CHAPTER 2 - LITERATURE REVIEW AND GENERAL CONSIDERATION

2.1 Introduction
The recent technological advancements and communication have caused the significant shift in sensor network research. Supporting real time communication is a major challenge due to limited energy, low node reliability, distributed architecture and dynamic network topology. Various kinds of query optimization method for reducing the redundant data and reducing the network energy become the focus of researchers. In WSN, queries are executing continuously and reaching the base station. There are different ways to categorize the types of queries, such as long-running continuous queries and one-shot queries, aggregate queries and nonaggregate queries, complex queries and simple queries, etc. There is no single solution for all type of queries and each and every one has to be treated separately.

2.2 Literature Review of Query Processing
Authors [12] considered sensor network as a database with respect to architectural view. This allows the user to fire the queries to one or more nodes. These queries can be one shot or continuous queries that produce huge volume of data. They suggested two ideas for implementing the database. First idea is in-network implementation in which grouping, aggregation and joins are greater importance. These are application dependent. Second, they discussed for relaxing semantics of queries for allowing the result. For energy-efficient implementations, query processing operators should be implemented within the sensor network.

In this paper [13], model for sensor database is defined and stored data are represented as relations. Long running queries are evolved over a sensor database for defining persistent view. This view is maintained during a given interval. They introduced the concept of sensor data. A sensor database consists of stored and sensor data. Sensor data are generated by signal processing
functions and are represented by the formulation of queries. They represented sensor data as sensor time series which is based on the sequence model.

Data storage can be divided into centralized, local, distributed storage and sensor network database. The wireless sensor network is considered as a distributed database [6] [2] and node is regarded as a perception data flow or data source. Sensor model had been transferred to virtual database table [6] and table includes various types of properties, references associated with the executable order sets on a query executor. In the query processing, put the data from sensor on each node into a tuple, these tuples can be transferred between the nodes of many routing, or gather data, or express code to server-side from the serial of network tip.

Information techniques are classified as event driven, time driven, query based and hybrid data acquisition [14] . Wireless sensor network produces enormous amount of data. Extracting information includes the parameters such as timeliness, accuracy, cost and reliability of the information. This process extracts the unstructured data from the nodes and filtered those data by using the sophisticated methods. Event driven is based on an event occurs, while data is periodically sent to the base station in the time driven. [14]. When the user is requesting for the data, query based approach is used and hybrid is the combination of all these approaches.

Cost, accuracy, Reliability of extracted information is solely based on nodes organization. They can be centralized or hierarchical. In the event driven approach, nodes are pre-configured with the threshold values. When the sensor node value is deviating from this threshold that is the indication of event and triggers the node to send the data to the sink.

Time driven is based on probabilistic model. In this model, sensor node periodically senses the environment and sends the data to the base station. Depends on the application requirements, time period can be pre-configured or decide by the end user.
Query based approach is following request and response model. End user posts the query at the base station and the query disseminated into the network for collecting the data. Based on the descriptions of the query, desired data can be collected from the network. Hybrid approach combines the algorithms of all the above approaches.

To reduce the communication cost of multiple query processing in WSN, authors [15] described the similarity among queries. This scheme constructs a set of shared immediate views for a given set of queries. Each view identifies a set of shared data among queries. They are processed only once, but reused by at least two queries. It verifies range similarities to merge queries. Achrekar et al. [1] proposed a multiple query optimization technique to minimize the average transmission time in sensor network for aggregation and data acquisition queries. The optimization takes place in two stages 1) Base station 2) Network. In the base station optimization, the user queries are rewritten into synthetic query by attaching the cost function to each of the query with predicates and combining results of computation at each stage using heuristics. The second tier optimization is implemented on sensor nodes. Communication cost is saved by reducing the number of radio messages sent between the base station and the sensor nodes. For satisfying the individual user queries, results coming from the sensors are filtered and extracted at the base station. In the case of traditional methods, queries are executed independently, that leads to the wastage of energy in the network.

In this paper [8], prediction based component is used for rewriting the queries. It is based on the semantic criteria and with the use of approximate data from the network map.

In general, data communication among the sensor nodes requires more energy that can be measured by descendent nodes reachable from the current node. This information helps in eliminating the unnecessary packet transmission by filtering.
out the child nodes for query dissemination or query. In this paper [16], authors proposed a technique for reducing the number of packets needed for query dissemination or query results. Context-aware routing tables are used to communicate with the neighbours.

Multiquery optimization is designed for optimizing the multiple queries and executing the operations only once. These are mainly used in decision support system. The approaches for multiuser optimization can be categorized as systematic, heuristic and semantic optimization [52]. In systematic optimization, optimizer estimates the cost of all possible plan and selects the best plan from that. In the second approach, for minimizing the resources, operator ordering is used. Amount of processing has to be minimized without affecting the output. Semantic optimization is the combination of both the optimization.

Multiple queries are resource constrained so we need some middleware as an interface between high level abstraction and system level programming. Tiny DB interface has several limitations and does not give optimized result in multiple query processing. It uses distributed processing architecture. The goal of the authors [17] is to propose a new three modules architecture for energy efficient processing.

