggregation operation is performed by the sink when receiving all aggregated data packets from its CH1s. Then, the sink scans the data index positions and the maximum values are kept as recovery result:

$$Data_{global} = \bigcup_{i=1,\ldots,l\ j=1,\ldots,N_{cluster}} (d_{ij}) \tag{11}$$

Where $\text{Data}_{\text{Global}}$ represents the global aggregated privacy-preserving sensor data packet, and $\text{n}_{\text{Cluster}}$ represents the total cluster number of $\text{CH}_{1\text{S}}$.

To verify the performance of the proposed scheme in terms of privacy-preserving, simulation is performed and the scheme is compared to traditional privacy-preserving data aggregation mechanisms such as CPDA (Conflict-free Periodic Data Aggregation) and KIPDA (K-indistinguishable Privacy Preserving Data Aggregation). Also, the network lifetime of the proposed protocol is simulated to validate its effectiveness.

The simulation results show that the computational complexity and the computation overhead of Sca-PBDA are extremely less than cryptographic CPDA and KIPDA. Thus, the proposed Sca-PBDA meets the application requirements of the computational complexity and scalability.

E. A CLUSTER-BASED DATA FUSION TECHNIQUE TO ANALYZE BIG DATA IN WIRELESS MULTI-SENSOR SYSTEM

Din *et al.* [45] proposed a new data fusion technique based on a hybrid algorithm for clustering and cluster member selection in wireless multi-sensor system. Also, the authors use a data fusion technique for partitioning and processing the collected data.

The authors based their work essentially on the following points:

- Develop a hierarchical framework to integrate the clustering algorithm with the routing technique.
- Propose an optimized routing technique to solve the void problem and facilitate the deployed nodes to achieve survivable network.

1) THE NETWORK ARCHITECTURE

The network architecture is based on four methods: node deployment, cluster formation, member node Selection criteria, and routing technique.

In the proposed architecture, the sensor nodes are responsible of sensing data, and forwarding them to the cluster heads. In the other hand, the cluster heads perform the computations and the communication. Then, the cluster heads forward the processed data to the base station using the routing technique. Another role of the cluster heads is to flood the control signals, and to find the energy efficient path to the base station.

In the organization of the network, the nodes can be deployed randomly or uniformly and start broadcasting control packet to their neighboring nodes to show their existence. When the neighboring nodes receive the broadcast message, they set up their neighboring table to mainly calculate the link quality for future decisions.

Nodes are deployed hierarchically and the cluster heads forward received data toward their neighboring cluster heads near to the base station. A rotation of cluster heads is performed after a speci?c interval to save the energy of the nodes.

The system model is evaluated through the power transmission calculation:

$$Y_{ij} = \frac{W_i}{N_{\|i,j\|^{\propto}}} \tag{12}$$

Where: W Represents The Power Of A Node i, NO represents the white noise, $\|i,j\|^{\propto}$ is the Euclidean distance between nodes i nodes j, and α represents the path loss [46].

Also, the data rate between the source and destination is calculated using the following equation:

$$C_D\left(s_i, d_i\right) = BLog_2\left(1 + Y_{s_i,d_i}\right) \tag{13}$$

Where, B represents the available bandwidth. also, in the proposed scenario authors used the technique of decode-and-forward [47].

The energy consumption is calculated as the overall sum of hops from all the CH.:

$$W = \sum_{n=1}^{k} W_n = \sum_{n=1}^{k} E\left(D_R^c + S_D\right) \tag{14}$$

Where E is the amount of energy consumed by a node, D_R^c represents the relationship between all the nodes and the overall summation of hops in a network, S_D is the total amount of data being collected by a node in one cycle, and wn is the energy consumption of node i per cycle.

2) THE CLUSTERING LAYER

The proposed technique aims to minimize the energy consumption during the cluster formation, increasing thereby the network life. In this technique, the neighbor node favors one node to another for the selection of different multi-step positions, where the node used to select the cluster head among the consequent layers (the decision node), the packet ID and the postfix counter represent the main factors in the design of the cluster. In addition, the nodes of the second layer select the nodes surrounded by the maximum number of nodes as decision nodes.

The nodes of the second layer interchange the information of their density to take part in the computation of becoming

TABLE 1. Big data challenges involved in data aggregation.

