Dynamic Data Aggregation Approach for Sensor-Based Big Data

Mohammed S. Al-kahtani
Dept. of Computer Engineering
Prince Sattam bin Abdulaziz University, Saudi Arabia

Lutful Karim
School of Information and Communications Technology
Seneca College of Applied Arts & Technology, Canada

***Abstract*—Sensors are being used in thousands of applications such as agriculture, health monitoring, air and water pollution monitoring, traffic monitoring and control. As these applications collect zettabytes of data everyday sensors play an integral role into big data. However, most of these data are redundant, and useless. Thus, efficient data aggregation and processing are significantly important in reducing redundant and useless data in sensor-based big data frameworks. Current studies on big data analytics do not focus on aggregating and filtering data at multiple layers of big data frameworks especially at the lower level at data collecting nodes (sensors) that reduce the processing overhead at the upper layer, i.e., big data server. Thus, this paper introduces a multi-tier data aggregation technique for sensor-based big data frameworks. While this work focuses more on data aggregation at sensor networks. To achieve energy efficiency it also demonstrates that efficient data processing at lower layers (sensor) significantly reduces overall energy consumption of the network and data transmission latency.**

Keywords*—Data aggregation; big data; sensor networks; energy efficiency; clustering

I. INTRODUCTION

The time of spreadsheet is over. A Google search, a barcode scan, a voice message, a picture of a car, a tweet among others all contains data that can be collected, analyzed and monetized. Indeed in today's time, we manage and store our life online. Data are gathered from smart phones, laptops and tablets that collect and transfer information on what people do. However, this is just the beginning. Most devices including our TVs, watches and even washing machines will collect and transmit messages. With the growing amount of information that exceed quintillion of bytes, new machines and techniques more powerful than the normal computer had to be created to allow us to make sense of the zeros and ones. Super computers and various algorithms have helped one so far in the real time analysis of those increasingly larger amounts of information. Nevertheless, for more efficient data mining, one always has to be on the chase for new methods.

The term Big Data refers to large volume of data sets. In the last few years, with the increase in the amount of digital information around us, the term has gained in popularity. As we speak, many professional in the field are working on finding better data mining ways to cope for the future. Sensors, mobile phones and other devices all generate big data. One can simply question what is the advantage of collecting so much information and how can it be useful for any company? The simplest example to answer such a question is the grocery stores/supermarkets. These stores offer various promotions and discounts upon using their cards such as Air Miles, Optimum card etc. These cards generate big data in the form of collected information in regards to demand and supply among various parameters stated in the contract signed by the customer. All the information are gathered and once processed, they help companies improve their businesses in various ways. Indeed, the primary goal of collecting these huge datasets is to look for meaningful patterns by using optimal processing.

Emergence of sensor networks also play a major role in the rise of big data as thousands of sensor network applications collect huge amount of data that require processing. Hence, sensors data processing can be considered as a part of big processing. As sensors produce redundant data we can aggregate data to reduce and represent them in a meaningful way in big data framework. However, works on big data presented in [9]-[13] do not talk on sensor-based big data aggregation, they mostly talk about architecture and network theory of big data, data mining, and application of big data.

As sensors-based big data aggregation is an important area of research to reduce computational cost as well as energy consumption this paper introduces a sensor data aggregation approach for a multi-tier big data framework. The proposed aggregation approach is designed in three layers to ensure that sensors data aggregation is facilitated at the lowest layer. As the proposed communication framework only consists three layers of communication and processing devices (i.e., sensors, gateway node that connects to Internet, and big data server) this data aggregation approach has three layers.

The proposed data aggregation allows both cluster-based and tree-based network topologies and thus, considered as a hybrid data aggregation approach. Clustering is used in most sensor network applications especially, they are greatly required for emergency or real-time applications such as rescue operations, health, and traffic monitoring to reduce data transmission latency (results in reduced data processing delay and overhead at big data server). On the other hand, tree-based approach achieves efficiency in non-real time applications where achieving energy efficiency is more important than data transmission delay. The proposed approach works by selecting a few nodes that work as active nodes [19] to collect and aggregate data for a certain period of time unless the residual energy of these nodes become critical. While most clustering algorithms [1], [4]-[8], [18]-[20] allow all member nodes of a cluster to actively work at any time instant the proposed

approach selects only a few nodes as active to work at any time instant that cover the whole network area. The proposed approach allows other nodes to work as alternative nodes that take the responsibility of active nodes only when any active node fails. This results in fault tolerance and energy efficiency. The rest of the paper is organized as follows.

Section II briefly presents literature on sensor data aggregation approaches. Section III briefly presents the working principle of the proposed data aggregation approach. Section IV analyzes the performance of the proposed data aggregation approach and compares it with tree and cluster-based approaches in terms of energy consumption and data transmission latency. Experimental (simulation) setup and results are presented in Section V. Finally, the summary of the paper and future works are presented in Section VI.

II. Related Work

Current research on big data analytics include distributed algorithms to process big data, network architecture and application of big data, MapReduce paradigm that works on big data [9]-[15]. The existing distributing algorithms to process and aggregation big data are mostly done at high performance big data server. These studies [9]-[15] do not consider data aggregation at multiple layers especially sensor data aggregation at the data collecting side as a way to reduce computational cost. Hence, we studied and presented a few literatures on sensor data aggregation as follows as a plan to integrate an improved sensor data aggregation approach in our proposed sensor-based big data framework.

Directed diffusion (DD) is a flat data aggregation approach where a node *A* broadcasts its interest and the node *B* that senses data related to the interest message transmits to *A* though multiple paths. Later, the node *A* selects the shortest path for further data transmission through a reinforcement packet. However, DD requires a large number of data transmissions. Hence, Cluster diffusion with Dynamic Data Aggregation Approach (CLUDDA) [3], [16] is introduced to only propagate event of interest and interest event between cluster head and cluster members. In case, the cluster head resides far from the cluster members, it consumes huge energy.

Tree-based approaches are good for small networks with fewer nodes. However, these algorithms suffer from a single point of failure where the failure of a single node disconnects the data transmission path from leaf node to the root. Among many tree-based approaches, energy aware distributed heuristic (EADAT) [17], Power efficient data gathering and aggregation protocol (PEDAP) [18] based on a spanning tree to maximize the lifetime of the network and Power-Aware PEDAP (PEDAP-PA) [18] are more popular. Chain-based data aggregation techniques, such as power efficient data gathering protocol for sensor information systems (PEGASIS), have been proposed [20] where each sensor transmits only to its closest neighbor. As this approach does not guarantee the shortest data transmission path from the furthest nodes of the chain to the sink a multiple-chain scheme is introduced in [20]. Again, this approach does not provide the shortest data transmission distance. Hence, the greedy chain construction algorithm, which constructs the chain by starting at the furthest node from the sink and considers it as a chain head, was proposed in [5]. Every time a non-chain node is added to the chain, this new node is considered as a new chain head until all nodes are added to the chain.

