Security of Big Data in Internet of Things

Rakesh Bandarupalli and H. Parveen Sultana

Abstract Presently 25 billions of devices are connected to the internet, and 50 billions of devices will connect to the Internet by 2020. These devices comprise of lots of sensors. These sensors, computers, tablets, and smart phones are generating twice as much as the data today as they generated two years ago. According to a survey, 90% of connected devices are collecting the information, and 70% of this data is transmitting without encryption. The rapid growth of the data brings tremendous changes in the humans' daily life. This data is providing new business opportunities. Many IoT devices are generating data related to the personal behavior. Lots of research is happening in big data security. This research is in the initial stage only. Up to right now, there is no specific method for providing security to the big data. Most of the data generated with IoT applications is unstructured data. Providing security to the unstructured data is more difficult than the providing security to the structured data. This chapter discusses various mechanisms for providing security to the big data generated by various IoT devices. This chapter describes the existing techniques for providing security at data generation, transmission and storage phases. The first methodology describes security to the data on Internet of Vehicles. This methodology uses a single sign-on algorithm. In this technique vehicle node and slink nodes need to register at a big data center for one time. This technique uses symmetric key cryptographic algorithms for encrypting the data. The second methodology describes providing the security with a dynamic prime number based security verification scheme. In this methodology, the prime numbers will be generated at regular intervals of time. The prime numbers will be generated at both source and receiver side. 128-bit symmetric key cryptography is used for this methodology. This paper also discusses about the advantages and disadvantages of these methodologies.

R. Bandarupalli
Sree Vidyanikethan Engineering College, Tirupathi, Andhra Pradesh, India
e-mail: bandarupallirakesh@gmail.com

H. Parveen Sultana (✉)
VIT University, Vellore, Tamil Nadu, India
e-mail: hparveensultana@vit.ac.in

N. Jeyanthi et al. (eds.), *Ubiquitous Computing and Computing Security of IoT*,
Studies in Big Data 47, https://doi.org/10.1007/978-3-030-01566-4_2

Keywords Internet of Vehicle (IoV) · Secure mechanism · Big data · Large-scale IoT · Big data analytics

1 Introduction

The computational capability is increasing drastically from the past decade. The development of networking is leading towards the rapid growth of the web technologies and data centers [1]. The development of Internet of things and big data is quickly accelerating and impacting all areas of technologies. This enhancement is benefiting many organizations as well as individuals [2]. The IoT devices are playing a vital role in this enhancement. The Big Data can be categorized depending upon three factors Velocity, Volume, and Variety. Gartner introduced these terms to describe the challenges of big data [3]. Massive opportunities are producing by analyzing this IoT data in the domain of smart cities, smart transportation, health care and much more.

The rapid growth of IoT devices triggers big data analytics as a challenging task. According to the estimation of IDC (International Data Corporation), the big data market will reach more than the US$125 billion by 2019. The big data analytics used to extract the useful information using various data mining techniques [4]. This information is useful in taking the business decisions as well as in revealing the recent trends.

The IoT data is different from the big data collected through the systems. As the IoT data is collected from various sensors, this data consists of a lot of noise, heterogeneity, and variety [1]. Various studies said that the sensors will increase to 1 trillion by 2030. This enhancement will cause the producing of huge amounts of data [3]. The area like smart traffic, smart grids, intelligent logistics management and intelligent buildings are the some of the applications of IoT and Big Data.

Big data is a term refers to large data sets. These data sets are complex in nature. The traditional data processing applications are not suitable for processing the data [2]. The big data consists of both structured as well as unstructured data. The data is generated from various sources like social networking sites, health care applications, and sensor networks and from many other organizations.

Internet of things is envisioned as the emerging trend. There is a lot of scope for research. This technology makes the human life comfortable. It shows solutions to the lot of problems related to logistics, transportation, urbanization, and environment. This technology enables to connect the physical world things and cyber world.

2 Introduction to Internet of Things and Big Data

2.1 Internet of Things

IoT is a platform where the devices are communicated with one another over the internet. These devices consist of various types of sensors. The IoT devices will share the information in a convenient manner. The IoT is termed as the next generation

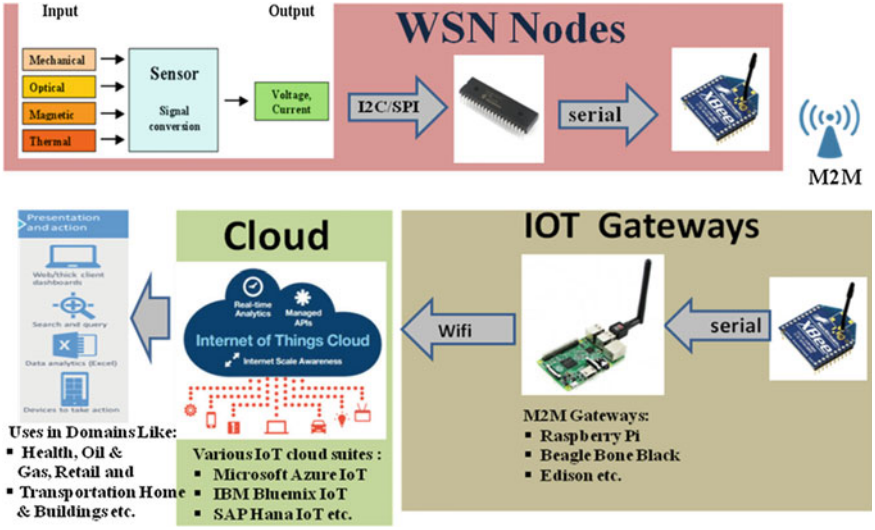


Fig. 1 Internet of Things big picture

revolution. IoT is adopted in various sectors such as smart cities, smart transportation, smart office, smart retail, smart energy and smart health care.

Figure 1 is representing the Internet of Things Big picture. It consists of various communication mechanisms used in IoT, Different IoT gateways, different storage mechanisms used for storing the data produced by IoT devices and also various applications that use Internet of Things.

The mobile devices, transportation vehicles, home appliances, health care devices etc. are used for data actuation [3]. Wrist watches, Doors, refrigerators, air conditioners and microwave Owens are the some of the IoT devices. These devices are deployed in various geographical locations [5]. These devices will acquire the real time data. These devices are connected with several communication mechanisms such as Bluetooth, Zigbee, WiFi etc. 50 billion devices such as laptops, smart phones, sensors will connect to the internet.

The graph in Fig. 2 is generated from the data provided by Cisco in the year of 2011. That graph is representing how the IoT devices are increasing rigorously in this decade.

Figure 3 is representing that by 2020 50 billion connected devices will be available for 7.6 billion people, and also the above figure is showing that the IoT is the combination of people, devices, and sensors interconnected with one another.

2.2 Big Data

The IoT devices and various other software applications will produce the data continuously. This data will consist of Structured, Unstructured and Semi structured data.