The proposed engine is divided into three modules such as client side, base station and wireless sensor network. Client sends the request in XML format and receives the response in XML format only. Base station consists of XQuery listener, XQuery parser, and SQLParser and XQuery response. The third module WSN is organized as a hierarchical network with multiple layers and each and every layer is having one or more sensor motes. One mote elected as a root node responsible for taking the SQL query as input and passing the same to each child sensor motes. This architecture is removing the heterogeneity from the sensor network in term of database and also prolong the lifetime of the sensor nodes. It is also supports multi-query environment instead of a single query.
Nan et.al [18] proposed a multiuser management framework for supporting historical and continuous queries. The idea is to reduce the common tasks in the collection of queries through merging and aggregation, according to the query region, attribute, time duration and frequency by executing the common queries only once. This scheme supports query partitioning, region aggregation, approximate processing, and time partitioning and historical database. The query execution plan generated by the query management scheme is based on collecting a group of queries, which will lead to latency in query answering, especially for the queries collected at the early stage. This scheme performs a significant amount of pre-processing and post processing work, which also directly affects the query response time.

A new query processing framework was presented in the form of compiler/optimizer stack by the authors [19]. They described the WSN is a platform for distributed computing. The presented framework distributed query processing for WSN that understands the expressive declarative queries and generates energy efficient query plan. They introduced the user level syntax and physical algebra for SNEEQL. They also discussed about the physical algebraic operators, energy analytical cost-estimation models, architecture for the optimization of SNEEQL queries, algorithms that supports integrated query planning that includes routing, placement and timing and empirical performance evaluation of the framework.

The events detected by the sensor generate a stream of data. Centralized join query processing incurs more communication overhead due to frequent exchange of data between the sink and the sensor nodes. Implementation of multiple data streams in a sensor network is challenging due to limited memory and battery power. The design of Distributed Nested Loop Join processing algorithm (DNLJP) [20] for efficient join processing, groups the sensors based on geographic locations to form a cluster and perform the query processing in a distributed manner over the data collected across different regions [20]. Query
cost of the centralized scheme and the proposed scheme are compared and the result shows that the query cost is low in the proposed scheme.

Authors [21] described the join processing by discarding tuples that do not join and continuous join filter is introduced in this approach. Join Filtering is that the size of a filter depends on the data of other nodes. Filters have to be installed on the nodes as they enable discarding tuples and at the same time, can prevent incorrect results. Optimizing filters require knowledge of the sensor readings of all nodes. The base station needs to compute continuously filters that minimize communication costs. For each execution, the base station has to decide which of the current filters are to be updated.

In wireless sensor network, joining or correlated filtering of sensor readings are one of the basic operations of query processing. In a join query [22] of in-network processing where join operators are pushed into the network such that the amount of data transmission could be considerably reduced is essential. A join query in WSNs is often characterized by its selection/join predicates on the auxiliary attributes. The auxiliary attributes such as the location of a sensor node and the timestamp of the sensor readings represent the spatio-temporal dimension of the sensor readings.

Yuan He et. al [23] described a query centric framework for collaborative heterogamous sensor networks. It consists of a central manager and numerous collaborative heterogeneous sensor networks. These are connected over the internet. In this, each sensor network is considered as a data source, which accepts the external queries and provide the instant information at the cost of internal resource. The data of one sensor correlates with other one. These are called implications. In this, complex query is processed by a pipeline of the sensor networks, so that implications can be utilized. They designed a heuristic approach Implication Aware Pipeline to minimize the energy cost caused by query processing in the distributed context.
Skyline query optimization [24] is used for multi-criteria decision making and can be used to monitor the sensed data under multiple criteria. Skyline query can be defined as $p = (p_1, p_2, p_d)$ and $q = (q_1, q_2, q_i)$ are two $d$-dimensional points, where $p_i$ and $q_i$ are the sensed values of $i^{th}$ dimension of $p$ and $q$. Each point in the equation refers to a sensor and each of its different sensing device readings corresponds to a dimension reading. Two algorithms are applied on a snapshot dataset. First is to partition the dataset into several disjoint subsets and produce the skyline points in each set. Another is to employ filter that consists of some points in the processed subsets to filter out the points from rest of the unexamined set.

Advancements in the mobile devices increase the high traffic on the internet. These devices are capable of storing considerable amount of data locally, increases the query optimization. It also decreases the network traffic, since the user can retrieve the result from the mobile database itself. Authors [25] proposed the query model for processing the data from the cached-database of mobile device. This can store a small portion of the database into the local memory so as to provide the result faster to the query interface. They introduced Sequential and Sorted Query optimization [25] schemes. In the first optimization, data is retrieved in the order of columns appear in the given query. In sorted optimization, first the columns are sorted out and then data is retrieved from the network. Execution time is the basic difference between these two schemes.

