Designing a Three Level Smart City Architecture based on Big Data Analytics

[1]G.Radhamani, [2]K.Rajarajeshwari
[1]Director, [2]Research Scholar
[1,2]Department of Computer Science,
[1,2] Dr. G. R. Damodaran College of Science, India

Abstract: The concept of Smart City creates a better living standard for Citizens using Information and Communication Technology (ICT). The amount of data in smart city is high in volume so the architecture uses Big Data concepts. The proposed architecture has three levels namely Data acquisition, Data processing & Event Handling, Application level. This architecture uses Hadoop and includes usage of dimensionality reduction algorithms in the data pre-processing stage to clean the data. The cleaned data is then stored in the HBase. HiveQL is used to query the database. The architecture uses Classification algorithms and Optimal routing algorithms based on the event called. The storage and working of the architecture happens in the Cloud environment. The proposed architecture helps in providing personalized services in smart city applications.

IndexTerms **- Big Data Analytics, Cloud Databases, Hadoop, MapReduce, Smart City.**

I. INTRODUCTION

The Smart Cities are aiming at providing all necessary facilities to its citizens based on their needs with usage of latest technology. The Smart City architecture integrates the concept of clustering the daily life management such as water management, power management, transportation, health care etc. Growing population with complex necessity led framing of smart city architecture having automated and optimized improved living with the help of Information and Communication Technology (ICT). The ICT concepts which involve in the smart city architecture proposed here are Big Data and Cloud concepts [1]. An example of Smart Data architecture would be managing water supply in the city by visualizing the water usage pattern in people and finding the water deficient area and distributing the water efficiently that the water is evenly distributed to areas maintaining the proper Demand/Supply paradigm. This example of water distribution explains the vision of the smart city that is all citizens have their basic needs supplied in a non-biased manner. Big data efficiently handles voluminous amount of data, with high velocity and variety of Data. The Hadoop Distributed System (HDFS) is used to store the data in distributed File Format [2]. So the processing of the data happens in parallel processing. The paper discusses about the review of literature in Section II, about Hadoop Distributed File System in Section III and Smart city applications in Section IV. The proposed architecture is discussed in Section V, the future work in Section VI and conclusion in Section VII and the references.

II. LITERATURE REVIEW

Smart City research is a current scenario and there are various researches being done on the topic with different dimensions [3]. This section of literature review will give an idea about the various researches that has taken place in Smart City development.

The data collection in smart city development happens using various technologies like sensors, radars, social media etc. There are various advanced sensors and sensing application available in market for data collection in smart city applications (Costa, Daniel G., et al. Sensors 17.1 (2017): 93.). Sensors can be used in applications such as health monitoring, Traffic management, waste management etc. The sensors are used in detecting the weather and in-house power management. Sensors and radars do not want human intervention and can record data in any given conditions without interruptions. The paper uses fuzzy logic to control the visual sensors to acquire data based on trigger, time or query. The paper recommends Fuzzy logic controller (FLC) to control the internal and external parameters in the environment while acquiring data. The controller can handle any number of visual sensors used /deployed in the Smart city environment. There are also Unmanned Aerial Vehicles (UAVs) used in smart city data collection, the paper [5], discusses on the usage of unmanned aerial Vehicles for the purpose. The Unmanned Aerial Vehicles and driverless cars help in Mobile Crowd Sourcing (MCS) which is an important source of information for the smart cities. The paper uses Deep Reinforcement Learning (DRL) to develop a highly effective control Unmanned Aerial Vehicle Framework. The framework uses neural network for feature extraction of needed information and takes decision according to Deep Q networks. (Mohamed, Nader, et al., 2018) reviews the applications of smart city where there is a possibility to use Unmanned Aerial Vehicles (UAVs). The paper also discusses about the recent technologies in UAVs and their integration in smart city applications.

There are various smart city applications implemented already and the papers [7][8][9] , gives an account of various smart city architectures using cloud, Big data. The paper [7] suggests architecture for smart cities based on Soft Sensing. The soft sensing refers to acquiring data through crowd sensing and social sensing. The paper also converses about various algorithms used in smart city applications, the Support Vector Machine (SVM) Algorithm used in Smart transportation. Quadratic Classifiers used for Smart parking, Binary Classification used for health care application. The paper [8] focuses on developing an architecture which performs complete set of actions starting from data collection to event handling. The paper proposes a four- tier architecture, in which it generates data, preprocesses it, filters, aggregates, classifies and takes decision. The paper [9] discusses about the big data

paradigms in smart city using cloud computing concepts. The proposed architecture uses MapReduce model to implement parallel processing. The paper also discusses about two case studies related to smart transportation.