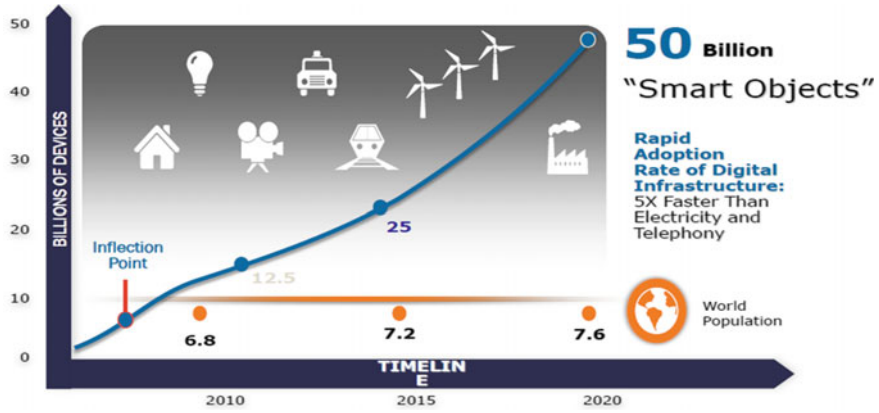


Fig. 2 Growing of Internet of Things

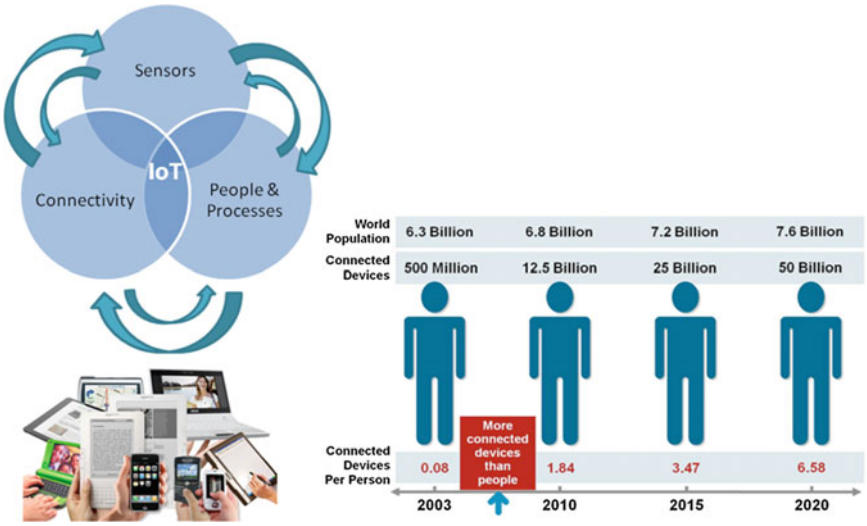


Fig. 3 Connected world

This huge amount of data is termed as "Big Data". The conventional data bases are not sufficient to store this huge amount of data [6]. In simple terms, the big data can be defined as the data that cannot be handled by the single system. The conventional data bases cannot be used for processing and analyzing this data that is growing rigorously.

Gartner proposed a model consisting of 3V's (Volume, Velocity, Variety). In other words Volume of the data producing, Variety of the data producing, Velocity or speed of the data producing [1]. Some investigations said that the volume is the main characteristic of the big data.

2.3 Big Data Analytics

The big data analytics examines large data sets consisting of a variety of data to provide useful business information, market trends. By analyzing this huge amount of data can help the organizations in getting the useful information. Big data analytics require tools and technologies that can transform structured, unstructured and semi structured data into more useful data [4–6]. The scientists can analyze large volumes of big data using the traditional tools.

2.4 Relationship Between Big Data Analytics and IoT

The big data analytics are used for decision making by analyzing the data produced by IoT devices continuously. The big data analytics are used to analyze the continuous data and store this large amount of data using various storage technologies. This large amount of data mostly consists of unstructured data [7]. Here the analytic tools need to analyze this data with lightning speed so that the business organizations can take the decisions immediately. Need of adopting big data in Internet of Things applications are increasing dramatically. Figure 2 is representing how the big data and IoT are interdependent on one another. As the usage of IoT devices is increasing the use of the big data will also increase proportionally [8–10]. The combinations of these two technologies are providing good business opportunities in the area of business and research.

Figure 4 is representing that how the big data analytics and Internet of Things are inter connected with one another. The Figure consisting of three phases, the first phase consists of IoT devices with sensors, these devices are interconnected with one another [11–14]. The second phase consists of different storage technologies. The data produced by IoT devices are stored on low-cost commodity hardware. This data can be called as big data, this data has mainly three properties i.e. volume, velocity, and variety. This data will be distributed among fault tolerant databases. The third phase is an analytical phase. In this phase, various tools will be used for analyzing the data such as MapReduce, Spark, Skytree, and Splunk. These tools require training data set. With the help of training data sets we use queries, then produce reports and result sets.

2.5 Architecture of IoT for Big Data Analytics

The architecture in Fig. 5 is representing the Architecture of Internet of thing for Big data analytics. This architecture is consisting of seven layers. The first layer consists of IoT devices which are having sensors. The second layer consists of communication devices such as Internet, Zigbee, WiFi, and Bluetooth etc.

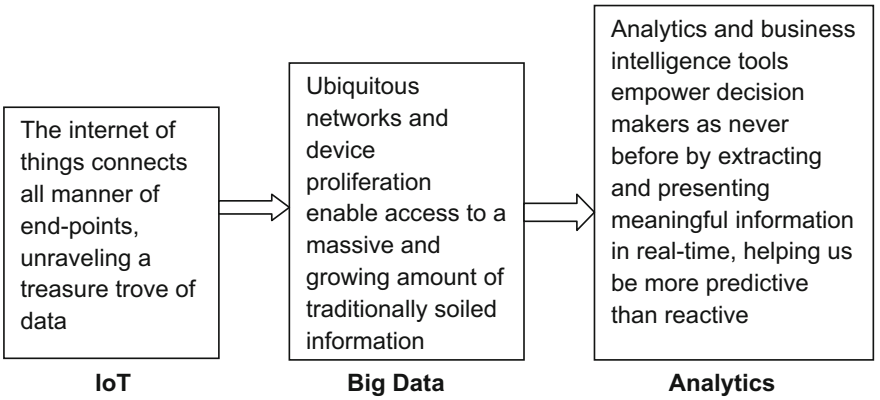


Fig. 4 Relationship between Internet of Things and big data analytic

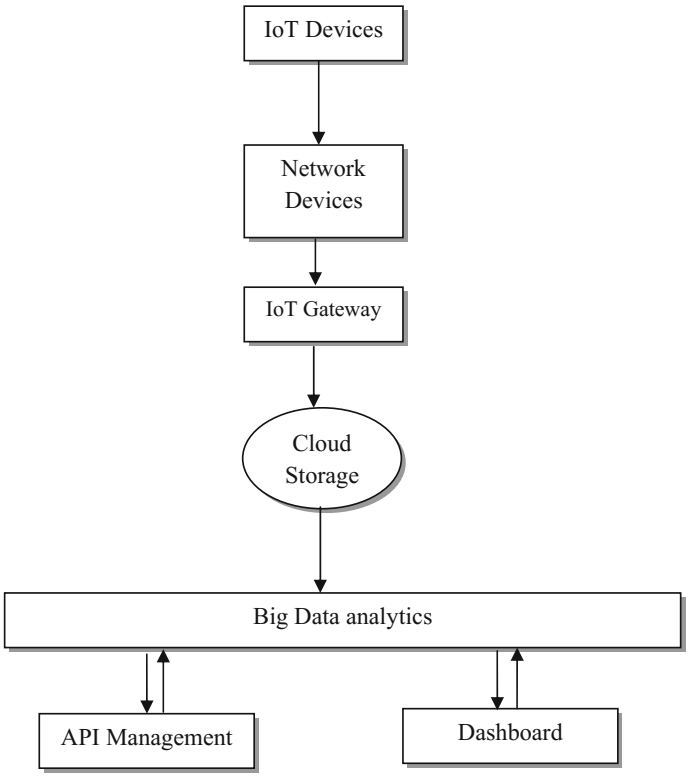


Fig. 5 Architecture IoT for big data analytics

The next layer is the cloud which is constructed with commodity hardware. The data generated by IoT devices will be stored in this cloud. The data will be received to the cloud through IoT gateway. The next phase consists of big data analytics phase. In this phase, a large amount of data will be processed which is stored in the commodity hardware. The major purpose of this architecture is to provide ample business solution.