Cache based optimization improves the execution time, minimizes the communication cost and traffic at centralized server. There are two schemes introduced in this optimization. One is event driven caching and other is location dependent caching. Event driven caching is based on frequently asked queries at specific time and location dependent is on the basis of frequently asked queries by referring the current location.
In-network query processing [53] [54] demands the detailed query details must be disseminated to all sensors, thus, energy would become a constraint. Without the necessity of query dissemination details, authors [26] purposed an energy efficient query processing. Node sends the data to the base station. But node can be the combination of router and processor with higher energy. Those nodes are not sensor nodes, only deployed for the routing purpose. They divided the query processing into five levels. In the first level, all the queries will be collected in the base station. In the next level, sensors are grouped into query domain macros [54] such as southern region, second floor etc.. They are then mapped to the corresponding sensors. In the third level, they introduced the concept of cache management. Before sending the query to the sensors for the processing, first cache in the base station will be checked for the data. Base station stores the Cached Data Element (CDE) corresponding to the sensors [54] and each one have four elements. First refers the last cached sensed data, second means round gap between two consecutive sensing. Third specifies the number of rounds left until the next sensing. This element might have negative value also. Fourth refers the accuracy of the data stored in first three element. These elements are updating on every round. If the values are assigned never in these elements, then they are considered as invalid.

In the fourth level of this architecture, base station send the query for sensing the data from the network. Sensors are selected from their query domain and CDE value is invalid. In the next level, all selected sensors are sensed the data and update their sensed Cache element (SDE) [54].The semantics of SDE is also following the same set of rules as per CDE. The fourth element is a local counter which is holding the value between Inchng and MaxChange. Applying an algorithm, SDEs are updated and routed the details to the base station. Now the base station constructs a tree and disseminates the query for processing.

Sensor networks become integral part of internet. Internet scale query processing [56] is needed for processing huge volume of data generated from the
heterogenous network. Local query refers the query within a single network and Global Query refers from the multiple networks [27]. Authors [53] introduced the concept of sensor web, it is an architecture for accessing distributed heterogeneous sensors from the internet. For processing these queries in the internet, in–network processing is a suitable approach. Whatever be the type of query, that has to be forwarded an appropriate network and selection of the network uses the spatial and temporal correlation of data is used.

There are two types of WSN: Atom WSN and Collaborative WSN

An atom-WSN [53] is the basic building block of this architecture. It consists of number of homogenous, heterogeneous nodes and sink nodes. It may include the fusion nodes also. All these nodes can be IP enabled for the communication with other service WSN domain. For node level operations, aggregate functions and interaction at sink level, IP enabled sensors are able to communicate within their levels. Internet protocol interface is designed to meet the various functionalities.

Collaboration WSN can be viewed as vertical collaboration and horizontal collaboration. Vertical collaboration refers communication between WSN, Service domain and across the IP networks. In the horizontal collaboration, communication between geographically overlapped WSNs. Since Service domain is not playing any role in this collaboration, traffic cannot be pushed into the IP based network. Traffic will be within the participating WSN only.

Query processing in this architecture first determines the type of query and decides the participating WSNs for executing the query. For the execution of the query, participating WSNs are ranked based on the correlation matrices. Nodes in every participating WSNs have temporal correlated and spatially correlated attributes. These are independent of query but playing a role in the selection of WSN for the query execution.
Metadata management:

Sensors are becoming essential for monitoring the environment. They collect geo-referenced data and analysed that data. Heterogeneity and dynamic changes in the network are the challenges for the researches for maintaining interoperability of WSNs. Any node failure or malfunctioning of a node changes the status of a network. So the system must follow the self-adaptive process for the re-configuration, optimization and repairing of a network itself without human’s interventions. Metadata elements are provided for maintaining the dynamic interoperability of WSNs. These elements can enable the understanding of the network itself [28]. Metadata provides the description of the sensed data and preserves the context of the sensing data. Metadata becomes integral part of the WSN and it defines as descriptive data used to depict the network, including the environment, the nodes, status of the nodes, sensing data and the whole system [28]. The metadata is mainly used for routing and query aggregation process. They must generate, update and change dynamically, then broadcasts the changes to the whole network.

Authors [28] introduced the concept of context and it is a box that contains two parts. WSN status describes the inside of the box and metadata elements and their data values are inside the box. WSN status in the domain of dynamic interoperability is represented by the context. Contextualising rules are used to infer the contexts of WSN and Bridge rules are used to allow the relationship among the different WSN contexts. The contexts described by the authors are Sensing Context, Node context, Network context and Organizational context.

Sensing context are describing the sensing conditions and its operations. It mainly focuses on the sensed data. Node contexts are describing the knowledge of individual nodes in the network. Network contexts relate to the communication, collaboration and processing functionalities among the nodes. Legal, privacy and security issues are considered in an organizational contexts.
Metadata elements are used for developing the contextualising rules to infer different WSN context.

Metadata are abstraction and provides the knowledge of network for energy efficient query processing. The type of metadata needed, formalization of metadata, managing the metadata and synchronizing base station metadata from the nodes are needed for the query processing. Various stages of query processing such as parsing, optimization, dissemination and execution of the query depend on the metadata. Different type of queries of the network need different type of metadata for the execution. Metadata management should provide adequate information for the query processing without compromising the cost for managing the metadata. Metadata can be classified as distribution and variability [29]. They can be stored statically or dynamically according to the requirements. Zhang et.al [29] introduced the metadata manager at the base station as well as at the cluster head. Collecting the metadata is a challenging job and models are created on the base station for managing the metadata. Push and Pull operations are adapted for the different roles. Nodes are classified into groups according to their position, phenomenon and other attributes. Aggregation and Compression of sensed data can be further used for reduction of the traffic.