A multiple chain scheme has also been proposed in [22]. In this approach, the network is divided into four zones and each zone is centered at the node that is closest to the center of the sensing region. A linear that ends at the centre node is created for each zone. The multiple chain schemes aim to decrease the total distance of transmitting data as nodes broadcasts. In the greedy chain construction scheme proposed in [12], the process starts by selecting the chain head. The farthest node from the sink is selected as the chain head. At each step, a non-chain node, *A* is added to the chain head if *A* is closest to the chain head. The procedure stops whenever all nodes are added to the chain. This approach is further improved by including the non-chain node to the chain as a chain leader that provides the shorted distance as compared to other nodes if included into the chain as a leader.

In the grid-based data aggregation method [18], each grid has a data aggregator and all sensors in a grid transmit data to the grid aggregator while in the in-network data aggregation, data are aggregated at parent nodes as they are being transmitted towards sink at the root of the tree. The work in [5] presents a hybrid data aggregation scheme that combines the best features of grid-based and In-network aggregation schemes. The network topology is initially constructed based on in-network data aggregation approach. Once an event is detected by a sensor, the sensor follows in-network data aggregation scheme if the data is received from a static sensor application. If data is from a mobile sensor application, grid-based approach is used for data aggregation. Among other approaches, the work done in [26] introduces a cluster-based data aggregation approach where cluster head uses three different approaches to reduce redundant data collected from neighboring nodes (i.e., huge processing burden on cluster head), [27] introduces identity-based aggregate signature (IBAS) scheme for sensor-based secure data aggregation that provides data integrity as well as reduce bandwidth usage.

In sensor network, nodes receive data only when they are in active state that introduces the idea of properly utilizing the limited number of active time slots of sensor nodes with the goal of reducing data aggregation latency. The minimum latency aggregation schedule (MLAS) in most duty cycle WSN allows low latency and collision free aggregation schedule. However, this approach uses fixed structure aggregation methods and requires all sensor nodes are always awake. The work done in [28] introduces a distributed aggregation algorithm for duty-cycle WSNs, in which the aggregation tree and a conflict free schedule are generated simultaneously without using any fixed aggregation structure. The work done in [29] introduces an approximation algorithm to construct a maximum lifetime data aggregation tree that uses an adjustable transmission power level to achieve higher network lifetime while most work consider fixed transmission power. In [30], authors introduce a cluster-based approach for in-network aggregation. This approach uses an energy efficient routing strategy that uses multi-path routing tree and performs data fusion and data aggregation at intermediate

nodes. While most data aggregation approached do not consider data security and privacy issues, Vakilinia et al. [31] presents data privacy preserving data aggregation/fusion approach for crowdsensing that uses linear transformation and homomorphic encryption scheme to obtain secured aggregated data. However, these approaches are complex and computationally expensive.

The work done in [32] presents several data fusion techniques such as approaches based on neural network, genetic algorithm, fuzzy logic, particle swarm optimization, steiner tree-based approach and data selection-based summation fusion. In [33] Yan M. introduces Forecast Algorithm of Data Aggregation (FTDA) data fusion algorithm based on the time prediction model, which predicts a time when data may differentiate from the data at current time. This model has the ability to proactively identify data redundancy and reduce energy consumption. However, approaches presented in [32], [33] work for small scale sensor networks, require more computational power and hence, have space to make them more energy efficient.

Most approaches that we have presented in this section do not consider selecting a fewer number of nodes as active nodes and allowing all other nodes to remain in sleep state (or idle) that reduce the network energy consumptions. Also they do not consider the type and priority of data packets for data aggregation. Hence, we introduce a multi-tier data aggregation approach that (1) uses both cluster and tree-based approaches, (2) selects only a few nodes as active node while keep all other nodes in sleep state, (3) assigns type and priority to each data packet.

III. Proposed Architecture and Approach

This section presents the high level architecture of the proposed data aggregation framework of big data along with the low level data aggregation and filtering scheme at sensor networks.

A. High Level Architecture

The proposed big data aggregation and filtering framework works in three layers, (1) Lower Layer: aggregates data at sensors (2) Middle Layer: aggregates data at base station (3) Upper Layer: aggregates data aggregation at big data server in distributed manner.

Fig. 1 illustrates such as a big data framework that only has three data communication layers. For instance, sensors at lower layers sense data and transmit those data to sink node or base station (BS). Then, the BS processes or aggregates data and transmit the aggregated data to the central big data sever through Internet. Finally, the big data server aggregates data by distributing it to commodity computers. Hence, the proposed hybrid data aggregation scheme has three data aggregation layers. The computational efficiency of big data sever at upper layer depends on data aggregation at data at middle and lower layers as low power nodes at these layers can aggregate and filter data to some extent even though nodes at upper the layer have higher computational power. However, existing big data aggregation approaches in literature are mostly only designed for upper layer at big data server. Hence, the computational cost or time at the server is not reduced as these approached do not consider any lower layers preprocessing of data (such as preprocessed at lower layers at sensors).

By designing efficient data aggregation approach at the lower level sensor nodes the overall computational costs at the upper layer big data server can be reduced, which is the objective of this paper as the data aggregation scheme reduces the volume of sensor's data that will be transmitted to the upper layer. Thus, this approach reduces data aggregation and processing overhead at the upper layer in NoSQL or other non-relational database systems for big data. The upper layer also consists of emergency response centre. The sink or base station at middle layer transmits emergency or time critical data to the emergency response centre before sending it to NoSQL database servers for processing/filtering and future storage.

Sensor networks are being used for many applications. These applications can be classified as (1) real-time and (2) non-real-time. Real-time applications such as health monitoring have more priority than non-real-time applications (i.e., real-time emergency data should have more priority than non-real-time data). Hence, data aggregation approaches should be designed considering the priority of sensor applications or data types. Most existing approaches [1], [4]-[8], [18], [23] do not consider this criteria to design a data aggregation approach.

Moreover, data processing at upper layer (i.e., at big data servers) should also consider the type of data so that data can be stored based on their categories for future use. Data aggregations at the lower and middle layers based on data types and sensor applications will ease the data processing at the upper layer. Thus, this paper introduces an energy efficient application dependent data aggregation approach for sensor-based big data frameworks. Sensors are programmed to have a data type field in their packets so that other sensors or devices that receive the data packet can identify the type of applications and perform data aggregation based on the data type [21]. This field also helps to store data at the appropriate locations in big data server for further processing and use.

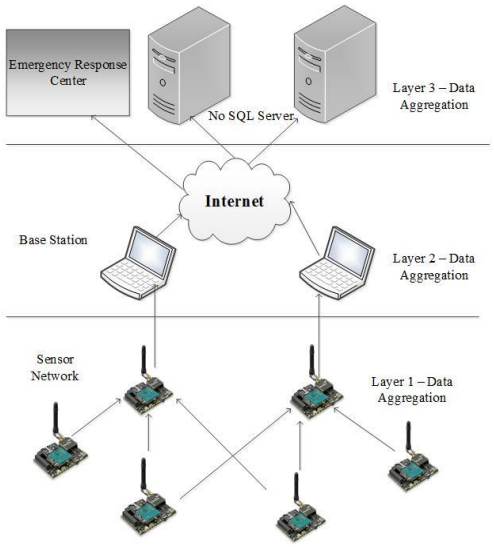

Fig. 1. 3-tier sensor-based big data aggregation framework.