3 Privacy of Big IoT Data

Sensitive information (Personal details) of users will be grabbed by the IoT devices. So, security is one of the major issues in the Big IoT data. These systems majorly depend on third party services. The traditional security solutions are not much effective in protecting this huge amount of data. The current existing security algorithms are majorly designed for providing security to the static data. The data produced from IoT devices are dynamic data. The data can be protected at generation phase, data storage phase, and data processing phase. Information privacy is protecting the information of a person or an individual from others. Security is protecting the data by using technology from recording, modifying, deleting.

3.1 Big Data Privacy at Data Generation Phase

The big data can be generated passively or actively. In active data generation, the data generated by the user will give to the third party. In passive data generation, the data will be generated by the user and the user will not have awareness about either the data is collected by the third party or not. The data can be protected at data generation phase either by access restriction of data or by falsifying the data.

a. **Access Restriction**

In most of the cases user not interested to share the sensitive information. If the user wants to share the data passively the user will take some precautions to secure the data by blocking the advertisements, blocking the scripts, and also by using some encryption techniques.

b. **Falsifying the Data**

In several cases it is very difficult to protect the sensitive information; in such cases, data falsification is used. In data falsification, the data will be distorted by using various tools. For example, while us using the credit card for online shopping Mask Me tool is used by most of the merchants.

3.2 Big Data Privacy at Data Storage Phase

The enhancement of big data technologies is leading to overcoming the storage problem. But if the big data storage system is compromised in security aspect, it will lead to a disclosure of the Users personal information. There are four categories in the traditional security mechanism. They are data security schemes at the file level, data base level, medium level and encryption scheme at the application level. The big data infrastructure should be scalable. By using storage virtualization we can accommodate more than one application dynamically. In this storage virtualization, more than one network storage devices are combined dynamically, so that we can assume that this is a single storage device. The data storage security and also computation auditing security can be provided with the help of SecCloud model.

3.3 Privacy Preservation Approaches for Cloud Storage

There are mainly three factors to be considered in storing the data securely n the cloud, i.e. integrity, confidentiality, and availability. The integrity and confidentiality are directly related to the security aspect of the data. The availability is representing the authorized persons can access the data whenever they required it. There are some basic methods to fulfill the security aspect of the data [15]. For example, the sender will encrypt the data with a public key and the receiver will decrypt the same data using a private key. The mechanisms for ensuring the privacy of the data are Attribute based encryption, Storage Path encryption, Homomorphic encryption and Using of Hybrid clouds.

3.4 Verification of Integrity of Data in Big Data Storage

When the data is stored in a third party cloud, the user will not have control over the data. So the data is at risk. In this scenario, the user needs to verify whether the data is stored in the cloud or not properly [16, 17]. This verification is called for checking the integrity of the data. To verify the integrity of the data there are several mechanisms provided. They are Message authentication code, Digital signatures, Checksums, trap-door hash functions, and Reed-Solomon code. We can also verify the integrity of the data available in the cloud by retrieving all the data stored in the cloud. The integrity verification is having the highest priority in security aspect.

3.5 Privacy Preserving of Big Data in Data Processing

Batch processing, machine learning, stream processing and graph processing are the big data processing paradigms [18, 19]. We can provide security to the data in two phases. In the first phase, the data should be protected from disclosing to the others. If the data is disclosed then the personal or sensitive information of the user will be at risk. In the second phase, the meaningful information needs to be extracted from the data without violating the privacy.

4 A Secure Mechanism for Big Data Collection on Internet of Vehicles

Internet of Vehicles is an extension of Internet of Things. The internet of vehicles is under smart transportation domain. On Internet of Vehicles, the vehicles get to connect with one another and also with the internet. This connection is leading producing of the data of different dimensionalities [20]. This data consists of the vehicles' location, a speed of the vehicle and the route in which the vehicle was traveled. This type of information will be collected by different sensors of the vehicles. The analysis of this information carries huge research interest. This research will be useful in traffic management [21]. This information may also consist of the users' personal information. If the security is not provided for this data then users' privacy will be at risk. There may be a chance that fraudulent information may be transmitted by the malicious vehicles to disturb the traffic system intentionally. So it is necessary to take the precautions to avoid the malicious vehicles.

Here in this security mechanism, first of all, the vehicles need to register at the big data centers to authenticate the vehicles. This authentication will be done by using the single sign-on algorithm. The basic architecture of the internet of vehicles is as represented in Fig. 6. This architecture consists of 4 blocks, i.e. Vehicle nodes, Road side units, Big Data centers and Storage module. This architecture also includes Satellite [22]. The vehicle nodes will communicate with one another and also with the Road side units. These road side units are also called as the SLINK nodes. These vehicles will be communicated with the help of the internet also, for that purpose the vehicles also interact with the satellites. The data generated with the help of vehicles will be collected by the SLINK nodes. This data will be transferred to the big data centers, where the collected data will be analyzed and again the information will be transmitted to the vehicle nodes. This information consists of in which route the traffic is huge and in which route the traffic is less [23]. The data will be finally stored in the storage module.

The diagrammatical representation of this scheme is shown in Fig. 7. As the no of vehicles increases, the data of different attributes will also increase. This data will be collected from different geographical locations. This data will be stored in the big data centers. These centers are distributed storage systems. These centers use Hadoop

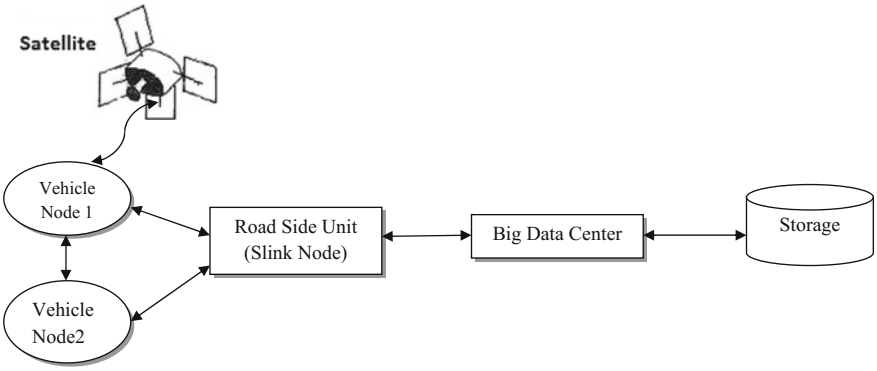


Fig. 6 Architecture of internet of vehicles

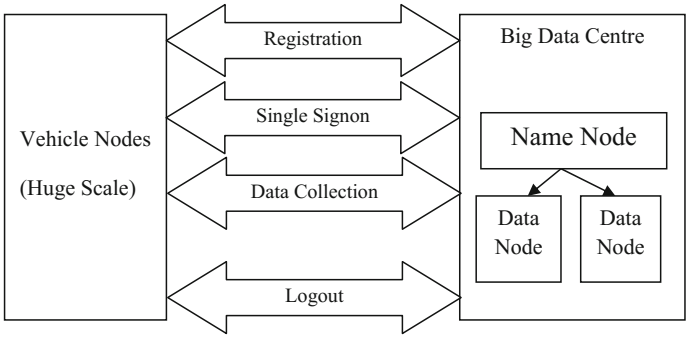


Fig. 7 Collecting data in a secure manner

Architectures. In the first phase authentication of all vehicle nodes will be done. Here the vehicle nodes will register at big data centers and required information will be exchanged with the big data centers. After registration phase, the vehicles will be log-on to the big data centers using a single sign-on algorithm. After that, the data will be exchanged continuously until up to the vehicles logouts from the system.