Reliability of the wireless sensor network is affected by faults due to hardware malfunctioning, environment hazards or software related issues. In this paper [5], authors presented the various types of faults and fault tolerance techniques for the enhancement of service.

Following major failures can occur in a WSN, i) Power Failure caused by fault in the base station ii) Network failure caused by faulty link iii) Transmission of arbitrary data caused by faulty nodes. The goal of fault detection is to ensure that WSN is working properly and collecting correct data.

Fault detection techniques are classified as Self-diagnosis, Group and Hierarchical detection. In Self – diagnosis, nodes itself identify faults in their
components. In the group detection, several nodes are monitoring the behaviour of the other node. In the hierarchical detection, hierarchical tree is formed and forward the results of child nodes to the parent nodes and up to the sink. Parent node aggregates the status of the child nodes and forward that to the next level.

2.2 Review Literature on Clustering and existing Clustering Protocol

Recent researches in the wireless sensor network are investigating the different approaches for energy efficient query processing.

Data management and Aggregation

Energy Efficient data collection is playing a very important role in wireless sensor network. Researchers are focusing on effective algorithms that reduce the energy consumption with effective in-network techniques.

Approaches used for data collection involves all the nodes of a network. In general, large number of queries needs only subset of those nodes only. So authors [30] concentrated on only selective queries and the task of data collection is an instance of NP-hard problem. They presented an algorithm called Pocket Driven Trajectories that optimizes the data collection path. They optimized the path by approximating the global minimal Steiner tree with the support of spatial knowledge. They identified the number of factors influencing the efficient data collection such as the distribution of nodes over the network, location, arrangement of clusters, location of sink etc…. They referred clusters as packets and this algorithm selects the set of packets for a given query. It aligns the aggregation tree to the spatially optimal path connecting these pockets. This path maximizes the use of participating nodes, thus reducing the number of non-participating nodes.

Sensor nodes are randomly deployed for data collection at specified location. Data collection requires sensing data from all the nodes and then collected data should be forwarded to the base station accurately [31] [32].
The efficient way of reducing the communication cost is data sharing among multiple applications. Authors [33] idea is to reduce the number of samplings at each node resulting in less communication cost. In the suggested method, each node samples the data for a continuous interval instead of sampling a discrete data point. This problem is formulated as a nonlinear, nonconvex programming problem.

Query optimization in WSNs is dealing with network schemes or distributed algorithms for reducing communication cost for aggregation queries. But the previous proposed work focuses on reducing the amount of transmitted data for each node. The proposed greedy 2 factor approximation algorithm is solving the above issue in the resource constrained sensor nodes. The time complexity of this algorithm is $O(n^2)$, and the memory complexity is $O(n)$. All the tasks are arranged in ascending order and identify the subset of tasks that overlap with the first task. Minimum interval that could be shared by these tasks are identified and remove the overlapped tasks including the first. Repeat the same steps for the remaining tasks until all the tasks are removed.

Liu proposed a mobile database structure [6] and discussed the overall design and implementation for the data management system. A query process is disseminated from top to bottom, till it reaches the leaf nodes. Result is collected from bottom to top and query processor is responsible for executing that. They introduced the sensor node manager which is managing the sensor mode of database. Sensor mode manager includes some components: Attr, Command, WsnMDBAttr, Tuple, WsnMDB Command, and Query Result. The function of Attr is used to get property values, Commands are to achieve various orders in the mode and WsnMDBAttr is the concentrator of properties to connect various fixed property components in the database.

In order to manage the large amount of sensed data in an energy-efficient manner, new ways of storage and data query methods are needed. Authors [2] surveyed about the most energy-efficient data storage, query techniques and its
related issues in this paper. They discussed about the various data reduction techniques and open issues related to query processing. Efficient meta-data management is very useful when one can manage a huge amount of distributed data. Efficient data management in WSN allows the transparent access to the sensed data and also to enhance the lifetime. In–network aggregation is the best technique for reducing the amount of data transmitted between sensor nodes.

Nandini et al. [34] introduced a framework for the data aggregation in the network. This framework acts as an interface for aggregating the data sensed by a number of nodes within a wireless sensor network. Simulation of the queries with and without aggregation have been shown using TOSSIM simulator. The result showed the efficiency of the aggregation queries in WSN. Author [35] introduced the algorithm that calculates the energy cost before and after aggregation according to the attribute historic data. Attributes are divided into two categories: inherent attributes and sense properties of nodes. Inherent attributes are static and sense attributes are such as temperature, pressure, sound etc. When several queries are sending the data to the base station, they can be merged and repeated data only need to send only once to the base station. Queries can be merged according to query attribute or query condition. Merging of query attributes can be done if attributes have repetition and query conditions are not related. Merging of query condition is possible when the attributes scope is overlapped between two query conditions.