Routing protocols can be proactive (periodic) and reactive (event-based). For periodic routing protocols, data are sensed and transmitted periodically – at a certain time interval. In reactive routing protocols, data are transmitted only when a certain event is triggered. Sensors will also be programmed to contain a field (i.e., routing type) in their data packet that data transmission mode. For instance, if the routing field is set to 1 it will represent the periodic data transmission of emergency/real-time applications. Otherwise, data transmission will be event-based. Data aggregation at sensors also depends on this field. In the proposed approach, emergency real-time data will be only aggregated or filtered at sensors to avoid transmitting redundant data (i.e., data with the same information that has already been transmitted) that will reduce network energy consumption and also allow the sink to transmit data faster to the emergency response centre. Moreover, more data aggregation and processing takes place at the middle layer (at base station or sink node) compared to that at the lower layer (i.e., at the sensor) since sensors have limited power and processing capabilities. Thus, big data servers at the upper layer are expected to receive partially structured data to reduce the overall processing overhead of big data framework.

B. Proposed Hybrid Sensor Data Aggregation Scheme

The proposed hybrid data aggregation scheme classifies sensor-based applications into the following categories.

1) Real-time, emergency, time critical applications – such as traffic monitoring, battlefield surveillance and health monitoring.

2) Non-real-time applications – agriculture, air pollution monitoring.

The lower layer sensors transmit data to the upper layers through gateway nodes. Fig. 2 illustrates such a scenario. However, data aggregation approaches may achieve energy and computation efficiency using dynamic network topologies based on the requirement of sensors applications. For example, sensors are programmed to form cluster-based topology for emergency real-time applications and tree topology for non-real-time applications (details of cluster formation, tree formation and CH selections are presented in [24]). In cluster-based topology, sensors collect and transmit data at their allocated timeslot to the cluster head (CH). Then the CH transmits to the gateway and end station. As this type of topology ensures the minimum number of hops to transmit to the end node data aggregation using cluster-based approach is expected to achieve computational, data latency as well as energy efficiency. In cluster-based data aggregation, once a cluster is formed and CH is selected the CH selects a minimum number of nodes as active node for any time instant while other nodes remain in sleep state (or idle). We use the work done in [19] to select active nodes. Active nodes of a cluster sense and transmit data to CHs while aggregates and filters data to discard redundant data. On the other hand, idle nodes (in sleep state) do not perform data sensing, transmission and aggregation. By discarding a large number of redundant and useless data in emergency applications this approach ensures faster transmission of data to the central server [1].

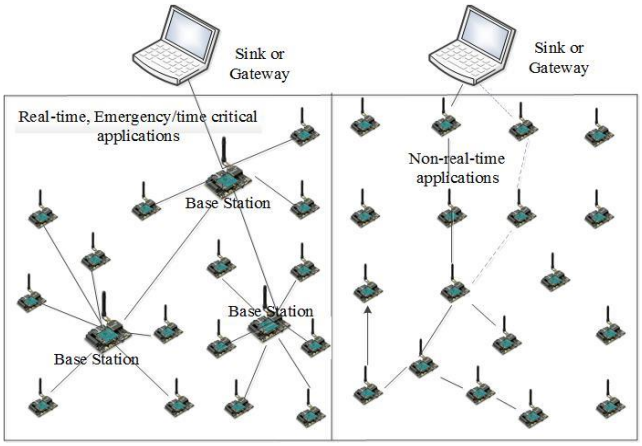

Fig. 2. Layer 1 data aggregation.

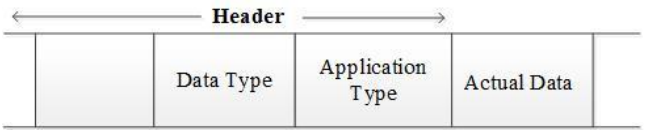

Fig. 3. Format of sensor data packet.

On the other hand, achieving energy efficiency is more important than achieving reduced end-to-end delay in non-real-time applications, such as agriculture, farming, pollution monitoring. Tree-based hierarchical topology may create the shorter path that uses more hops as compared to cluster-based topology. As distance is less the energy consumption will be less (energy consumption is directly proportional to the distance between two nodes [2], [3]). Thus, in tree-based data aggregation approaches, a sensor transmits data through the shortest path from itself to the sensor gateway.

In tree-based approach, nodes are identified to locate at different levels of the hierarchy considering the gateway node is the root of the hierarchy. Nodes residing one-hop away from the gateway can be considered to locate at the level 1 and so on. Then, the shortest path from the sensor gateway node to the active leaf nodes will be created using the method presented in [19]. Data transmission starts at sensors of the lowest level. For instance, active sensor nodes (or leaf nodes) sense and transmit the event of interest to the active nodes at the upper level. Parent nodes in this tree structure always perform data aggregation using different aggregation functions such as MAX, MIN, MEAN, MEDIAN and SUM and transmit again to the active nodes at the upper level until data reaches at the sensor gateway at root. Thus, this energy consumption of the active nodes in this approach is well distributed and the total network energy consumption is expected to be lower even though the number of hops from the sensor to the gateway is more as compared to the cluster-based counterpart of this proposed approach. However, the tree-based data aggregation may result in increased end-to-end data transmission delay as data from a node passes through several number of hops and is processed at each node for a certain time period. Thus, the proposed hybrid, dynamic and application-based data aggregation scheme offers a trade-off between energy efficiency and data transmission delay.

TABLE I. REPRESENTING TERMINOLOGIES BY SYMBOLS

Name of the terminology	Symbol
Sensor Network	$G(V, E)$
Non-real-time applications	nr
Real-time applications	r
Data type	Dt
Application type	AP
Tree-based topology	Tr
Cluster-based topology	Cl
Cluster head	CH
Level in a hierarchy	L
Active nodes	AN
Alternative nodes	Al
Gateway Node	G

Normally, sensor networks are used for a specific application by forming a specific network topology. Using the proposed data aggregation scheme, the sensors in a network can be reused to other applications and are able to change their topology if the application changes. Data packets have a number of fields and one field is used to set the application type. Once sensors receive a data packet from the gateway with the changed application field, it reconstructs the topology. Fig. 3 illustrates a sensor data packet that contains fields to identify data type and application type for the proposed data aggregation framework.

Sensor networks are mostly designed for a specific application and hence, a data aggregation scheme (cluster or tree-based) can be pre-established. However, the data aggregation scheme can also be constructed on-demand based on the types of packets that sensors transmit. This dynamism allows sensor networks to be used or re-used in multiple applications. Algorithm 1 presents the pseudo-code for the proposed sensor data aggregation approach. Table I lists the symbol used for different terms in Algorithm 1.