In this methodology, all the vehicles will register at the big data center for entering into the network. In the second phase authentication will be done by using a single sign-on algorithm [24]. In the next phase the collected information will be transferred securely and efficiently. In the final phase the collected information will be stored using distributed storage.

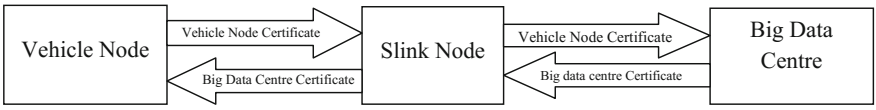


Fig. 8 Exchanging of messages at initialization phase

4.1 Initialization Phase

Here each and every vehicle is equipped with a certificate given by a third certification authority. In this phase, the vehicles need to register with the big data centers. The vehicles and big data centers generate a public key and private key among themselves. As shown in the Fig. 8 certificates and public keys will be exchanged in between the Vehicle node and big data center. The slink node acts as a mediator [25]. After that the certificates will be verified then the vehicle will get registered in the big data centers.

4.2 First Time Log-on

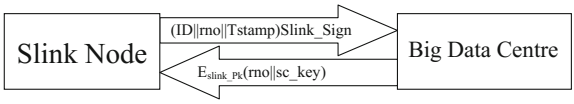
This section describes different procedures for slink node and vehicle login using single sign-on algorithm.

In the slink nodes' sign-on phase ID of the slink node, random number to fight against the replay attack, Message time stamp and slink nodes signature will be sent to the Big Data center. The big data center will check all these details. If these messages are from valid slink nodes then the big data centers will generate a session key. This session key is a unique key. The big data center will forward a packet consisting of a random number and unique session key (sc_key). This packet will be encrypted with the public key of the slink node. The slink node will decrypt the private key and acquires session key. Table 1 list out the symbols used in Fig. 9.

Table 1 Symbols notations

Symbol	Description
Rno	The random number to fight against replay attack
Tstamp	Message time stamp
Slink_Sign	Slink nodes signature
slink_Pk	Slink nodes public key
sc_key	Session key between big data center and slink node

Fig. 9 Logging on of slink nodes' for the first time



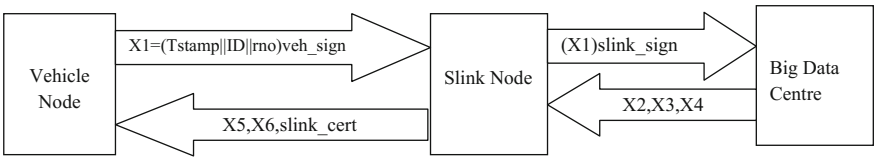


Fig. 10 Vehicle nodes' first-time log-on

In the Vehicle nodes' sign-on phase, the vehicle node sends the ticket to the slink node with its signature. The ticket has a time stamp of the message, vehicle node ID and random number generated by the slink node for fighting against the replay attack. Then the slink node sends the same ticket to the big data center [26]. This ticket consists of signature of the slink node. Big data center validates these details and three tickets will be passed to the slink node. The first ticket consists of Time stamp, ID of the big data center and a Random number generated by the big data center along with the big data center signature (See Fig. 10).

$$X2{:}(\text{Tstamp}\|\text{ID}\,\|\text{rno}\,)\text{cen\_sign}$$
$$X3{:}E_{\text{veh\_pk}}(\text{vc\_key})$$
$$X4{:}E_{\text{sc\_key}}(X2\|X3)$$
$$X5{:}E_{\text{vs\_key}}(\text{vs\_key})$$
$$X6{:}E_{\text{veh\_pk}}(X2\|X3)$$

The second ticket consists of session key between vehicle node and big data center. This session key is encrypted with vehicles public key [27]. The third ticket consists of both first and second tickets these tickets are encrypted with session key between slink node and big data center. The slink node generates a session key between vehicle node and slink node. The X2 and X3 are encrypted with this session key and this packet will be forwarded to the vehicle node. The session key also forwarded to the vehicle node by encrypting it with the public key of the vehicle. Table 2 has the description of various symbols used in this scheme.

4.3 Once Again Log-on

As the vehicle nodes are in the moving condition, the vehicle nodes need to log-on to the next arriving slink nodes by leaving the current log-on slink node. When the vehicle nodes want to access another slink node by leaving the first log-on slink node we need to follow the scenario discussed in this section [28]. Figure 11 is representing the communication between the slink node and vehicle node. In this communication process session key will be updated. The stored ticket X2, vehicle certificate will be forwarded to the Slink node with vehicle signature.

Table 2 Symbols notations of vehicle node

Symbol	Description
Tstamp	Time stamp
Rno	Random number for fighting against replay attack
Veh_sign	Vehicle nodes signature
Slink_sign	Slink nodes signature
Cen_sign	Big data centers signature
Veh_pk	Vehicle nodes public key
Vc_key	Session key between big data center and vehicle node
Sc_key	Session key between big data center and slink node
Vs_key	Session key between slink node and vehicle node
Veh_pk	Vehicle nodes public key

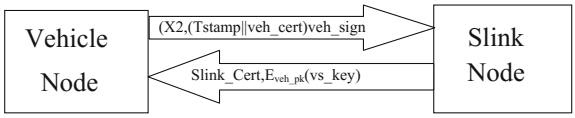


Fig. 11 Later log-on of vehicle node

The ticket X2 consists of the big data center signature. This signature shows that the ticket is issued by the big data center. After that, the session key (vs_key) of the slink node will be encrypted with the vehicle nodes public key. This session key (vs_key) will be forwarded along with the slink node certificate (Slink_cert).

4.4 Collecting Data Securely

In the previous scenarios the secure connection will be established between the Vehicle node, Slink node and big data centers. The data will be divided into two categories, business data, and confidential data. The business data will be exchanged in the plain text format and confidential data will be exchanged securely [29]. The business data consists of the information like temperature. The X4 can be calculated by concatenating vehicle nodes' Id with business message M1. The hash value of M4 is utilized for calculating HMAC. HMAC helps in stop the tampering of data. So that, the data will be sent to the receiver without any loss. The same scheme will be used to transfer the data from a big data center to the Vehicle node. Here M2 is the business message to be sent from the big data center to the vehicle node (See Fig. 12).

But the confidential data need to be exchanged securely. So here we are encrypting the confidential data and converting into the cipher text format. Here a random key, Z is utilized for encrypting. For sharing the random key Z with the slink node and big data center, vc_key and vs_key will be used (See Fig. 13).