WSN consists of a number of sensor nodes and a sink node. Nodes are combined together known as clusters for supporting scalability. One of the node will be elected as leader and called as cluster head (CH). Cluster head is responsible for processing the queries as well as routing the sensed data to the target nodes.

Authors [36] objective was to achieve a balance between utilization of in-network processing power and communication. They also introduced the query data volume for finding the cost associated with each query. There is a high
degree of redundancy on cluster based queries according to their simulated result.

Authors [37] presented an efficient request-oriented coordinator methods for hierarchical sensor networks. Power consumption of a node is lesser than the energy of the cluster head. Energy consumption of communication based on transmission distance which is exponentially incremental according to the distance. Random selection of a cluster head is not able to balance the energy among sensors. They suggested the selection of cluster head based on energy information.

The node with the higher energy in a cluster will become the next cluster head. All energy aware routing protocols are based on cluster based but power of the cluster heads are drained rapidly compared to other member nodes. So cluster must be reconfigured at regular intervals. It requires a lot of packets to dismantle old cluster and set up a new cluster.

Nodes can control their transmission power to transmit data at different distances by the protocol suggested by the authors [38]. Nodes in the cluster may not send their packet to the cluster head directly and send their packets to the intermediate nodes in different portions to distribute energy dissipation of all nodes uniformly.

Yong et. al [10] introduced the cougar approach for the network processing. In the previous sensor networks, sensors are pre-programmed and send the data to the base station for further query processing and analysis. In the previous approach, user cannot change the behaviour [10] of the system dynamically and battery power is also a design constraint. In the cougar approach, declarative queries are introduced and without the knowledge of the network, users can submit the query.

In the wireless network, the queried nodes are able to process the multiple queries in their limited memory and computational capacity. Traditional SQL query does not satisfy the constraints of WSN.XML [39] [40] is a standardized
language for data interchange in the wireless sensor network. It supports complex data management and heterogeneous networks also. Authors [39] proposed a XML based XQuery [39] [40] based query processing architecture for processing multiple queries. Due to its hierarchical presentation, XQuery architecture is well suited for mutliquery optimization.

XML usage and evaluation of XML queries in the wireless networks are discussed by the researchers [40]. They presented the Path engine on updateable compressed XML data for sensor nodes. XML data management would be very useful in the heterogeneous network, WWW integration and service oriented architectures. They developed the XOBE Sensor Network integrated framework for the integration of XML into WSN. This network allows the usage of XML variables in C and checks the type at compile time and transforms them into string operations.

Multi-hop mode of communication is used to forward the data to the base station since the sensor nodes are energy constraint devices. Existing literature showed that sending and receiving data needs more energy than any other operation of the network. So good routing protocols are to be developed for enhancing energy efficiency. Each node may route messages using either short hops or long hops. [41]. Authors proposed the algorithm for optimizing the length of these hops and therefore energy can be saved with the extended lifetime. Minimum hops length would be considered for the retransmission in the case of transmission failure.

Various routing protocols are introduced for this purpose. Cluster based routing [42] [36] protocols are based on Hierarchical routing. In this protocol, nodes are organized into the clusters and node with higher energy will be elected as cluster head.

There are so many clustering protocols which are studied in the literature such as LEACH [43], HEED [44], TEEN [45], PEGASIS [46], and EEHC [47] for efficient processing of the queries.
In LEACH, cluster formation is in a distributed manner. In this algorithm, cluster head is rotated among all the nodes and any sensor node can be the head randomly during each round. LEACH [43] [48] [47] [49][45] is a specialized hierarchical clustering protocol for minimizing the energy dissipation in the network. In this protocol, cluster set up and operation is well coordinated with the localized control and selection of the cluster head is based on random rotation. Local compression techniques are used for reducing global communication from the cluster head to the base station.

Each and Every cluster, local computation is performed, thus reducing the communication cost for sending the data to the base station.

It is a self-organized [43] , distributed protocol and energy is distributed uniformly among the sensors in the network. The operation is divided into two phases: set-up phase and steady state phase. In the first phase, selection of clusters and cluster head is decided for further processing. Cluster head selection is based on percentage of the cluster head for this round and number of times that particular node has been selected as a cluster head. Cluster head is rotated between all nodes in each and every round.

Once the cluster head selection is over, the message will be communicated to all the nodes in that particular cluster. After this phase, each node decides to join a particular cluster based on the signal strength. [43].After the creation of cluster head and cluster, cluster head starts collecting the data from the nodes in their particular cluster. In this algorithm, nodes does not require the knowledge of the entire network.

Another distributed, hierarchical energy efficient clustering protocol HEED [44] [47, 48] [49] selects the cluster head periodically. The parameter for selecting the cluster head is sensor node energy. In this protocol also, life time is enhanced by distributing the energy in the entire network. The cluster head selection process will be terminated after a particular number of iterations. In this approach, authors [44] assumed that the sensor nodes can control their
transmission power level. Heed protocol allows single hop communication within the cluster and multi hop communication between cluster head and base station.