Algorithm I: Proposed Hybrid Data Aggregation Scheme

Randomly pick a node *i*
Set $node_i \leftarrow active$
$activenodeset \leftarrow \{i\}$
while $WholeNetCovered \neq TRUE$
  pick node j randomly
  if $j \neq i$ & *not in activenodeset* & $NetCoverage(node_i) \cap NetCoverage(node_j) = Null\ or\ Minimal$
then $node_j \leftarrow active$
    $activenodeset \leftarrow \{i, j\}$
  else
    $node_j \leftarrow alternative$
    $alternativeenodeset \leftarrow \{j\}$
  end if
end while
Remaining nodes $\in$ *sleep-mode*
If $AP = nr$ **then**
  $G(V, E) \leftarrow Tr$
  form shortest path with *AN* in leaf nodes to *g*
  AN at different *L* transmit multi-hop
else if a $AP = r$ **then**
  $G(V, E) \leftarrow Cl$
  select CHs from *AN*s
  AN in each cluster transmit towards *CH*s
end if
while $G(V, E)$ in work **do**
  if $AP = r$ & $Dt = r$ **then**
    for each AN_i in cluster j **do**
      transmit data to *CH*
      CH filters redundant data & transmit to *g*
    end for
  else if $AP = r$ & $Dt = nr$ **then**
    reconstruct $G(V, E) \leftarrow Tr$
    aggregation level $\leftarrow$ *L*i
    CH aggregates using *MAX, MIN, SUM, REDUCE* & other functions based on *AP*
    CH transmits aggregated data to *g* directly or through other *CH*s
  else if $AP = nr$ & $Dt = nr$ **then**
    aggregation level $\leftarrow$ *L*i
    CH aggregates using *MAX, MIN, SUM, REDUCE* & other functions based on *AP*
    CH transmits aggregated data to *g* directly or through other *CH*s
  else if $AP = nr$ & $Dt = r$ **then**
    reconstruct $G(V, E) \leftarrow Cl$
    for each AN_i in cluster j **do**
      transmit data to *CH*
      CH filters redundant data & transmit to *g*
    end for
  end if
end while

IV. PERFORMANCE ANALYSIS

In this section, the performance of the proposed data aggregation scheme will be analyzed in terms of networks energy consumption and data transmission delay. Then we will set up the network simulator based on some assumptions and measure the performance of the proposed hybrid data aggregation scheme as compared to the tree and cluster-based approaches.

A. Energy Model

The energy model in [2], [3] is used to evaluate the performance of the proposed data aggregation approach as we only consider data transmission and reception energy consumption in this evaluation. This model considers that energy consumption is proportional to data transmission distance. The energy consumption of a node for transmitting

data of n_{data} bytes to another node, which are at distance d apart is

$$E_{TX} = n_{data} \times \varepsilon_{data} + n_{data} \times d^2 \times \varepsilon_{air} \quad (1)$$

However, the energy consumption of a node for receiving a data packet is independent of distance and is denoted as follow.

$$E_{RX} = n_{data} \varepsilon_{data} \quad (2)$$

Where ε_{data} is the energy consumption of a sensor node in its electronic circuitry and ε_{air} represent the energy consumptions in RF amplifiers for propagation loss.

B. Estimation of Energy Dissipation

Let us assume that the number of sensor applications = n_{app} and the number of non-real-time applications that use tree topology = n_{nr}

The number of real-time applications that use cluster-based topology = n_r .

$$\therefore n_{nr} + n_r = n_{app} \quad (3)$$

Let us assume that each network has the same number of nodes, n_{node} .

Therefore, the total number of nodes = $n_{node} \times n_{app}$.

1) Existing cluster-based method

Let us assume that the number of clusters in each network is n_{cl} .

Therefore, the number of nodes in each cluster,

$$n_{nodecl} = \frac{n_{node}}{n_{cl}} . \quad (4)$$

Let us assume that each network has 2 level or hierarchy. We denote these levels as L_1 and L_2 . Also, we consider that the level that is closer to the gateway is L_2 . So, the number of clusters in each level is $\frac{n_{cl}}{2}$.

Let us assume that the distance between an active member node and CH = d_{avg}

The average size of a data packet that is transmitted from a member node to CH is n_{data} .

Therefore, the total network energy consumption for transmitting a data packet to a cluster is

$$E_{TX1CL} = \left(\frac{n_{node}}{n_{cl}} - 1 \right) \times (n_{data} \times \varepsilon_{data} + n_{data} \times d_{avg}{}^2 \times \varepsilon_{air}) \quad (5)$$

The energy consumption of a CH for receiving data from an active member node is

$$E_{RXCL} = \left(\frac{n_{node}}{n_{cl}} - 1 \right) \times (n_{data} \times \varepsilon_{data}) \quad (6)$$

Similarly, the energy consumption of a CH to transmit data packet to the sensor gateway is given as

$$E_{TX\,2CL} = n_{agdatacl} \times \varepsilon_{data} + n_{agdatacl} \times d_{CH}{}^2 \times \varepsilon_{air} \quad (7)$$

Where the aggregated data size at CH is $n_{agdatacl}$ and the average distance between CH and sensor gateway is d_{CH}

Thus, the total transmission energy consumption in a cluster-based data aggregation scheme is

$$E_{TXCL} = n_{cl} \times (E_{TX1CL} + E_{TX\,2CL}) = n_{cl} \times (n_{data} \times \varepsilon_{data} + n_{data} \times d_{avg}{}^2 \times \varepsilon_{air}) + n_{cl} \times (n_{agdatacl} \times \varepsilon_{data} + n_{agdatacl} \times d_{CH}{}^2 \times \varepsilon_{air}) \quad (8)$$

2) Proposed hybrid approach

This section presents the proposed data aggregation scheme both for when (1) modifications are done based on cluster-based topology for real-time applications, and (2) modifications are done based on tree-based topology for non-real-time applications.

Proposed approach is based on cluster-based topology for real-time applications

Let us assume that the number of nodes that reside in sleep mode = n_{idle}.

Therefore, the number of active nodes in a cluster including CH is

$$n_{activeprcl} = n_{nodeprcl} - n_{idle} . \quad (9)$$

If we substitute (9) into (5) we find the energy consumption of active nodes in a cluster for transmitting data to CH, which is given as follows:

$$E_{TX1PRCL} = (n_{nodeprcl} - n_{idle} - 1) \times (n_{data} \times \varepsilon_{data} + n_{data} \times d_{avg}{}^2 \times \varepsilon_{air}) \quad (10)$$

Similarly, if we substitute (9) into (6) we obtain the energy consumption of a CH for receiving data from an active member node of a cluster, which is given as follows:

$$E_{RXPRCL} = (n_{nodeprcl} - n_{idle} - 1) \times (n_{data} \times \varepsilon_{data}) \quad (11)$$

The energy consumption of a CH for transmitting a data packet to the sensor gateway is given as

$$E_{TX\,2PRCL} = n_{agdataprcl} \times \varepsilon_{data} + n_{agdataprcl} \times d_{CH}{}^2 \times \varepsilon_{air} \quad (12)$$

In (12) the aggregated data size at a CH is $n_{agdata-prcl} \leq n_{agdata-cl}$ and the average distance between a CH and sensor gateway is d_{CH} .