Challenges / Data aggregation strategies	Clustering	Processing		Securing	Energy saving
		Gathering	Analysis		
Distributed Compressive Data Aggregation in Large-Scale Wireless Sensor Networks	Yes	No	No	No	No
Sensor Data Aggregation in a Multi-layer Big Data Framework	Yes	No	No	No	Yes
Lifting Wavelet Compression Based Data Aggregation in Big Data Wireless Sensor Networks	Yes	No	No	No	Yes
Data Aggregation with Principal Component Analysis in Big Data Wireless Sensor Networks	Yes	No	Yes	No	Yes
Scalable privacy-preserving big data aggregation mechanism	Yes	No	No	Yes	No
A Cluster-Based Data Fusion Technique to Analyze Big Data in Wireless Multi-Sensor System	Yes	No	No	No	Yes

cluster head. They turn on their transceiver through the time division multiple access (TDMA) technique.

To select the member nodes in each cluster, the cluster head initially broadcasts a message of Join Request which indicates the availability of the cluster head. Then, the node that receives the Join Request message broadcasts in return a Join Accepts message to become a member of the cluster. If more than one Join Request is received by the nodes, the decision is made based on the load on the cluster head.

3) DATA FUSION MODEL

The data fusion model proposed in [48] is modified based on the architectural requirements. The proposed data fusion technique is composed of five levels.

Level 0: Initially, the data is received by cluster head. Then, data are aligned and divided into sub-blocks where each one is processed separately at each cluster head.

Level 1: Data are refined at each cluster head and various types of data are converted into the consistent structure (images, texts …).

Level 2: Contextual description is provided for each subblock based on environmental data.

Level 3: Each cluster head identifies future threats and susceptibilities for operations based on the computational complexities and the designed algorithm.

Level 4: The processing performance is continuously monitored at each cluster head.

The life cycle of the data fusion model in the multi-sensor system is defined. It consists of four interdependent attributes:

- Raw data are collected and organized in a meaningful form.
- The undesirable data are discarded.
- The data blocks are fused and analyzed.
- Finally, the processed data are transmitted over the network.

The proposed scheme is implemented using Java iterations and the Hadoop. The simulation results show the proposed technique is energy efficient in all the proposed scenarios.

In our proposed classification for big data challenges in wireless sensor networks, we assume that all the challenges are interrelated and complement each other.

As the last section of our paper is dedicated to big data aggregation and fusion mechanisms in wireless sensor networks, we present in table1 the main challenges considered in addition to big data aggregation challenge.

V. CONCLUSION

Big sensor data continue to increase every day. Their variety, volume and velocity are also expanding. The big data paradigm in wireless sensor networks requires energy efficient clustering, processing, and securing. These requirements represent the main big data challenges in wireless sensor networks. The data aggregation is one of the principle big sensor data processing challenges. In this paper, we introduced big data in wireless sensor networks. We presented a view of big data concepts and analytic tools and survived the works proposed for integrating them in wireless sensor networks. We also proposed a classification for big sensor data challenges and reviewed the proposed solutions for these challenges. As big sensor data aggregation represents our principle point of interest, we survived its proposed strategies in detail. In the future, we aim to propose novel strategies for big data challenges and issues in heterogeneous wireless sensor networks.

REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey," *Comput. Netw.*, vol. 38, no. 4, pp. 393–422, 2002.

[2] J. Hill, R. Szewczyk, A, Woo, S. Hollar, D. Culler, and K. Pister, "System architecture directions for networked sensors," *ACM SIGOPS Oper. Syst. Rev.*, vol. 34, no. 5, pp. 93–104, Nov. 2000.

[3] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Netw. Appl.*, vol. 19, no. 2, pp. 171–209, 2014.

[4] K. U. Jaseena and M. Julie David, "Issues, challenges, and solutions: Big data mining," *Comput. Sci. Inf. Technol.*, vol. 4, pp. 131–140, Dec. 2014.

[5] H. Harb, A. Makhoul, A. K. Idrees, O. Zahwe, and M. A. Taam, "Wireless sensor networks: A big data source in Internet of Things," *Int. J. Sensors, Wireless Commun. Control*, vol. 7, no. 2, pp. 93–109, 2017.

[6] A. A. Tole, "Big data challenges," *Database Syst. J.*, vol. 4, no. 3, pp. 31–40, 2013.

[7] A. K. Bhadani and D. Jothimani, "Big data: Challenges, opportunities, and realities," in *Effective Big Data Management and Opportunities for Implementation*, M. K. Singh and G. D. Kumar, Eds. Hershey, PA, USA: IGI Global, 2016, pp. 1–24.