There are two parameters playing a major role in the cluster head selection. Energy is used for initializing the cluster heads and intra communication cost is the deciding factor for joining a cluster. This cost is basically calculated on node’s proximity or cluster density.

In this algorithm, clustering process will be initiated in a particular interval for the selection of cluster heads. Probability value is estimated for the cluster head selection and that should not be beyond the threshold value. If the probability value is less than one, then the node becomes a cluster head tentatively and becomes permanent cluster head if the value reaches one. The parameters such as minimum selection probability and network operations interval can be tuned to optimize the resource usage [44]. This approach uses the availability of multiple transmission power levels at sensor nodes.

TEEN [45] [49] is designed for reactive networks. In this protocol, closer nodes form the cluster and repeats until it reaches the base station. In every round, cluster head broadcasts hard threshold and soft threshold values to the other nodes in their cluster. Hard threshold [45] value indicates the sensed attribute value and beyond this value, the node transmits the value to the cluster head. This is called sensed value. Soft threshold refers the small change in sensed value which triggers the node to transmit its value.

The node transmits the current sensed value only if the value is greater than hard threshold and Sensed value is set equal to the current value of the attribute. Both Hard threshold and Soft threshold value is used to reduce the number of transmission by checking the value against sensed value.

PEGASIS [46] [49] [9] [47][45] is a chain based protocol and a single node in a chain is used to send the data to the base station. Each node can communicate
to its closest neighbour and chain is constructed in a greedy way. Then, data is
gathered from node to node and moved in a chain. Designated node aggregates
and transmits the data to the base station. Cluster head selection in this protocol
is not considering about the energy of nodes and location of the base station.
Nodes takes turns to transmit the fused data to the base station for balancing the
energy depletion [46].

Data storage is a major component of wireless sensor network. Storage can be
local or distributed within the network. Distributed Data storage [50] is the
major component of storage which protects the critical data from the failures.
Authors [50] proposed a protocol called DISC for distributed storage and
collection in the wireless network. A distributed Information Storage and
Collection (DISC) protocol selects the backup node randomly. Bloom filters are
used to store the data inside the network. This makes faster retrieval and
searching of data from the network. They suggested primary cluster head
(PCH), backup cluster head (BCH) and reader node in the architecture. Reader
node acts as a base station in this protocol.

The function of the protocol is divided into data dissemination and data
collection process. PCH and BCH nodes are performing data dissemination
process and data collection is executed by BCH and Reader node as well.
Backup cluster head is selected randomly from the nodes. Sensor nodes are
reporting the sensed value to the Primary cluster head which in turn forwards
the data to the neighbouring Backup node after aggregating the data. A unique
descriptor is included in the aggregated data. It includes epoch duration of the
data, region identifier and type of data. Bloom filter [50] is used to obtain one
element from the above parameters. After receiving the data, BCH does not
store the data locally but distributes among the sensor nodes. BCH is managing
trace – table which store the identifier of the node keeping a particular piece of
information. [50] This protocol uses four types of messages for initializing,
forming primary, backup cluster, discover the backup aggregators and disseminate the data.

Clustering is playing a vital role in the wireless sensor network. Lifetime of WSN can be enhanced by the proper selection of cluster head. Authors [38] proposed the cluster head selection by Randomness with data recovery. In this approach, cluster heads and vice cluster heads are selected randomly and heterogeneity in energy is maintained. If the main cluster head drains in the power, then these vice-cluster heads can be selected as head node. Cluster head selection is based on three phases. In the set up phase, base station maintains the details of node location and power. Randomness will be calculated and if in the range between -2.04383 and 2.839258[38], the node will be selected as cluster head. In the next phase, selection of vice-cluster head process starts and checks for each and every node. If the node is a cluster head, it transmits the request for power and location from its neighbouring nodes within one hop distance.

From the list of one hop distance nodes, again randomness range will be checked. If it falls within that range, that node will be added in the VCH list. In the third phase, cluster head checks for the power signal and the value is lesser than threshold value, first VCH from the list will be elected as a cluster head and all required data will be sent to that VCH. This information is broadcasting to all the nodes of the network.

This approach supports recovery mechanism also. Any node or cluster head gets crashed, broadcasts this information to all the nodes and new cluster head will be selected from the list of VCH. All the entries in the database table are updated and uncommitted values are removed from the database.

Sensor web is used by the Open Geospatial Consortium [51] for location aware sensing devices that can report data through the Web. All the nodes are connected and communicated using Internet and controlled by Web interface. In sensor data management, sensors network is considered as distributed
database systems. Saving energy is a critical point in the sensor data management. Semantic representation of sensory data is significant since ontologies used for domain knowledge and their relationship. In the ontology architecture, system is divided into three layers for managing heterogeneous data sources. Each and Every layer is independent and does not affect the implementation of other layers. The first Data Layer is for Data finding, collection and aggregation. Next processing layer is responsible for aggregated data and Semantic layer for the annotations and ontological descriptions. Semantic layer is optional and it provides the enhanced interfaces and query processing within a sensor domain. Semantic layer is provided with Rules, Ontology model, Reasoning Server and Semantic Queries.