Thus, the total transmission energy consumption in the cluster-based proposed data aggregation approach is:

$$E_{TXPRCL} = E_{TX1PRCL} + E_{TX2PRCL} = (n_{nodeprcl} - n_{idle} - 1) \times (n_{data} \times \varepsilon_{data} + n_{data} \times {d_{avg}}^2 \times \varepsilon_{air}) + n_{agdataprcl} \times \varepsilon_{data} + n_{agdataprcl} \times {d_{CH}}^2 \times \varepsilon_{air} \quad (13)$$
$$= n_{activeprcl} \times (n_{data} \times \varepsilon_{data} + n_{data} \times {d_{avg}}^2 \times \varepsilon_{air}) + n_{agdataprcl} \times \varepsilon_{data} + n_{agdataprcl} \times {d_{CH}}^2 \times \varepsilon_{air}$$

Proposed approach is based on tree-based topology for non-real-time applications

Again let us assume that the number of levels from leaf nodes to the sensor gateway is 2.

The number of active nodes in each level is $n_{activeprtr}$.

The proposed data aggregation approach that uses tree topology creates the shortest path from a leaf node to the sensor gateway. We assume that the size of a data packet that is sensed at a leaf node is $n_{dataprtr}$ and the size of aggregated data packets at the upper level nodes is $n_{agdataprtr}$. The average distance between the nodes at level 1 and level 2 is d_{L1prtr} and between the nodes at level 2 and the sensor-gateway is d_{L2prtr}

Therefore, the average distance (shortest) between the leaf node and the sensor-gateway node is given as

$$d_{L1prtr} + d_{L2prtr} \quad (14)$$

Thus, the energy consumption of active nodes at $L1$ for transmitting data to the nodes at $L2$ is given as

$$E_{TX1PRTR} = n_{activeprtr} \times (n_{data} \times \varepsilon_{data} + n_{data} \times {d_{L1-prtr}}^2 \times \varepsilon_{air}) \quad (15)$$

The energy consumption of active nodes at $L2$ for receiving data from nodes at $L1$ is given as follows

$$E_{RXPRTR} = n_{activeprtr} \times (n_{data} \times \varepsilon_{data}) \quad (16)$$

Similarly, the energy consumption of all active nodes at $L2$ for transmitting data packets to the sensor-gateway is given as

$$E_{TX2PRTR} = n_{activeprtr} \times (n_{agdataprtr} \times \varepsilon_{data} + n_{agdataprtr} \times {d_{L2prtr}}^2 \times \varepsilon_{air}) \quad (17)$$

Thus, the total energy consumption for transmitting data in the tree-based proposed data aggregation approach is given as:

$$E_{TXPRTR} = E_{TX1PRTR} + E_{TX2PRTR} = n_{activeprtr} \times (n_{data} \times \varepsilon_{data} + n_{data} \times {d_{L1prtr}}^2 \times \varepsilon_{air}) + n_{activeprtr} \times (n_{agdataprtr} \times \varepsilon_{data} + n_{agdataprtr} \times {d_{L2prtr}}^2 \times \varepsilon_{air}) \quad (18)$$
$$= n_{activeprtr} \times \varepsilon_{data} (n_{data} + n_{agdataprtr}) + n_{activeprtr} \times \varepsilon_{air} (n_{data} \times {d_{L1prtr}}^2 + n_{agdataprtr} \times {d_{L2prtr}}^2)$$

3) Existing tree-based method

Let us assume that the number of nodes at level of the tree = n_{tr} and the number of hops to transmit data from a leaf node to the sensor-gateway = 2

Let us assume that the average distance from L1 nodes to L2 nodes = d_{L1tr}

The average distance from L2 nodes to sensor-gateway= d_{L2tr}

The size of data sensed at the lowest level leaf nodes= n_{datatr}.

Then, the size of aggregated data at L2 nodes = $n_{agdatatr} \geq n_{agdataprtr}$ (19)

In this approach, all nodes are kept in the inactive mode. Transmission energy consumption of L1 nodes as given in (20).

$$E_{TX1TR} = n_{tr} \times (n_{data} \times \varepsilon_{data} + n_{data} \times {d_{L1tr}}^2 \times \varepsilon_{air}) \quad (20)$$

Similarly, reception energy consumption of nodes at L2 is given as:

$$E_{RX-TR} = n_{tr} \times (n_{data} \times \varepsilon_{data}) \quad (21)$$

And energy consumption for transmitting data from nodes at L2 to the sensor-gateways deduced using (22).

$$E_{TX2TR} = n_{tr} \times (n_{agdatatr} \times \varepsilon_{data} + n_{agdatatr} \times {d_{L2tr}}^2 \times \varepsilon_{air}) \quad (22)$$

Thus, the total transmission energy consumption is

$$E_{TXTR} = E_{TX1TR} + E_{TX2TR} = n_{tr} \times (n_{data} \times \varepsilon_{data} + n_{data} \times {d_{L1tr}}^2 \times \varepsilon_{air}) + n_{tr} \times (n_{agdatatr} \times \varepsilon_{data} + n_{agdatatr} \times {d_{L2tr}}^2 \times \varepsilon_{air}) \quad (23)$$
$$= n_{tr} \times \varepsilon_{data} (n_{data} + n_{agdatatr}) + n_{tr} \times \varepsilon_{air} (n_{data} \times {d_{L1tr}}^2 + n_{agdatatr} \times {d_{L2tr}}^2)$$

4) Comparison of energy consumption among cluster-based, tree-based and hybrid approach

Case 1: Non-real-time sensor applications using tree-based topology.

Since it is known that $n_{activeprtr} < n_{tr}$ we can conclude from (18) and (23) that

$$E_{TXPRTR} < E_{TXTR} \quad (24)$$

Similarly, we can conclude from equations (16) and (21) that

$$E_{RXPRTR} < E_{RXTR} \quad (25)$$

Case 2: Real-time sensor applications that use cluster-based topology.

It has been shown that $n_{active-prcl} < n_{cl}$, so, we can conclude from (8) and (12) that

$$E_{TXPRCL} < E_{TXCL} \quad (26)$$

Similarly, $E_{RXPRCL} < E_{RXCL}$ (27)

Case 3: Comprises of both real-time and non-real-time applications. Let us assume that the number of non-real-time and real-time applications are n_1 and n_2, respectively. Then, the transmission energy consumption for the proposed data aggregation approach will be given as

$$n_1 \times E_{TXPRTR} + n_2 \times E_{TXPRCL} \quad (28)$$

Where the transmission energy consumption for the cluster-based approach will be denoted as

$$n_1 \times E_{TXCL} + n_2 \times E_{TXCL} \quad (29)$$

Similarly, the transmission energy consumption for the tree-based approaches will be given as

$$n_1 \times E_{TXTR} + n_2 \times E_{TXTR} \quad (30)$$

Since $E_{TXPRCL} < E_{TXCL}$ comparing (28) and (29) we find that transmission energy consumption of the proposed approach will be less than the transmission energy consumption of the cluster-based approach. Similarly, as $E_{TXPRTR} < E_{TXTR}$ comparing (28) and (30), we find that transmission energy consumption will be less than that of tree-based approach.