The role of Internet of Things (IoT) in smart city architecture is commendable and the following papers [10],[11],[12] concentrates on implementing Internet of Things in Smart City architecture. The paper [10] proposes a urban planning architecture which uses Internet of Things (IoT). A variety of datasets were used to validate the architecture. The Smart Home requirements and technologies needed to build the smart home are discussed in detail in Paper [11]. The paper discusses about the smart home heterogeneous network and fog computing architecture used in Smart City. The Paper [12] suggests a Fog-supported smart city network architecture called Fog Computing Architecture Network (FOCAN), a multi-tier structure in which the applications are running on things that jointly compute, route, and communicate with one another through the smart city environment.

A Smart City architecture developed for handling Big data is expected to have a Quality of Experience (QoE) better than the conventional Smart City Architectures. The papers [13],[14],[15] discusses about various Smart City architectures which has better QoE. Paper [13] proposes architecture for smart city which has the work segmented as planes namely data storage plane, data application plane and data processing plane. The data processing plane does the main function analyzing and processing the data using Machine Learning algorithms. The paper [14] presents a smart bot based on IoT, which collects massive amount of data and process it to give solution to the user. The architecture uses Mobile Cloud Computing and big data. Paper [15] suggests an architecture which will take the ethics, law and tradition of the people living in the smart city into consideration. The paper also proposes an architecture which is Ethics-Aware Object-Oriented Smart City Architecture (EOSCA). The work is based on the Object Oriented features, which will take real world objects into Object oriented feature. The papers [16, 17, 18, and 19] uses LDA and PCA in large datasets which are of variety of structures including traffic data and ECG data.

III. HADOOP ARCHITECTURE

Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model [12]. Hadoop is divided into two HDFS and MapReduce. HDFS is used for storing the data and MapReduce is used for the Processing the Data. HDFS has five services as follows, Name Node, Secondary Name Node, Job Tracker, Data Node, Task Tracker.

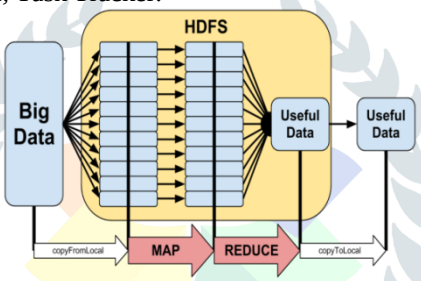

Fig.1. Big Data and Hadoop Architecture

The Hadoop Architecture is shown in the Figure 1, the master node is at the top, servicing nodes are the Slaves node at the bottom. Name Node is the Master node and Data node is the save node and they can communicate with each other. The term 'MapReduce' refers to two step process that Hadoop programs perform. The Map job takes a set of data and converts it into another set of data; individual elements are broken down into tuples. The Reduce job uses Map job's output as input and combines those tuples into a smaller set o tuples. A reduce job is always done after the map job.

IV. SMART CITY APPLICATIONS

The Smart cities are developed with a mission to offer the citizens with best amenities and personalized in a way that they feel their choices are fulfilled within the system. The core infrastructure consists of the following elements, which are considered as key elements in the formation of a smart city.

- Adequate Water supply
- Customized Power Management
- Personalized Health care
- Efficient Transportation
- Solid Waste Management

The proposed paper discusses about two features and how efficiently the data in the application can be utilized to provide users with better services. The two core elements discussed in the paper are Transportation and Health Care.

4.1 Smart Transportation

The transportation of People and goods is one of the main element in the Smart City. This section discusses about, how a smart transportation system can be formed with the use of technology. Here the problem is divided into two parts/segments, 1. Efficient resource (Vehicle) allocation based on the Demand/Supply Chain 2. Optimizing the travel path. In the resource allocation part, the paper visualizes and predicts the need (Demand) and allocates (Supply) the resources according to the need. This Smart transport will ensure the resources are equally distributed and resources are utilized wisely.

4.2. Smart Health Care

The health care is moving in a fast phase development that disease centric approach is moving towards person centric/personalized health care system. In the Smart city architecture the health department application will play a significant role in providing services to its citizens. The services provided in the smart city application will be a personalized service so that the

user feels that the service follows him. The health care system gives the government an idea about citizens' health, account of epidemic diseases so that the Government can take preventive measures in advance.