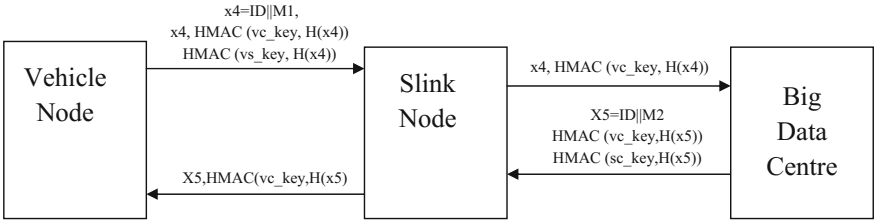


Fig. 12 Exchanging of business messages for big data collection

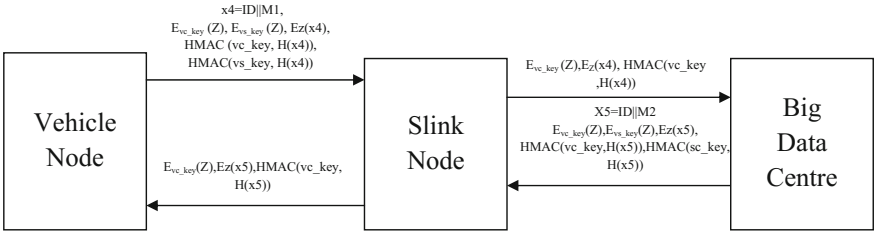


Fig. 13 Exchanging of Confidential messages for big data collection

4.5 Data Storage Security

In the previous sections we have discussed the secure connection establishment and secure data collection [30]. This section describes the storing of important data in big data center securely. All the information related to the vehicle need not be stored in the big data center. Some important information needs to be stored in the big data center securely [31]. Table 3 is the data structure of the information need to be stored in the big data center. It has the data structures related to both slink node and vehicle node. The first field of the data structure is ID, the second field is the certificate. This ID and certificate fields are used for identification purpose. The third field is statue field. This field consists of two values "on" and "off". If any abnormal situation occurs, the status field will change from on to off. The next field is the time stamp period field. If the time stamp period expires the vehicle node and slink nodes need to register once again as the new nodes. The session key and public key are important in providing confidentiality to the data.

The business information will be stored as a plain text in the big data center [32]. While storing the confidential information, the data need to be encrypted with the

Table 3 Data structure of big data center for storing slink node and vehicle node data

Nodes	ID	Certificate	Statue	Validity period	Encrypted session key
Slink node	Slink_ID	Slink_cert	Off/on	Period_TStamp	Sc_key
Vehicle node	Veh_ID	Veh_cert	Off/on	Period_Tstamp	Vc_key

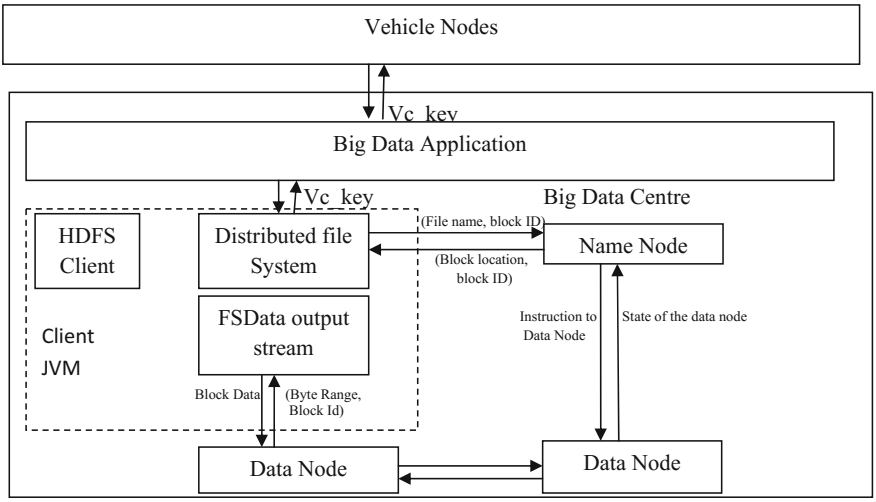


Fig. 14 Distributed storage system for big data collection

session key between vehicle node and big data center (vc_key). When the vehicle node itself interacts with the big data center, the data will be retrieved [33]. Otherwise, the data cannot be retrieved from the big data center. When the vehicle node interacts with the big data center the data will be decrypted with the session key between vehicle node and big data center (See Fig. 14).

As the no of vehicles is increasing day to day, the data collected from the vehicles also increasing rapidly. The Hadoop Distributed File System (HDFS) is a famous system for storing the big data. In this HDFS system we will have one name node and remaining all are data nodes. In the HDFS system the data will be replicated to more than one location to avoid the fault tolerance. Whenever a vehicle node wants to access the data, the vehicle node interacts with the big data center, the client JVM request for the file name and block ID through the distributed file system [34]. This distributed file system interacts with the Name node. The name node will acknowledge the block location and block id to the distributed file system [35]. Finally, FS Data output stream sends the Block ID and byte range to the data node to acquire the data [36]. If the acquired data is business data then the data will be sent to the vehicle node as a plain text. Otherwise, the data will be sent as the cipher text. The cipher text needs to be decrypted by using the session key between the vehicle node and big data center (vc_key).

5 Providing Security to Big Sensing Data Streams Using Dynamic Prime Number Based Security Verification

Real time data processing schemes require in many applications such as social networking applications like Facebook and Twitter, large scale sensors, web exploring, financial data and surveillance data analysis [37]. Stream processing engines are introduced with an aim to process the sensing data streams with the small delay. These engines are used to process the data in real time rather than processing after storing the data. But these engines are not suitable for processing the big stream data [38]. These large quantities of data contain various data, i.e. both structured and unstructured data. As these big data streams are continuous in nature, this data needs to be processed in real time [39]. The velocity and volume of this data are huge, we cannot store this data. So, the conventional computing models are not suitable.

5.1 Security Verification of Data Streams

The big sensing data streams are used in some critical applications such as military, these data need to be secured. The sensors are having the low processing power, less power, low storage and also very less energy [40]. These data streams need to be processed during the transmission phase itself. Here providing the security to the data is a very important aspect. For providing security to the data cryptographic model is used. There are two cryptographic models such as Asymmetric and Symmetric [41]. The asymmetric algorithms are much slower when compared to the symmetric cryptographic algorithms. But these symmetric cryptographic algorithms are failed in many cases when providing the security to the streamed data.

As the symmetric key cryptographic algorithms are failed in many cases of big data streaming, the dynamic prime number based security verification scheme will address those challenges. In this scheme, the key will be generated with the help of prime numbers synchronously. This key is generated at regular intervals of the time [42]. This prime number generation will be done at both sensing device side as well as at data stream manager side. As the key is generated at both source and DSM sides, it reduces communication overhead [43]. Here the key is of 64-bit size. This smaller key helps in faster processing of the streamed data by not compromising the security. The key is updated dynamically at both source and DSM side.

5.2 Architecture of Secure Data Stream

a. **Data Stream Processing**

The data stream processing is a revolutionary area. Many applications are using this data stream processing. In data stream processing huge amounts of data need to

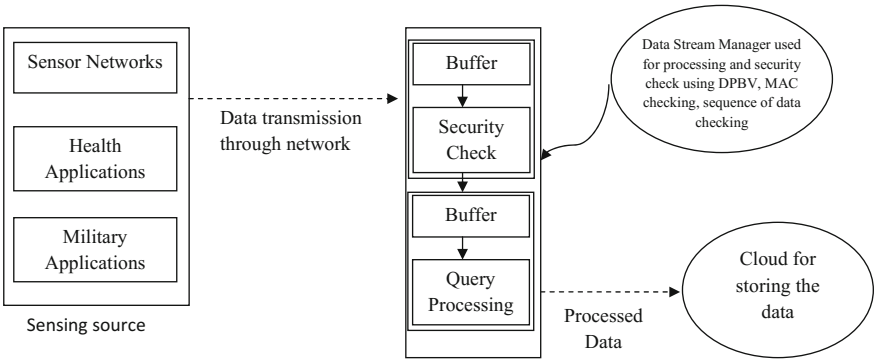


Fig. 15 Architecture of data stream

analyze with a small delay. In conventional mechanisms data is analyzed after storing it [44]. Here the data will be generated from various sources. It is very difficult to handle the data generated from various sources at a time [45]. And also in DSM, the data blocks need to go under the security verification.