We will find the similar result for data reception energy consumption (i.e., reception energy consumption of the proposed approach will be less than that of the cluster and tree-based approaches)

C. Analysis on Data Transmission Latency

In the cluster-based method, the active member nodes of a cluster transmit data packets to the CH. Then the CH aggregates and transmits the processed data to the sensor-gateway. If the time allocated to the active member node and CH are T_c and T_{ch}, $T_{ch} > T_c$ as the CH performs data sensing, data transmission, reception and aggregation.

The data transmission latency for the cluster-based method will be as presented in (31).

$$D_{cl} = n_{cl} \times \left(\left(\frac{n_{node}}{n_{cl}} - 1 \right) \times T_c + T_{ch} \right) \quad (31)$$

1) Proposed hybrid approach

The data transmission latency for the proposed cluster-based approach

$$D_{prcl} = n_{cl} \times \left(n_{activeprcl} \times T_c + T_{ch} \right) \quad (32)$$

The data transmission latency for the proposed tree-based method is presented in (33).

$$D_{prtr} = n_{activeprtr} \times T_{L1prtr} + n_{activeprtr} \times T_{L2prtr} \quad (33)$$

The number of active nodes in each level of the proposed tree-based method is presented in (34).

$$n_{activeprtr} < n_{node} \quad (34)$$

2) Existing tree-based method

The number of nodes in each level is assumed to be same $= n_{node}$ and duration of timeslot allocated to each node at the lowest level is T_{L1tr}.

The duration of timeslot allocated to each node at the upper level is $T_{L2tr} > T_{L1tr}$.

This is because the upper level nodes perform data aggregation and transmit aggregated data to the sensor-gateway.

Thus, the data transmission latency for tree-based approach will be

$$D_{tr} = n_{node} \times T_{L1tr} + n_{node} \times T_{L2tr} \quad (35)$$

3) Comparison of data transmission latency

Case 1: If all sensor applications of the proposed approach are non-real-time and use tree-based topology

By comparing (41), (42) and (43) we can conclude that $D_{prtr} < D_{tr}$

Case 2: If n_1 sensor applications of the proposed approach are non-real-time and n_2 applications are real-time, the data transmission latency will be

$$n_1 \times D_{prtr} + n_2 \times D_{prcl} = n_1 \times (n_{activeprtr} \times T_{L1prtr} + n_{activeprtr} \times T_{L2prtr}) + n_2 \times n_{cl} \times \left(n_{activeprcl} \times T_c + T_{ch} \right) \quad (36)$$

For tree-only approach the data transmission latency will be

$$(n_1 + n_2) \times D_{prtr} = n_1 \times (n_{node} \times T_{L1tr} + n_{node} \times T_{L2tr}) + n_2 \times (n_{node} \times T_{L1tr} + n_{node} \times T_{L2tr}) \quad (37)$$

As $n_{activeprtr} < n_{no\det r}$ and $T_{L1prtr} < T_{L1tr}$ we can conclude from (36) and (37).

$D_{pr} < D_{tr}$ (i.e., data transmission latency of proposed approach is lower than tree-based approach).

$D_{pr} < D_{cl}$ (i.e., data transmission latency of proposed approach is lower than cluster-based approach).

From the above analysis, we conclude that the energy consumption and data transmission delay of the proposed sensor-based data aggregation approach at layer 1 is less than that of traditional cluster and tree-based schemes.

D. Computational Complexity

If the number of active nodes at each level l in the proposed tree-based approach is $n_{l(activeprtr)}$ and the number of levels in the network is L_{prtr} the total number of active nodes will be $\sum_{l=1}^{L_{prtr}} n_{l(activeprtr)}$.

Thus, the number of packets transmitted by each active node of the network at their predefined timeslot is $\sum_{l=1}^{L_{prtr}} n_{l(activeprtr)}$.

If we define the complexity of the algorithm based on the number of message transmission, which is a function of the number of nodes from each level at the predefined timeslot then the processing complexity of the proposed approach based on tree topology is O (n) where n is the number of nodes transmitting data packets.

Similarly, we can show that the processing complexity of proposed approach based on clustering will O (n).

V. Validation of the Proposed Approach

To validate our proposed hybrid data aggregation and filtering technique for sensor-based big data frameworks we considered the scenarios presented in the section.

A. Simulation Setup

We designed and implemented a simulator to implement the proposed data aggregation approach using C programming language rather than using the existing simulators, NS-2, OPNET, NS-3 many sensor network and big data functionalities are not available in these simulators. Moreover, we have more control on implementing the new concept of sensor-based big data.

Real experiments or testbed always give accurate result as compared to simulation. However, real experiments are not always possible due to the unavailability of sensors and other components. Hence, simulation is being used to replace experimental work in sensor networks and other fields to a great extent. Hence, we decided to perform simulation to evaluate the performance of the proposed data aggregation scheme that works at layer 1 of the big data architecture and compared with the traditional cluster and tree-based approach as presented before. We use network energy consumption, network lifetime and data transmission latency as the performance metrics. Each time the simulator was run for a certain number of rounds and we run the simulator a certain number of times. The outputs are calculated as an average of these results. We define the performance metrics and related terms as follows:

Round – is a period of time comprises a number of network setup and operation phases.

Data transmission latency – is considered as the end-to-end data transmission delay, i.e., the time required to transmit data from an active node to the sensor gateway or base station.

Energy consumption – is the total energy consumed by a sensor to transmit, receive and aggregate data.

We simulated an area of size 100 meters x 100 meters as the network size. As this network area is considered as small, the network is divided into only 4 clusters and 20-30 nodes are randomly deployed on an average into each cluster (100 nodes in total into the network). For this small network area deploying 100 sensors can be considered as a large number of sensors that collect huge amount of data, i.e., big data. The proposed data aggregation approach still works even if we increase the size of the network and the number of sensors in this ratio (large scale). Simulation parameter and their respective values of our paper [25] are also used in this paper.

The simulator was run for rounds between 5000 and 30000 for different experiments to compare energy consumption between low (5000 rounds) and high (30,000 rounds) number of network setup phases. The sensor gateway is placed at the outside of the network area which is located at the co-ordinate (55, 101). During the network operation phase, cluster head allocates a number of timeslots to each node. However, each nodes receives different number of timeslots based on their distance or level from the sensor gateway. For instance, nodes which are closer to the sensor gateway require more time to sense data, receive data from lower levels of nodes, aggregate and transmit data. Hence, these nodes require more time (i.e., timeslots) as compared to the nodes that reside far from the sensor gateway at the lower levels. Table I lists the parameters and their values that are used in the simulation.