[8] *Welcome to Apache Hadoop!* Accessed: Dec. 29, 2014. [Online]. Available: http://hadoop.apache.org/

[9] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[10] R. Lämmel, "Google's MapReduce programming model—Revisited," *Sci. Comput. Program.*, vol. 70, no. 1, pp. 1–30 2008.

[11] S. Boubiche, D. E. Boubiche, and B. Azzedine, "Integrating big data paradigm in WSNs," in *Proc. Int. Conf. Big Data Adv. Wireless Technol. (BDAW)*, Nov. 2016, p. 56.

[12] S. Farrah, H. El Manssouri, E. Ziyati, and M. Ouzzif, "An approach to analyze large scale wireless sensors network data," *Int. Res. J. Comput. Sci.*, vol. 2, no. 5, pp. 1–6, May 2015.

[13] E. Capriolo, D. Wampler, and J. Rutherglen, *Programming Hive*. Sebastopol, CA, USA: O'Reilly Media, 2012.

[14] L. G. Rio and J. A. I. Diguez, "Big data infrastructure for analyzing data generated by wireless sensor networks," in *Proc. IEEE Int. Congr. Big Data*, Jun. 2014, pp. 816–823.

[15] M. S. Rudresh, S. V. Shashikala, and G. K. Ravikumar, "Efficient handling of big data analytics in densely distributed sensor networks," *Int. J. Innov. Sci., Eng. Technol.*, vol. 2, no. 2, pp. 214–221, Feb. 2015.

[16] M. Vodel, M. Caspar, and W. Hardt, "Critical parameters for the efficient usage of wake-up-receiver technologies," in *Proc. ICCAN*, 2011, pp. 100–105.

[17] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of big data challenges and analytical methods," *J. Bus. Res.*, vol. 70, pp. 263–286, Jan. 2017.

[18] T. Dr AbdulRazak, R. Rajakumar, and M. Rameeja, "Improving wireless sensor network performance using bigdata and clustering approach," *Int. J. Sci. Res. Publ.*, vol. 4, pp. 1–7, Aug. 2014.

[19] G. S. Kunal *et al.*, "An efficient EM-algorithm for big data in wireless sensor network using mobile sink," *Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 5, pp. 2201–2205, 2016.

[20] J. Zhou, Y. Zhang, Y. Jiang, C. L. P. Chen, and L. Chen, "A distributed k-means clustering algorithm in wireless sensor networks," in *Proc. Int. Conf. Inform. Cybern. Comput. Social Syst. (ICCSS)*, Aug. 2015, pp. 26–30.

[21] Doreswamy and G. S. Kunal, "DGC-SOM Clustering algorithm for efficient big data gathering in densely distributed wireless sensor network," *Int. J. Latest Trends Eng. Technol.*, pp. 40–47, 2017.

[22] L. D. Alahakoon, S. K. Halgamuge, and B. Srinivasan, "A structure adapting feature map for optimal cluster representation," in *Proc. Int. Conf. Neural Inf. Process.*, 1998, pp. 809–812.

[23] D. Alahakoon, S. K. Halgamuge, and B. Srinivasan, "A self-growing cluster development approach to data mining," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, vol. 3. Oct. 1998, pp. 2901–2906.

[24] A. S. Pattanshett and N. D. Kale, "A survey on big-data gathering using mobile collector in densely deployed wireless sensor network," *Int. J. Eng. Res. Technol.*, vol. 3, no. 12, pp. 108–112, Dec. 2014.

[25] D. Takaishi, H. Nishiyama, N. Kato, and R. Miura, "Toward energy efficient big data gathering in densely distributed sensor networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 388–397, Sep. 2014.

[26] M. Wu, L. Tan, and N. Xiong, "A structure fidelity approach for big data collection in wireless sensor networks," *Sensors*, vol. 15, no. 1, pp. 248–273, Jan. 2015.

[27] S. Arivoli and V. Chitra, "Big data gathering in wireless sensor network using hybrid dynamic energy routing protocol," *Int. J. Manage., Inf. Technol. Eng.*, vol. 4, no. 4, pp. 59–68, Apr. 2016.