The Fig. 15 is representing the architecture of secure data stream. In this architecture, the data stream flows from various sensor devices to the cloud. Here the architecture mainly focuses on three aspects, collecting the data, processing the data and storing the data. The security and query related processes are done in DSM (data stream manager). In this architecture first of all security verification will be done after that query processing will be done [46]. Small buffers will be maintained for both activities. In the final stage the data that is processed will be stored in the cloud [47]. Here the queries used for processing the data are continuous in nature, as the data is flowing continuously.

5.3 Purpose of Symmetric Key Cryptography

The size of symmetric keys is much smaller in size when compared with the asymmetric keys, so they require less computation power. A 128-bit symmetric key provides the equal strength compared with the 3248-bit asymmetric key. The main aim of big sensing data streams is to provide security to the data streams in real time. So, the symmetric key cryptography is the best choice in this scenario. The symmetric key cryptography is 1000 times faster than other public key cryptographic algorithms. As the size of the symmetric key cryptography is much smaller, the attacker can easily attack the data which is encrypted with symmetric key cryptography. To overcome this disadvantage, the keys are generated synchronously with dynamic number based algorithm [48]. Here the keys will be generated at both sensing devices end as well as at Data Stream Manager at regular intervals of time. The Fig. 16 will demonstrate about this key generation scheme.

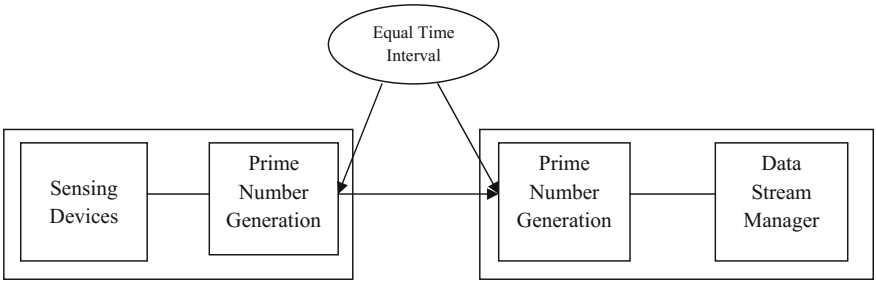


Fig. 16 Relative dynamic prime number generation

5.4 Setup of DPBSV System

Here the system is completely untrusted. The Data stream Manager (DSM) should maintain the entire sensor ID's and also secret keys (See Table 4 and Fig. 17).

Here in this process, first of all, sensors ID and a pseudo random number will be sent to the DSM. The DSM receives these details from Sensor. After that, the DSM retrieves secret key (key_s), with the help of Retrieve key function. After that, the session key (key_{si}) will be generated with the help of random key function. This session key will be combined with the secret key. This combination will generate a key (Key_{enc}) used for authentication purpose. The generated key and session key will be encrypted with the shared key (key). The hash value 1(H) will be computed with the help of hash function. The computed hash value and sensors private key will be passed to the sensor. The below given are the steps for computing hash value1.

Table 4 Symbols notations and descriptions

Symbol	Description
SID	Sensor's ID
Rno	Pseudo random number
Key_{enc}	Key generated for authentication
$H/H'/H''$	Computed hash value
Enc()	Function used for encryption
Key	Shared key initially used for authentication of DSM and sensor
Key_d	DSM secret key
It	Prime number generation interval time
RP	Prime number generated randomly
PF(RP)	Function used for generating the prime number randomly
Key_{sh}	DSM and sensor computed secret key
Key_s	Secret key of the sensor
M/M'	User authentication key which is encrypted with sensors' secret key

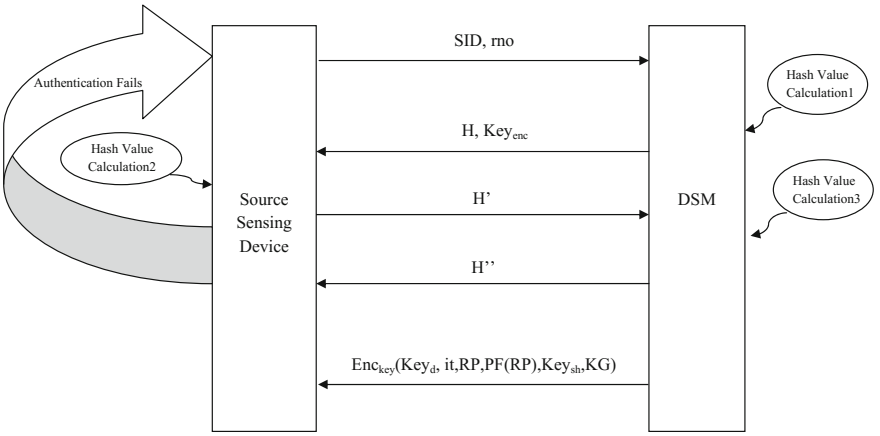


Fig. 17 Secure authentication procedure between DSM and source sensing device

$$\begin{aligned} &\text{Key}_s \text{ <- Retrieve (SID)},\\ &\text{Key}_{si} \text{ <- random ()},\\ &\text{Key}_{enc} \text{ <- Key}_s \oplus \text{Key}_{si}\\ &\text{M} = \text{Enc}_{key}\left(\text{Key}_{si,}\text{Key}_{enc}\right)\\ &\text{H} = \text{Hash}\left(\text{Key}_{enc}\|\text{M}\|\text{rno}\right) \end{aligned} \tag{1}$$

After that, the sensor will get the hash value (H) and key used for authentication purpose (key_{enc}). The DSM will get these details and finds its own secret key based on the authentication key. The sensors secret key and authentication key (Key_{enc}) will be encrypted for users' authentication. The hash value will get with the equation Hash ($\text{Key}_{enc}\|\text{M}'\|\text{rno}$). And validates whether the hash value generated by it and the hash value generated by DSM are equal or not. If $\text{M} = \text{M}'$ and the hash values are equal, then the authentication of DSM is successful by the sensor. If the authentication process is failed then again the process will begin from step 1.

$$\begin{aligned} &\text{key}_{si} = \text{Key}_{enc} \oplus \text{Key}_s\\ &\text{M}' = \text{Enc}_{key}\left(\text{key}_{si}, \text{key}_{enc}\right)\\ &\text{Hash}\left(\text{key}_{enc}\|\text{M}'\|\text{rno}\right)\\ &\text{M} = \text{M}', \text{ for authentication of DSM}\\ &\text{H}' = \text{Hash}\left(1\|\text{Key}_{enc}\|\text{M}'\|\text{rno}\right) \end{aligned} \tag{2}$$

After that the H' will be forwarded to the DSM, the DSM compares the received value with the Hash($1\|\text{Key}_{enc}\|\text{M}\|\text{rno}$), If these two are equal then the sensor is authenticated successfully. If the authentication is failed then the protocol will be

terminated. In this way, both the sensor and DSM authenticate each other. After successful validation, the DSM sends another hash value to fulfill the protocol. The hash value H″ will be calculated as shown below.

$$H'' = \text{Hash}(2 \| \text{Key}_{\text{enc}} \| M \| \text{rno})$$

5.5 Handshaking of DPBSV

During the calculation of prime numbers, we need to take care of communication overhead. The communication overhead must be reduced. The PF(RP) function used to generate the prime numbers randomly at both sides. These prime numbers have to be generated at regular intervals of the time. The DSM transmits the algorithms related to generating of the prime number and keys like (Key_d,it,RP,PF(RP),KeyGen, Key_{sh}) to each and every individual sensor by encrypting them with the shared key generated initially. This transferred information will be stored in the trusted part of the sensor.