B. Simulation Results

Fig. 4 shows that the energy consumption of the proposed data aggregation approach is much lower than that in traditional tree and cluster-based approaches because the proposed approach selects only a few active nodes and most other nodes remain in idle state whereas the traditional approaches consider all nodes as active. Moreover, the proposed approach uses both cluster and tree-based approaches based on the type of data it senses and balances the energy consumption. Fig. 5 demonstrates that the data transmission latency of the proposed data aggregation scheme are less than that of the tree and cluster-based data aggregation approach because the CH receives data from a few active cluster member nodes in cluster-based approach and the parent node receives data from a few active child node, which require less time for the CH and a parent node to process and further transmit data to the next level.

From the result presented in Fig. 4 about the network energy consumption we can deduce that the network lifetime of the proposed scheme is expected to be more than those of cluster and tree-based approaches. Figure 6 demonstrates our claim that the network lifetime of the proposed hybrid data aggregation approach is much more than that in the traditional tree-based and cluster-based data aggregation approaches. We can further justify the presented results as follows:

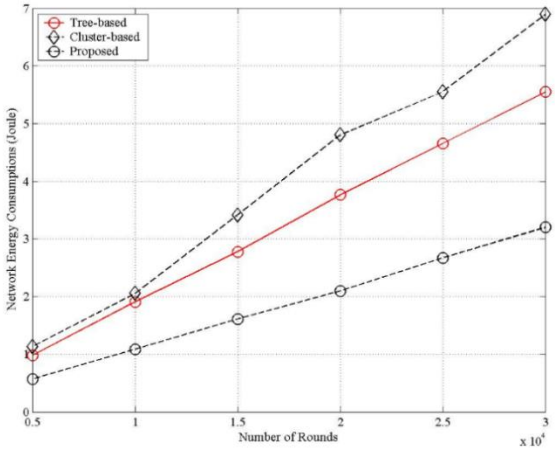

Fig. 4. Comparison of network energy consumption.

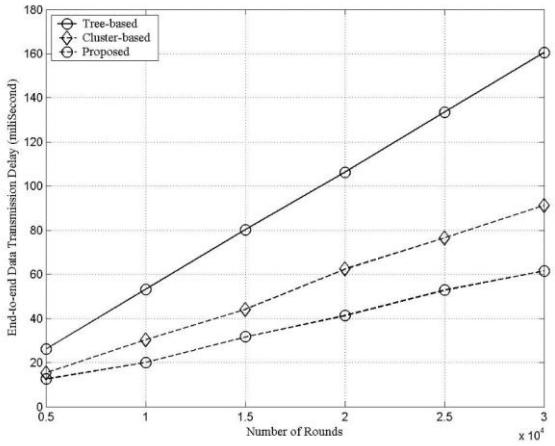

Fig. 5. Comparison of data transmission delay.

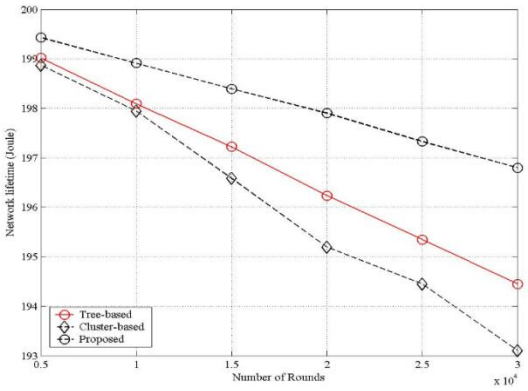

Fig. 6. Comparison of network lifetime.

In tree-based data aggregation schemes, upper level nodes wait until nodes at the lower levels transmit data to the upper levels. This results in higher data transmission latency. Moreover, a large number of active nodes at each level results in data redundancy, and data processing overhead. Cluster-based approach allows all cluster members to transmit data to the cluster head (CH). Thus, the CH requires much energy to process the received data. As some of the CHs might be far away from the sensor gateway, it consumes much energy of the CH to transmit the large aggregated data. In its own case, the proposed data aggregation scheme selects only a few active nodes that cover the whole network, this provides lower processing overhead and reduce the total network energy consumption (i.e., higher network lifetime). Processing and transmitting data from a fewer active nodes will also result in less data transmission latency. In summary, Table II compares the existing tree and cluster-based data aggregation approaches with the proposed hybrid approach based on different features.

VI. Conclusion and Future Works

We introduced a sensor-based big data aggregation approach in this paper. This approach works in multiple layers. However, we focus on aggregating redundant and unstructured sensors data at the lowest level of this framework at sensor nodes. The proposed hybrid data aggregation scheme uses either an efficient cluster-based data aggregation when data are transmitted from real-time or emergency sensor applications or a tree-based approach for non-real-time sensor applications. Experimental results demonstrate that the proposed hybrid and dynamic data aggregation scheme is better than traditional cluster and tree-based schemes in terms of network energy consumption, network lifetime and data transmission latency. This results in less amount of (unprocessed) data by big data server at upper layers to further faster data aggregation and filtering. In future, we plan to design and implement and efficient (computational) data aggregation scheme for upper layers at big data server. Also, we plan to implement the proposed approach in testbed (real experiments) and compare with more existing approaches to justify its effectiveness. Securing sensor data aggregation approaches against attacks, i.e., Sybil, wormhole, blackhole, bogus information, modification of sequence number through the use of public and private key cryptography and encryption mechanisms is significantly important even though those approaches require more computations. Hence, we plan to implement computation-efficient secure data aggregation approaches as part of our future research in this direction.

TABLE II. Comparison of Different Data Aggregation Methods

Features	Tree – based	Cluster-based	Proposed
Flooding interest propagation	√	X	X
Initially, sink receives data through multiple paths	√	X	X
All nodes in the network are active (i.e., they sense, send and transmit data)	√	√	X
A few active nodes cover the whole network area	X	X	√
Form clusters and send event of	X	√	√

interest to CH			
Dynamic data aggregation	X	X	√
Static data aggregation	√	√	X
Support fault tolerance	X	X	X
Tree structure with a single point of failure	√	X	X
Name of existing approaches	DD, FEDA, DABDR, TAG	CLUDDA SUMAC, OCABTR	PROPOSED HYBRID

REFERENCES

[1] Karim L., Nasser N. and Salti T., "Routing on Mini-Gabriel Graphs in Wireless Sensor Networks", *IEEE WiMob*, pp. 105-110, China, Oct 2011.

[2] Heinzelman W.B., "Application Specific Protocol Architectures for Wireless Networks", *PhD thesis*, Massachusetts Institute of Technology, June 2000.

[3] W. R. Heinzelman, A. Chandrakasan and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, 2000, pp. 10 pp. vol.2. doi: 10.1109/HICSS.2000.926982

[4] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann and F. Silva, "Directed diffusion for wireless sensor networking," in *IEEE/ACM Transactions on Networking*, vol. 11, no. 1, pp. 2-16, Feb 2003.

[5] R. Rajagopalan and P. K. Varshney, "Data-aggregation techniques in sensor networks: A survey," in *IEEE Communications Surveys & Tutorials*, vol. 8, no. 4, pp. 48-63, Fourth Quarter 2006.

[6] Takaishi, D., Nishiyama, H., Kato, N., Miura, R., "Toward Energy Efficient Big Data Gathering in Densely Distributed Sensor Networks," in *Emerging Topics in Computing, IEEE Transactions on* , vol.2, no.3, pp.388-397, Sept. 2014.