[28] L.-M. K. Ang, J. K. P. Seng, and A. M. Zungeru, "Optimizing energy consumption for big data collection in large-scale wireless sensor networks with mobile collectors," *IEEE Syst. J.*, vol. 12, no. 1, pp. 616–626, Mar. 2018. [Online]. Available: http://ieeexplore.ieee.org/document/7914656/

[29] B. Saneja and R. Rani, "An efficient approach for outlier detection in big sensor data of health care," *Int. J. Commun. Syst.*, vol. 30, no. 17, p. e3352, Nov. 2017.

[30] E. Ahmed *et al.*, "The role of big data analytics in Internet of Things," *Comput. Netw.*, vol. 129, pp. 459–471, Dec. 2017.

[31] B. Liu, J. Cao, J. Yin, W. Yu, B. Liu, and X. Fu, "Disjoint multi mobile agent itinerary planning for big data analytics," *EURASIP J. Wireless Commun. Netw.*, vol. 1, p. 99, Dec. 2016.

[32] L. Singh and D. Kumar, "A big data analysis for risk identification on wireless sensor network," *World Wide J. Multidiscipl. Res. Develop.*, vol. 3, no. 8, pp. 337–343, 2017.

[33] O. Younis and S. Fahmy, "HEED: A hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks," *IEEE Trans. Mobile Comput.*, vol. 3, no. 4, pp. 366–379, Oct. 2004.

[34] H. Kaur and R. Rjput, "Big data analysis on WSN for risk analysis on different data," *World Wide J. Multidiscipl. Res. Develop.*, vol. 3, no. 7, pp. 143–148, 2017.

[35] J. Xu, S. Guo, B. Xiao, and J. He, "Energy-efficient big data storage and retrieval for wireless sensor networks with nonuniform node distribution," *Concurrency Comput. Pract. Exper.*, vol. 27, no. 18, pp. 5765–5779, 2015.

[36] T.-Y. Tsai, W.-C. Lan, C. Liu, and M.-T. Sun, "Distributed compressive data aggregation in large-scale wireless sensor networks," *J. Adv. Comput. Netw.*, vol. 1, no. 4, pp. 295–300, Dec. 2013.

[37] L. Karim and M. S. Al-kahtani, "Sensor data aggregation in a multi-layer big data framework," in *Proc. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, Oct. 2016, pp. 1–7.

[38] M. S. Al-Kahtani, "Efficient cluster-based sleep scheduling for M2M communication network," *Arabic J. Sci. Eng.*, vol. 40, no. 8, pp. 2361–2373, 2015.

[39] L. Karim, N. Nasser, and T. El Salti, "Routing on mini-gabriel graphs in wireless sensor networks," in *Proc. IEEE WiMob*, Wuhan, China, Oct. 2011, pp. 105–110.

[40] L. Karim, N. Nasser, and T. Sheltami, "A fault-tolerant energy-efficient clustering protocol of a wireless sensor network," *Wireless Commun. Mobile Comput.*, vol. 14, no. 2, pp. 175–185, 2014.

[41] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proc. 33rd Annu. Hawaii Int. Conf. Syst. Sci.*, vol. 2. Jan. 2000, p. 10.

[42] L. Cheng, S. Guo, Y. Wang, and Y. Yang, "Lifting wavelet compression based data aggregation in big data wireless sensor networks," in *Proc. IEEE 22nd Int. Conf. Parallel Distrib. Syst.*, Wuhan, China, Dec. 2016, pp. 561–568.

[43] J. Li, S. Guo, Y. Yang, and J. He, "Data aggregation with principal component analysis in big data wireless sensor networks," in *Proc. 12th Int. Conf. Mobile Ad-Hoc Sensor Netw.*, Dec. 2016, pp. 45–51.

[44] D. Wu, B. Yang, and R. Wang, "Scalable privacy-preserving big data aggregation mechanism," *Digit. Commun. Netw.*, vol. 2, no. 3, pp. 122–129, 2016.

[45] S. Din, A. Ahmad, A. Paul, M. M. U. Rathore, and J. Gwanggil, "A cluster-based data fusion technique to analyze big data in wireless multi-sensor system," *IEEE Access*, vol. 5, pp. 5069–5083, 2017.

[46] D. Bol *et al.*, "Green SoCs for a sustainable Internet-of-Things," in *Proc. IEEE Faible Tension Faible Consommation (FTFC)*, Jun. 2013, pp. 1–4.

[47] T. Han and N. Ansari, "Heuristic relay assignments for green relay assisted device to device communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2013, pp. 468–473.