After successful completion of handshaking process, the data needs to be transmitted securely. This secure transmission and verification can be done by using several functions and keys. As discussed earlier this scheme utilizes dynamic prime number generation process. This dynamic prime number generation can be done at both sensor and DSM side. Each and every sensor will have its own key. Initially shared key and prime numbers will be generated by the DSM itself. Next prime number will be generated depending upon the current prime number and the interval time. The shared key will be generated by the sensors depending upon the formula $\text{Key}_{sh} = \text{Hash}(\text{Enc}(RP,\text{key}_d))$. Here, each and every data block consists of two parts. The first part consists of the encrypted data. This data will be encrypted with the help of secret key Key_i and shared key $\text{key}_{sh.}$ These three things will be mutually exclusively ORED. DATA $\oplus$ Key_i $\oplus$ $\text{key}_{sh.}$ This encryption is mainly used for integrity checking. The second part is used for authentication checking. $S_i \oplus \text{key}_{sh.}$ So finally the resultant block is

$$(\text{DATA} \oplus \text{Key}_i \oplus \text{key}_{sh}) \| (S_i \oplus \text{key}_{sh.})$$

Lets $I_d = \text{DATA} \oplus \text{Key}_i \oplus \text{key}_{sh}$

$A_d = S_i \oplus \text{key}_{sh}$

In the next step the sensor will send the encrypted format of the above data block

$$\text{Enc}_k(A_d \| I_d)$$

5.6 Security Verification of DPBSV

The security verification should be done in real time. The main aim of the security verification is to provide the end to end security. This security verification will be done at DSM side. In the security verification the DSM verifies whether the data is modified or not. And also it verifies whether the data is from the authenticated node or not. First of all the DSM will decrypt the data block to check the integrity and authenticity. First of all the DSM authenticates each and every block. And the integrity will be checked at the arbitrary interval blocks. The interval may vary from 0 to 6. i.e. the interval blocks may be 6 at most or 0 at least.

The Fig. 18 representing the updating of shared and also the security verification of the data. The updating of the shared key will be done at both sources sensing device side as well as at Data stream manager side. But the security verification will be done only at DSM side.

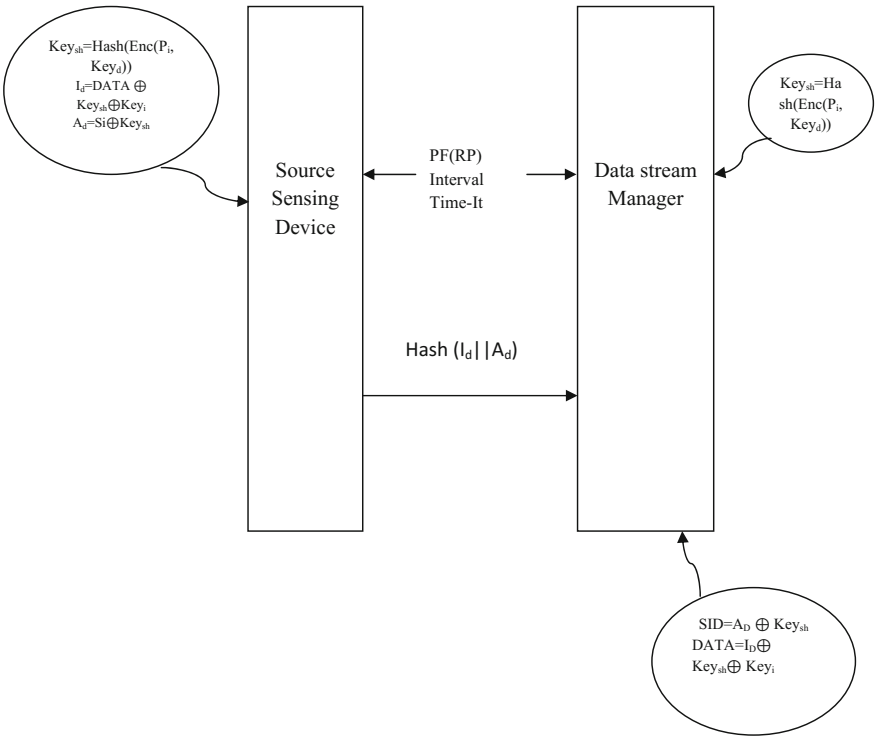


Fig. 18 Updating of shared key and verification of security

6 Conclusion

In this chapter, we discussed two security algorithms. The first algorithm is used to provide the security to the vehicular data. In this methodology, the vehicular nodes and slink nodes need to register at big data center. In this methodology, single sign-on algorithm was used for login to the big data center. Symmetric key cryptography was used in this methodology. The second algorithm is for providing the security to the Sensor data. Here the dynamic prime number based security scheme was used for proving security to the big data. This prime number generation will be done at both sources as well as at big data center side.

References

1. Guo L, Dong M, Ota K, Li Q, Ye T, Wu J, Li J (2016) A secure mechanism for big data collection in large scale internet of vehicle. IEEE Internet Things J, https://doi.org/10.1109/jiot.2017.2686451
2. Guerrero-ibanez JA, Zeadally S, Castillo JC (2015) Integration challenges of intelligent transportation systems with connected vehicle, cloud computing, and internet of things technologies. IEEE Wirel Commun 22(6):122–128
3. Puthal D, Nepal S, Ranjan R, Chen J (2017) A dynamic prime number based efficient security mechanism for big sensing data streams. J Comput Syst Sci, Science Direct 22–42
4. Walravens C, Dehaene W (2012) Design of a low-energy data processing architecture for WSN nodes. In: Proceedings of the conference on design, automation and test in Europe, March 2012, pp 570–573
5. Walravens C, Dehaene W (2014) Low-power digital signal processor architecture for wireless sensor nodes. IEEE Trans Very Large Scale Integr (VLSI) Syst 22(2):313–321
6. Kaddoura I, Abdul-Nabi S (2012) On formula to compute primes and the nth prime. Appl Math Sci 6(76):3751–3757
7. Perrig A, Szewczyk R, Tygar J, Wen V, Culler DE (2001) SPINS: security protocols for sensor networks. In: Proceedings of ACM MobiCom'01, 2001, pp 189–199
8. Akyildiz I, Su W, Sankarasubramaniam Y, Cayirci E (2002) Wireless sensor networks: a survey. Comput Netw 38(4):393–422
9. Hempstead M, Lyons MJ, Brooks D, Wei G (2008) Survey of hardware systems for wireless sensor networks. J Low Power Electron 4(1):11–20
10. Burke J, McDonald J, Austin T (2000) Architectural support for fast symmetric-key cryptography. ACM SIGOPS Oper Syst Rev 34(5):178–189
11. Puthal D (2012) Secure data collection and critical data transmission technique in mobile sink wireless sensor networks. M. Tech thesis, National Institute of Technology, Rourkela
12. TCG Trusted Platform Module (TPM) specification, https://www.trustedcomputinggroup.org/specs/tpm/. Accessed on 04 Aug 2014
13. Nepal S, Zic J, Liu D, Jang J (2011) A mobile and portable trusted computing platform. EURASIP J Wirel Commun Netw 2011(1):1–19
14. Gulisano V, Jimenez-Peris R, Patino-Martinez M, Soriente C, Valduriez P (2012) Stream cloud: an elastic and scalable data streaming system. IEEE Trans Parallel Distrib Syst 23(12):2351–2365
15. Guo L, Wu J, Xia Z, Li J (2015) Proposed security mechanism for XMPP-based communications of ISO/IEC/IEEE 21451 sensor networks. IEEE Sens J 15(5):2577–2586
16. Aalm KM, Saini M, Saddik AE (2015) Toward social internet of vehicles concept, architecture, and applications. IEEE Access 3:343–357