Sensor network requires the organization of nodes for routing information. Hierarchical clustering algorithms are widely used for energy efficient routing. Cluster formation and Cluster head selection is very important for reliability, energy efficiency and load balancing of the network. Passive clustering is on demand formation protocol and cluster formation is dynamic in this approach. The exchange of flooded messages can be optimized for energy efficient processing in the network. In this approach, node in any of the cluster is not sending the data within a timeout period, they will be out of the clustered area.

In PCEEC [51], there are six states are defined they are dead, initial, ordinary, cluster_head ready, gateway and alternate cluster head. Initially, all nodes are in the initial state. This state continues till the time node receives the packet. When a node receives the packet, it checks for the state of the sender. If the sender is not a cluster head, receiver switches to cluster head ready, otherwise switches to gateway or ordinary node. Then it switches to gateway state in the case of number of cluster heads are greater than or equal to the number of gateways. In the case of failure, node with the highest level of energy will be
elected as next cluster head. Ordinary state node will change to alternate state if the energy of that node is higher than all other nodes.

Artificial Immune system (AIS) [7] [53] mimics the human immune system in solving the problems. It is a distributed adaptive system using models and principles and concerned with the abstracting the structure and function of the immune system to computational systems. Through antigen recognition process, it protects us against infection by producing antibodies. Common models of this algorithm include Clonal Selection algorithms, Negative Selection Algorithm, Immune network models and bone narrow models.

In all types of hierarchical clustering protocol, nodes are communicating to the cluster head at the same time irrespective of their conditions. Cluster Head Scheduling mechanism (CHSM) [7] is a scheduling mechanism inspired of AIS and the nodes with more information have a chance of communicating with cluster head. This mechanism is applied after the setup phase only. Selection of cluster members for the communication to the cluster head will be decided by using AIS algorithm. Each and every member can send the packet only at a particular interval. The cluster heads in turn store the transmission history from various nodes at different intervals as bits. These are called case history and after the setup phase all these bits are equal to zero. Whenever the cluster head initiates the communication, it allocates 0 or 1 at the most significant bit of the history bits after right shift. In this mechanism, cluster head communicates to the nodes with affinity value more than threshold value. In this algorithm, some of the nodes will be selected randomly after particular time slices and mutates the case history.

Authors [7] also mentioned about the QoS aware History Scheduling mechanism in which the distribution of the nodes in the cluster are not taken into consideration. The nodes selection is purely based on affinity value. Clusters are divided into sub clusters for improving distribution and CHSM is applied in each and every sub cluster. Nodes are static after the deployment.
2.3 Survey of Existing Routing Protocol

Routing:

Flooding algorithm is widely used for the query routing because of its distinctive advantages such as high reliability, high node coverage rate and low response delay. However, the flooding algorithm always introduces too much message consumption due to its redundant message forwarding between nodes in the network. [54]

To save the message consumption of the routing while still maintaining the time constraints to the routing, Bi-Filtered Forwarding (BFF) a kind of limited flooding algorithm for routing the query submitted by any node to all the other nodes in a wireless sensor network, with real time routing and less message forwarding traffic caused. In this algorithm, top and bottom filters are proposed by the authors [54]. Top filter is used to guarantee that only the node with high connectivity degree could become a forwarder to route the query message to all the other nodes so as to reduce the message redundancy. Bottom filter is used to guarantee that the node with low connectivity degree (at least 2) could also be a forwarder.

The performance of wireless sensor network is influenced by the physical characteristics. In this paper [55], authors developed the model for optimizing the network in a truck for cold chain monitoring. All nodes in the model are homogenous and energy constrained. They introduced the Virtual Base station and Energy aware Wait and See protocol [44]. Without the base station, life time of the network is enhanced by sharing the energy among sensors. VBS will be selected for a given duration as soon as the event is received. The major role is diffusion, updating and replicating that event in the sensor database. The purpose of protocol is avoidance of collisions and enhances responses to update request packets.
In this approach, all the nodes in the network must have the knowledge of collected data. Any new event occurs, that will be notified to the virtual base station. This station includes the sequence number to the event before broadcasting to other nodes. The database gets updated with the new information. If any node is not received the event details can query the VBS for the updating its data. Thus all sensors have the same information and user can query any sensor without locating each and every sensor. In Energy aware protocol, node calculates the time duration after receiving the update request. Within this duration, the node waits for other nodes answer depends on their energy level. After receiving Update Request Packet, sensor node waits for a particular duration without answering. It sends the response packet after confirming that none of the neighbours have answered for this request.

Energy Aware Routing Protocol [56] is based on clustering and cluster heads are grouped and routing tree is formed among the cluster heads. The nodes in this are static and only base station is deployed for controlling the network. Sensor nodes are not aware of their location and radio power can be controlled. In this approach also, each round consists of set-up phase and working phase. Routing tree is constructed in the set up phase. Collected data are sent to the base station in the next phase. Every node in this protocol maintains the information about their neighbour nodes. All the nodes are cluster head nodes and broadcast the messages within its range. Each node receives the message and updates its neighbourhood table. Authors introduced the concept of intra-cluster coverage in which some of the nodes are active within the clusters while maintaining the coverage expectation of the cluster.