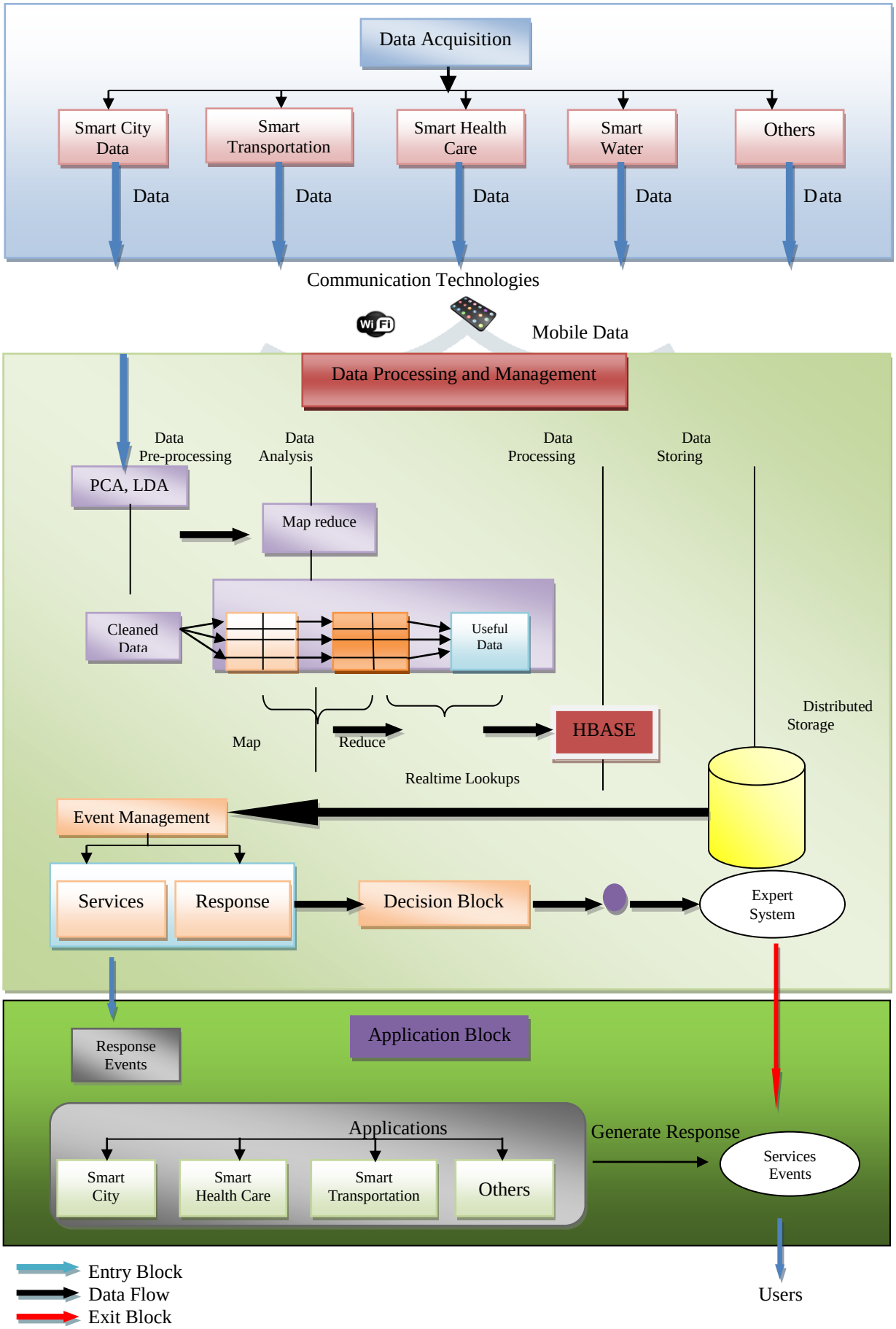

Figure 4: Three Level Architecture of the Proposed System

V. PROPOSED ARCHITECTURE

The Smart City comprises of providing personalized services to citizens that using technology they are able to save time, cost and energy. So building smart city architecture involves processing of voluminous amount of data, for which the conventional RDBMS structure doesn't suit since data are usually unstructured in Big Data. The features include handling voluminous amount of Data, Variety of Data with high velocity. To handle this challenge cloud computing is being used and by framing proposed architecture in cloud.