17. Cecchinel C, Jimenez M, Mosser S, Riveill M (2014) An architecture to support the collection of big data in the internet of things. In: Proceedings of the IEEE 10th World Congress on services, Anchorage, June 2014, pp 442–449
18. Su Z, Xu Q, Qi Q (2016) Big data in mobile social networks: a QoE-oriented framework. IEEE Netw 30(1):52–57
19. Guo L, Dong M, Ota K, Jun W, Li J (2015) Event-oriented dynamic security service for demand response in smart grid employing mobile networks. China Commun 12(12):63–75
20. Tracey D, Sreenan C (2013) A holistic architecture for the internet of things, sensing services and big data. In: Proceedings of the 13th IEEE/ACM international symposium on cluster, cloud, and grid computing, Delft, May 2013, pp 546–553
21. Salahuddin MA, Al-Fuqaha A, Guizani M (2015) Software-defined networking for RSU clouds in support of the internet of vehicles. IEEE Internet Things J 2(2):133–144
22. Li H, Lu R, Zhou L, Yang B, (Sherman) Shen X (2014) An efficient Merkle tree based authentication scheme for smart grid. IEEE Syst J 8(2):655–663
23. Li H, Lin X, Yang H, Liang X, Lu R, (Sherman) Shen X (2014) EPPDR: an efficient privacy-preserving demand response scheme with adaptive key evolution in smart grid. IEEE Trans Dependable Secure Comput 25(8):2053–2064
24. Li H, Yang Y, Luan TH, Liang X, Zhou L, (Sherman) Shen X (2015) Enabling fine-grained multi-keyword search supporting classified sub-dictionaries over encrypted cloud data. IEEE Trans Dependable Secure Comput
25. Liu C, Zhang X, Liu Z, Yang Y, Ranjan R, Georgakopoulos D, Chen J (2013) An iterative hierarchical key exchange scheme for secure scheduling of big data applications in cloud computing. In: Proceedings of the 12th IEEE international conference on trust, security and privacy in computing and communications, Melbourne, July 2013, pp 10–16
26. Adluru P, Datla SS, Zhang X (2015) Hadoop eco system for big data security and privacy. In Proceedings of the 2015 IEEE Long Island Systems, Applications and Technology Conference (LISAT), Farmingdale, May 2015, pp 1–6
27. Jam MR, Khanli LM, Akbari MK, Javan MS (2014) A survey on security of Hadoop. In: Proceedings of the 4th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Oct 2014, pp 716–721
28. Xu L, Jiang C, Wang J, Yuan J, Ren Y (2014) Information security in big data privacy and data mining. IEEE Access 2:1149–1176
29. Wang H, Qin B, Wu Q, Xu L, Ferrer JD (2015) TPP: traceable privacy-preserving communication and precise reward for vehicle-to-grid networks in smart grids. IEEE Trans Inf Forensics Secur 10(11):2340–2351
30. Soares J, Borges N, Canizes B, Vale Z (2015) Probabilistic estimation of the state of electric vehicles for smart grid applications in big data context. In: 2015 IEEE Power & Energy Society General Meeting, Denver, July 2015, pp 1–5
31. Mershad K, Artail H (2013) A framework for secure and efficient data acquisition in vehicular ad hoc networks. IEEE Trans Veh Technol 62(2):536–551
32. Gulisano V, Jimenez-Peris R, Patino-Martinez M, Valduriez P (2010) Streamcloud: a large scale data streaming system. In: Proceedings of 30th International Conference on Distributed Computing Systems, ICDCS, 2010, pp 126–137
33. Arasu A, Babcock B, Babu S, Datar M, Ito K, Nishizawa I, Rosenstein J, Widom J (2003) STREAM: the stanford stream data manager (demonstration description). In: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, 2003, p 665
34. Carney D, Çetintemel U, Cherniack M, Convey C, Lee S, Seidman G, Stonebraker M, Tatbul N, Zdonik S (2002) Monitoring streams: a new class of data management applications. In: Proceedings of the international conference on very large data bases, 2002, pp 215–226
35. Abadi DJ, Carney D, Çetintemel U, Cherniack M, Convey C, Lee S, Stonebraker M, Tatbul N, Zdonik S (2003) Aurora: a new model and architecture for data stream management. VLDB J 12(2):120–139
36. Chandrasekaran S, Cooper O, Deshpande A, Franklin M, Hellerstein JM, Hong W, Krishnamurthy S, Madden SR, Reiss F, Shah MA (2003) TelegraphCQ: continuous dataflow processing.

In: Proceedings of ACM SIGMOD International Conference on Management of Data, 2003, p 668
37. Tatbul N, Çetintemel U, Zdonik SB (2007) Staying fit: efficient load shedding techniques for distributed stream processing. In: Proceedings of international conference on Very Large Data Bases, VLDB, 2007, pp 159–170
38. Jin M, Zhou X, Luo E, Qing X (2015) Industrial-QoS-oriented remote wireless communication protocol for the internet of construction vehicles. IEEE Trans Industr Electron 62(11):7103–7113
39. Puthal D, Sahoo B (2012) Secure data collection & critical data transmission in mobile sink WSN. In: Secure and energy efficient data collection technique. LAP Lambert Academic Pubilishing, Germany, ISBN978-3-659-16846-8
40. Kumar N, Rodriguesand JJPC, Chilamkurti N (2014) Bayesian coalition game as-a-service for content distribution in internet of vehicles. IEEE Internet Things J 1(6):554–555
41. Fu J, Chen Z, Sun R, Yang B (2014) Reservation based optimal parking lot recommendation model in internet of vehicle environment. China Commun 11(6):38–48
42. Cheng J, Cheng J, Zhou M, Liu F, Gao S, Liu C (2015) Routing in internet of vehicles a review. IEEE Trans Intell Transp Syst 16(5):2339–2351
43. Dua A, Kumar N, Bawa S (2014) A systematic review on routing protocols for vehicular ad hoc networks. Veh Commun 1(1):33–52
44. Li B, Zhao C, Zhang H, Sun X (2013) Characterization on clustered propagations of UWB sensors in vehicle cabin: measurement, modeling and evaluation. IEEE Sens J 13(4):1288–1300
45. Kumar N, Misra S, Rodrigues J, Obaidat MS (2015) Coalition games for spatio-temporal big data in internet of vehicles environment: a comparative analysis. IEEE Internet Things J 2(4):310–320
46. Zhou Y, Chen S, Zhou Y, Chen M (2015) Privacy-preserving multi-point traffic volume measurement through vehicle-to-infrastructure communications. IEEE Trans Veh Technol 64(12):5619–5630
47. Wu Q, Ferrer JD, Nicolas ÚG (2010) Balanced trustworthiness, safety, and privacy in vehicle-to-vehicle communications. IEEE Trans Veh Technol 59(2):559–573
48. Li H, Liu D, Dai Y, Luan TH (2015) Engineering searchable encryption of mobile cloud networks: when QoE meets QoP. IEEE Wirel Commun 22(4):74–80
49. Cárdenas AA, Manadhata PK, Rajan SP (2013) Big data analytics for security. IEEE Secur Priv 11(6):74–76