Because of congestion in WSN leads to the dropping of data [57]. Some applications data rate will be higher than other type of applications. So, differentiated service must be provided for avoiding congestion in these type
of applications. In this paper [46], congestion aware routing protocol (CAR) is introduced and it identifies most appropriate route dynamically based on congestion and data priority. Higher priority packets are routed through congestion free zone by CAR and AODV protocol is used for lower priority packet. In the first phase of this protocol is design phase of congestion free zone (conzone) [46] and second phase is routing the packets into this zone.

Routing table gets updated after conzone is formed and higher priority data will be routed through this zone only. Spanning tree approach is used for routing the data to the base station.

Routing protocols can be categorized into proactive, reactive and hybrid routing protocol. Each and every protocol has their own merits and demerits. Adaptive routing protocols are based on redundancy node and dual routing [58]. Network can be divided into working and sleeping nodes. Both will be using two different types of protocol.

Energy efficient routing and Quality of service routing is a major issues in the wireless sensor network. D.Raj et.al [59] proposed an Enhanced energy efficient multi path routing protocol using cuckoo algorithm (EEEMRP). It is an extension of Ad-hoc on demand multipath distance vector routing protocol.

The Cuckoo search algorithm (CSA) [59] is a bio inspired, meta-heuristic optimization algorithm to solve the problem and provides energy efficient routing in the wireless sensor network. This algorithm considers several parameters such as throughput, packet delivery ratio, energy consumption and delay for finding an optimal solution. Breeding behaviour of cuckoo spices for their production motivates the authors for introducing this algorithm.

Distributed Quality of Service routing [60] selects a route with minimum delay and sufficient resources in a dynamic environment. For finding the best path, alternate paths are also searched for improving the quality of
service. Using reinforcement learning techniques, routing path can be optimized in order to improve the network performance [49]. Probabilistic nature of routing table entries are the main focus of Reinforcement learning algorithms.

The role of optimal routing and energy optimization is significantly affect the performance of the WSN. Due to the energy constraints, a smart routing is needed for balancing the energy. Soft computing paradigms are extensively adopted by the researchers for optimizing the routing. Some of them are Reinforcement Learning, (RL), Swarm Intelligence (SI), Evolutionary Algorithms (EA), Fuzzy Logic (FL), Neural Networks (NN) and Artificial Immune System (AIS) [61]

RL specifies the Q-Learning algorithm for the distributing and dynamic routing and clustering. It enforces the minimal communication and computational requirements.

SI is a system of collective intelligent groups of agents. They can interact with each other and outside environment. In this, all individual agents are following the rules that leads to global behaviour. Managing the collective behaviours of highly dynamic and distributed environment in decentralized and self-deployed systems are addressed by Swarm Intelligence.

GA mimics the process of natural evolution. It is one of the EA algorithms and does not suit for the optimal routing.

FA deals with the uncertainty of data. Since WSN is an uncertain environment, FA is useful in the case of decision making. It provides greater flexibility for dealing with uncertainty and avoids complex mathematical modelling.

Neural network is based on supervised learning methods in different environments. It is an arithmetical algorithm which maps input and output. [51] AIS mimics the human intelligent behaviour in computerized tasks and
problem solving. It can be applied for optimizing the routing and clustering of WSN.

**Multipath Routing**

All existing routing protocols are focusing on shortest path based on energy. If the energy of the node drains, then alternate path has to be selected for the completion of routing process. Instead of single path, multipath selection distributes the traffic among the alternate paths. Number of paths and selection of these paths impacts the performance of a multipath scheme. Energy efficient adaptive multipath routing [62] utilizes the available energy and received signal strength for identifying multiple paths to the destination.

In the first generation of wireless network, routing protocols and network architectures were introduced. Energy efficiency is achieved with the integration of in-network aggregation and data dissemination architectures. WSN is considered as a distributed database in the second generation and users can interact with the help of SQL queries. Researchers were developed algorithms for executing the queries efficiently. Multiquery optimization and aggregation and multidimensional queries are attracted the researchers in the third generation. There is a drastic change in the storage with the introduction of NAND flash memory. Sensor node can store the several gigabytes, thus reducing the communication and computational cost. Authors [63] studied the various generations of WSN.

Data dissemination architectures can be cluster based, grid, tree and chain based. These are called push architecture and query based are called pull architecture. Query dissemination is solely depends on tree based structure. Data storage can be sink model and distributed indexing and storage. Data storage can be internally in the nodes or externally at the base station. Indexing facilitates the faster retrieval of data. Distributed indexing is used
for the user queries. Query execution can be push, pull or hybrid. Push approach is a proactive approach and sensor readings are pushed out of the network. In the pull approach, queries are executing reactively and results are pulling from the network. The combination of push and pull approach leads to Hybrid approach.

In this chapter, detailed reviews of query processing techniques, existing routing protocols were studied in detail. Keeping the goal of energy conservation, the study of various existing clustering protocols and their comparison on the basis of cluster count, cluster head selection and methodology adopted for choosing CH is reviewed and explained in detail.