The Hadoop framework is being used in processing the data as it uses Map/Reduce concepts. The data acquired through sensors, tweets, historical DB etc are being cleaned, missing values filled and stored in the databases. It applies clustering algorithms to the tuples to form clusters of certain groupings to identify the interested groups. The research work concentrates on finding which algorithms performance will be best suited for the application and further decide on the modification needed in the current algorithm. After clustering, it uses the co-ordinates of the location to find the optimal path of the destination. It computes distance for each optimal path with cost so the best optimal path which would reduce energy and cost can be identified.

The proposed architecture acquires the data using various sources. The data is collected using various applications. The data collected are of various forms mostly unstructured. The data from the data acquisition level is transferred to the data processing and event handling level. In the data processing and event handling level, the cleaning of Data is done using the data cleaning algorithms such as Linear Discriminant Analysis; Principal Component Analysis the noise removal and dimensionality reduction is being done. These two algorithms are being selected to be used in the proposed architecture based on the results of various research paper which states they are best suited to handle unstructured data.

5.1Data Pre-processing

The data pre-processing (dimensionality reduction) is done by the algorithms Principal Component Analysis and Linear Discriminant Analysis. The working of these algorithms is explained in this section.

5.1.1. Principal Component Analysis

A Principal Component Analysis aims at reducing the dimensions of the database without losing much of information. The PCA algorithm is an unsupervised learning algorithm since it ignores the class labels. The PCA works on the principle of maximizing the variance in a dataset and it is a linear transformation algorithm [20]. PCA algorithm is best suited for large dataset and the dimensionality reduction is done without losing information. PCA can be further extended if there is a change in the dimensions of the database. The PCA can be combined with incremental methods and can be used to compute the co-variance with already computed variance as the new data arrives. This is called online PCA or incremental PCA. PCA can be extended to non-linear projections easily and called as non-linear PCA in which the projections can be generated based on the non-linear function chosen by the user [21]. The algorithms will be applied to the selected dataset.

5.1.2. Linear Discriminant Analysis

The Linear Discriminant Analysis algorithm's goal is to project a dataset onto a lower dimensional space with classes separated thus avoiding over fitting and reducing the cost [22]. In contrast to PCA, LDA maintains the class discriminatory labels. PCA and LDA algorithms are widely used algorithms in dimensionality reduction. Since LDA works on maintaining the class labels it is often known to be superior to PCA but there are also some areas where the PCA suits the application than LDA. PCA is an unsupervised algorithm that works on the principal of ignoring the class labels and concentrating on maximizing the variance of the dataset While on the other hand the LDA works on the principal of maintaining the class labels and performing the function of dimensionality reduction.

After the cleaning of data is done the Map function is performed and then the Reduce functions are done. The Map divides the data into tuples, and the Reduce task combines the set of tuples into smaller set of tuples. After Map/Reduce the data is stored in HBASE and the Hadoop Distributed File System (HDFS) is used to store the data distributed. The HBase is a non-relational database which runs on top of the Hadoop and provides with random data access and querying. Since the decision making is done in real-time it needs real-time lookups, in-memory caching etc which makes HBase suitable for the architecture. The distributed data storage is accessed by the Event Handling level where the events generated are either Service oriented or Resource Oriented.

5.2 Data Processing and Event Handing

The events generated are handled, categorized and decisions are taken based on the intelligent system. The decisions are implied on the application level. The algorithms are called based on the application and the response block. The algorithms and its working are explained in this section.

The events are generated usually in the following ways 1. Based on the triggers generated due to abnormal data readings (example: spike in Blood pressure, detection of traffic congestion in a route) 2. Based on the user input (example: The user wants to know the best optimal path from Source A to Destination B). The response block is responsible for the generated events. Figure 3 represents the pictorial representation of the event handling model, the respective response module handles the event and uses the query system to query the database and provides the response. If the event is raised by the health care module, the response is

usually performing a Classification with the help of supervised classification algorithms such as Support Vector Machine and K-Nearest Neighbor and if the patient belongs to the high risk category then the alerts are given to the doctors and the concerned patient.

5.2.1. Support Vector Machine (SVM) Algorithm

The Support Vector Machine Algorithm is a supervised learning algorithm which plots each item as point in the n-dimensional data space and the value of the data item as the co-ordinate. Then it performs classification by the hyper-plane which differentiates the two classes. SVM is used for Classification because SVM works very well with clear margin of separation, effective in high dimensional spaces. It is highly effective in situations where number of dimensions is greater than number of samples; uses support vectors so it is also memory efficient. It is highly effective in situations where number of dimensions is greater than number of samples. The SVM algorithm will be best suited for the proposed architecture where the clinical data used will be high in dimensions. The classifier will consider all the test reports (dimension) of the patient and classify them as high, low and moderate risks.