[48] E. F. Nakamura, A. A. F. Loureiro, and A. C. Loureiro, "Information fusion for wireless sensor networks: Methods, models, and classifications," *ACM Comput. Surv.*, vol. 39, no. 3, p. 9, 2007.

SABRINA BOUBICHE received the Magister degree in distributed and mobile computer sciences from the National Computer Science Engineering School, Algiers, Algeria, in 2015, and the B.Eng. degree in computer engineering from UHLB University, Algeria, in 2002. She is currently an Assistant Professor with the Computer Sciences Department, University of Constantine 2, Algeria. Her research interests include wireless communication, big data, and sensor networks. Her research has been published in various renowned international journals, such as *IEEE Communications*, *Wireless Personal Communications*, *Telecommunication Systems*, and *Sensors*.

DJALLEL EDDINE BOUBICHE received the HDR Diploma (Habilitation to conduct research) degree from the University of Batna 2, Algeria, in 2015 and the Ph.D. degree in computer science from UHLB University, Algeria, in 2013. He is currently an Associate Professor with the Computer Sciences Department, University of Batna 2, where he is also a member of the Lastic Laboratory and the Scientific Committee, Computer Sciences Department. His research interests include wireless communication, emended system, intelligent multi-agent system, distributed computing, and sensor networks. His research has been published in various renowned international journals, such as the *International Journal of Sensor Networks*, *IEEE Communications*, *Wireless Personal Communications*, *Telecommunication Systems*, *Future Generation Computer Systems*, *Sensors*, and so on. He frequently serves as a Program Committee Member such as: ICACIS'12 ICCMIT'15, ECARS'15, and ISNCC 2016. He was a Program Chair of the AWICT'15 and IPAC'15 international conferences, and a General Chair and Co-General Chair of BDAW'16 and ICC'16. He is also a reviewer of some renowned international journals. He is a Guest Editor of several special issue organized in international indexed journals such as: special issue on high-performance information technologies for engineering applications organized in the *International Journal of Computational Science and Engineering*, special issue on advanced information processing in communication organized in the *International Journal of Computational Science and Engineering*, special issue on artificial intelligence and knowledge computing organized in the *International Journal Artificial Intelligence and Soft Computing* and special issue on advanced industrial wireless sensor networks and intelligent IoT organized in *IEEE Communications Magazine*.

AZEDDINE BILAMI received the Ph.D. degree in 2005. He is currently serving as a Professor (full position) with the Department of Computer Science, and the Head of the Lastic laboratory, University of Batna 2. He has over 50 publications in international journals and conferences, including IJCA (ACTA press), IJSNet (Inderscience), the IEEE Communications Letters, Springer Verlag, IGI Global, and Elsevier publications. His current research interests include network security, routing protocols, mobility, QoE and QoS in wireless and mobile networks, WSNs, and Internet of Things. He also served as a member of editorial boards, steering committees, and technical committees in many international conferences and journals, including COMNET, COMCOM (Elsevier), HINDAWI journals, the *KSII Transactions on Internet and Information Systems*, IAJIT, Journal: *Journal of King Saud University*, and NGNS.

HOMERO TORAL-CRUZ received the B.Sc. degree in electronic engineering from the Instituto Tecnológico de la Laguna Coahuila, Mexico, in 2002, and the M.S. and Ph.D. degrees in electrical engineering, telecommunication option, from the Center for Research and Advanced Studies, National Polytechnic Institute (CINVESTAV), Jalisco, Mexico, in 2006 and 2010, respectively. He served as an Assistant Research with the Electrical Engineering Department, Telecommunication Section in CINVESTAV, and he was a Professor with the Electronic Engineering Department, Instituto Tecnológico Superior de Las Choapas, Veracruz, Mexico. He is currently an Associate Professor with the Sciences and Engineering Department, University of Quintana Roo, Mexico. He has published around 20 research papers in renowned international journals and conferences, and eight book chapters published by CRC Press, Taylor & Francis Group, and IGI Global. His research interest includes VoIP technologies, QoS and network measurements, convergent networks, Internet technologies, IP traffic modeling, network performance evaluation, and WSN. Furthermor e, he received a national recognition as a Researcher (SNI level C) by CONACYT and has been elected as a member of the Mexican Academy of Sciences and the Networking and Distributed Computing Laboratory, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia. He has served as a technical program committee member of some international conferences and workshops. He is also an active reviewer of some renowned international journals. He has served as a guest editor of some international journals of Springer-Verlag, Hindawi, and Inderscience.