[7] Paulo Jesus, Carlos Baquero and Paulo Sergio Almeida, "A Survey of Distributed Data Aggregation Algorithms", *IEEE Communication Surveys and Tutorials*, Vol. 17, No. 1, First Quarter 2015.

[8] Yu Du, Fengye Hu, Lu Wang and Feng Wang, "Framework and challenges for Wireless body area networks based on big data," in *Digital Signal Processing* (*DSP*), *2015 IEEE International Conference on*, pp.497-501, 21-24 July 2015.

[9] Andreu-Perez, J.; Poon, C.C.Y.; Merrifield, R.D.; Wong, S.T.C.; Yang, G.-Z., "Big Data for Health," in *Biomedical and Health Informatics, IEEE Journal of*, vol.19, no.4, pp.1193-1208, July 2015.

[10] Chao Wu, Birch, D., Silva, D. Chun-Hsiang Lee, Tsinalis, O. and Guo, Y., "Concinnity: A Generic Platform for Big Sensor Data Applications," in *Cloud Computing, IEEE*, vol.1, no.2, pp.42-50, July 2014.

[11] Fu Xiao, Chongshen Zhang, and Zhijie Han, "Big Data in Ubiquitous Wireless Sensor Networks", *International Journal of Distributed Sensor Networks*, Volume 2014, Article ID 781729.

[12] Michael, K., Miller, K.W., "Big Data: New Opportunities and New Challenges", in *Computer*, vol.46, no.6, pp.22-24, June 2013.

[13] Hunter, P., "Journey to the centre of big data," in *Engineering & Technology*, vol.8, no.3, pp.56-59, April 2013.

[14] Fan Ye, Alvin Chen, Songwu Lu, Lixia Zhang, "A Scalable Solution to Minimum Cost Forwarding in Large Sensor Networks", *10th International Conference on Computer Communications and Networks* (*ICCCN2001*), Scottsdale, Arizona USA. October 15-17, 2001.

[15] Fan Ye, Haiyun Luo, Jerry Cheng, Songwu Lu, and Lixia Zhang, "A two-tier data dissemination model for large-scale wireless sensor networks", In *Proceedings of the 8th annual intl conference on Mobile computing and networking* (MobiCom '02). New York, USA, pp. 148-159, 2002

[16] S. Chatterjea and P. Havinga, "A Dynamic Data Aggregation Scheme for Wireless Sensor Networks". *In ProRISC 2003, 14th Workshop on Circuits, Systems and Signal Processing*, 26-27 November 2003, Netherlands.

[17] Ding, M.; Xiuzhen Cheng; Guoliang Xue, "Aggregation tree construction in sensor networks," in *Vehicular Technology Conference, 2003. VTC 2003-Fall. 2003 IEEE 58th*, vol. 4, pp. 2168-2172 6-9 Oct. 2003.

[18] Hüseyin Özgür Tan and Ibrahim Körpeoğlu, "Power efficient data gathering and aggregation in wireless sensor networks" *SIGMOD Rec.* 32, vol. 4, pp. 66-71, December 2003

[19] L. Karim, N. Nasser, T. Sheltami. A fault-tolerant energy efficient clustering protocol of a wireless sensor network. *Wireless Communications & Mobile Computing*, vol. 14, Issue 2, 2012, pp. 175-185.

[20] Ahmed A. Ahmed, Hongchi Shi, and Yi Shang, "Survey on Network Protocols for Wireless Sensor Networks," in *Proc. Intl. Conf. Information Technology: Research and Education*, 11-13 Aug. 2003.

[21] Mohammed S. Al-kahtani, "Efficient Cluster-Based Sleep Scheduling for M2M Communication Network", *Arabian Journal for Science and Engineering*, August 2015, Vol 40, Issue 8, pp. 2361-2373.

[22] Harichandan, P., Jaiswal, A., and Kumar, S., "Multiple Aggregator Multiple Chain routing protocol for heterogeneous wireless sensor networks," in *Signal Processing and Communication (ICSC), 2013 International Conference on*, vol., no., pp.127-131, 12-14 Dec. 2013.

[23] Karim L., Nasser N., Abdulsalam H. and Moukadem I., "An Efficient Data Aggregation Approach for Large Scale Wireless Sensor Networks", *GLOBECOM 2010*, pp. 1-6, Miami, USA.

[24] Karim, L., Mahmoud, Q. H., Nasser, N., and Khan, N, "An integrated framework for wireless sensor network management", *Wireless Communications and Mobile Computing*, 2014, *14*(12), 1143-1159.

[25] Lutful Karim and Mohammed S. Al-kahtani, "PDDA: Priority-based, Dynamic Data Aggregation Approach for Sensor-based Big Data Framework", *The 7th IEEE Annual Information Technology, Electronics and Mobile Communication Conference* (IEEE IEMCON), University of British Columbia, Vancouver, Canada, 13 - 15 October 2016, pp. 1-7.

[26] H. Harb, A. Makhoul, S. Tawbi and R. Couturier, "Comparison of Different Data Aggregation Techniques in Distributed Sensor Networks," in *IEEE Access*, vol. 5, pp. 4250-4263, 2017.

[27] L. Shen, J. Ma, X. Liu, F. Wei and M. Miao, "A Secure and Efficient ID-Based Aggregate Signature Scheme for Wireless Sensor Networks," in *IEEE Internet of Things Journal*, vol. 4, no. 2, pp. 546-554, April 2017. doi: 10.1109/JIOT.2016.2557487

[28] Q. Chen, H. Gao, S. Cheng, J. Li and Z. Cai, "Distributed non-structure based data aggregation for duty-cycle wireless sensor networks," *IEEE INFOCOM 2017*, Atlanta, GA, 2017, pp. 1-9.

[29] H. C. Lin and W. Y. Chen, "An Approximation Algorithm for the Maximum-Lifetime Data Aggregation Tree Problem in Wireless Sensor Networks," in *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3787-3798, June 2017. doi: 10.1109/TWC.2017.2688442

[30] R. Vinodha and S. Durairaj, "Data gathering cluster-based approach for in-network aggregation," *2016 Intl Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*, Pudukkottai, 2016, pp. 1-4.

[31] I. Vakilinia, J. Xin, M. Li and L. Guo, "Privacy-Preserving Data Aggregation over Incomplete Data for Crowdsensing," *2016 IEEE Global Communications Conference (GLOBECOM)*, Washington, DC, 2016, pp. 1-6.

[32] Jayashri B. S and G. R. Rao, "Reviewing the research paradigm of techniques used in data fusion in WSN," *2015 International Conference on Computing and Communications Technologies (ICCCT)*, Chennai, 2015, pp. 83-88. doi: 10.1109/ICCCT2.2015.7292724

[33] M. Yang, "Data aggregation algorithm for wireless sensor network based on time prediction," *2017 IEEE 3rd Information Tech and Mechatronics Engineering Conference (ITOEC)*, Chongqing, 2017, pp. 863-867.