5.2.2. K-Nearest Neighbor

The K- Nearest Neighbor (KNN) algorithm is a non- parametric lazy learning algorithm. KNN stores all available cases and classifies new cases based on similarity measure. The new instance or an object is classified by majority of the votes of its neighbor classes. The reason for choosing KNN for our architecture is that it is one of the most popular algorithms in classification. It can be applied to data from any distribution and KNN gives a good classification if the number of samples is large enough. Here the proposed architecture uses KNN considering the large amount of data involved in the system. The results of both the algorithms will be analyzed and a detailed study on their performances in the system will be discussed in future.

5.2.3. Dijkstra's Algorithm

Dijkstra's Algorithm is used widely to find the optimal path from a given source to destination. Dijkstra's Algorithm is an uninformed algorithm that it does not know the target node beforehand. Dijkstra's Algorithm picks the next node based on the lowest cost of the distance, so it can be used in large graph area covering multiple targets. These characteristics of Dijkstra's Algorithm make it suitable for the application.

5.2.4. Genetic Algorithm

Genetic algorithm is used to solve both constraint and unconstraint optimal path problems. It uses natural selection and process it in steps from biological evolution. Genetic Algorithm generates new population at each iteration; hence it is suited for large dataset. The genetic algorithm works on three main rules namely Selection rules, Crossover rules and Mutation rules. These algorithms are used in finding the optimal path for the transportation application where the system calculates optimal path from Source to destination. A hybrid algorithm combining Dijkstra and Genetic Algorithm to be used in the architecture.

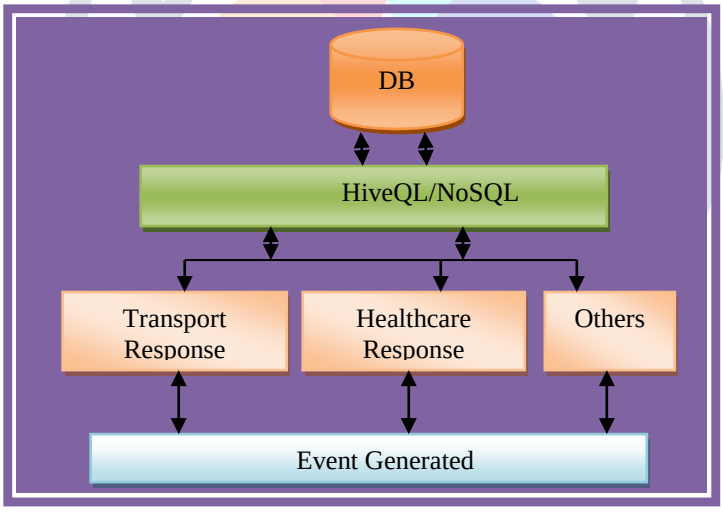

Fig 3. Event Process Model

The proposed architecture uses Cloud computing, which enhances speed, saving in cost and readily available. Cloud Computing offers a big platform for smart cities and there are various services offered, which can be used by smart cities. The service ranges from Smart devices, sensors, storage, middleware components. In the proposed architecture we use the middleware and storage in the Cloud infrastructure.

5.3 Applications level

The applications level consists of smart city applications such as traffic control, health department etc. The corresponding event gets generated and the service reaches the end user. The application level takes care of the user interfaces. Each application has a different interface. The application level acts as an intermediate between user and the proposed architecture. The user preferences are recorded by the application blocks and respective response is being transferred to the user.

VI. CONCLUSION & FUTURE WORK

A smart city is an urban development vision to integrate multiple Information and Communication Technology (ICT) and Internet of Things (IoT) solutions in a secure fashion to manage a city's assets. This includes data from citizens, devices and assets that is processed and analyzed to monitor and manage traffic and transportation systems, power plants, water supply networks, waste management, law enforcement, information system etc. The entire working model of the smart city includes collection of data from citizens, radars, sensors, social networking sites which are of various types and huge volume. So to deal with this type of unstructured data which is of huge volume, a big data framework can be used and to perform decentralized computing, cloud computing concepts when applied can be useful. The proposed architecture is divided into three levels namely Data acquisition, Data processing & Event handling, Applications level. The data pre-processing does the job of dimensionality reduction in the acquired data using PCA and LDA algorithms. The cleaned data is stored in HDFS. Depending on the events occurred the respective response block is called. The response block uses classification algorithms such as SVM and KNN and Optimal Path Routing algorithms such as Dijkstra's and Genetic algorithm. The architecture proposed in the paper, integrates Big Data Analytics and Cloud Computing.

The future work consists of developing the smart city applications with user friendly environment, developing the necessary algorithms which will help in creating the events and taking actions according to the events generated. The algorithms stated in the architecture are applied to the dataset and the complexity of the algorithms is calculated. The complexity of the algorithms are compared, if the complexity is low then the algorithm will best suit the large dataset used. The existing algorithms can be enhanced and the complexity can be recomputed. The architecture can be modified to use multiple or hybrid algorithms to obtain best results. The technologies such as IoT (Internet of Things), Deep Learning, Mobile Cloud computing etc can be incorporated to make the architecture robust.

REFERENCES

[1] Su, Kehua, Jie Li, and Hongbo Fu. "Smart city and the applications." "electronics, Communications and Control (ICECC), 2011 International Conference on. IEEE, 2011.

[2] Zikopoulos, Paul, and Chris Eaton. Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media, 2011.

[3] Balakrishna, Chitra. "Enabling technologies for smart city services and applications." "ext Generation Mobile Applications, Services and Technologies (NGMAST), 2012 6th International Conference on. IEEE, 2012.

[4] Costa, Daniel G., et al. "A fuzzy-based approach for sensing, coding and transmission configuration of visual sensors in smart city applications." Sensors 17.1 (2017): 93.

[5] Zhang, Bo, et al. "Learning-based Energy-Efficient Data Collection by Unmanned Vehicles in Smart Cities." IEEE Transactions on Industrial Informatics 14.4 (2018): 1666-1676.

[6] Mohamed, Nader, et al. "Unmanned aerial vehicles applications in future smart cities." Technological Forecasting and Social Change (2018).

[7] Habibzadeh, Hadi, et al. "Soft Sensing in Smart Cities: Handling 3Vs Using Recommender Systems, Machine Intelligence, and Data Analytics." IEEE Communications Magazine 56.2 (2018): 78-86.

[8] Rathore, M. Mazhar, et al. "Urban planning and building smart cities based on the internet of things using big data analytics." Computer Networks 101 (2016): 63-80.

[9] Massobrio, Renzo, et al. "Towards a cloud computing paradigm for big data analysis in smart cities." Programming and Computer Software 44.3 (2018): 181-189.

[10] Babar, Muhammad, and Fahim Arif. "Smart urban planning using Big Data analytics to contend with the interoperability in Internet of Things." Future Generation Computer Systems 77 (2017): 65-76.

[11] Hui, Terence KL, R. Simon Sherratt, and Daniel Díaz Sánchez. "Major requirements for building Smart mes in Smart Cities based on Internet of Things technologies." Future Generation Computer Systems 76 (2017): 358-369.

[12] Naranjo, Paola G. Vinueza, et al. "FOCAN: A fog-supported smart city network architecture for management of applications in the internet of everything environments." Journal of Parallel and Distributed Computing (2018).

[13] He, Xiaoming, et al. "Qoe-driven big data architecture for smart city." IEEE Communications Magazine 56.2 (2018): 88-93.

[14] Singh, Praveen Kumar, Rajesh Kumar Verma, and PESN Krishna Prasad. "IoT-Based Smartbots for Smart City Using MCC and Big Data." Smart Intelligent Computing and Applications. Springer, Singapore, 2019. 525-534. [15] Sholla, Sahil, Roohie Naaz, and Mohammad Ahsan Chishti. "Ethics aware object oriented smart city architecture." China Communications 14.5 (2017): 160-173.

[15] Varatharajan, R., Gunasekaran Manogaran, and M. K. Priyan. "A big data classification approach using LDA with an enhanced SVM method for ECG signals in cloud computing." Multimedia Tools and Applications 77.8 (2018): 10195-10215.

[16] Seng, Jasmine Kah Phooi, and Kenneth Li-Minn Ang. "Big Feature Data Analytics: Split and Combine Linear Discriminant Analysis (SC-LDA) for Integration Towards Decision Making Analytics." IEEE Access 5 (2017): 14056-14065.

[17] Xu, Xinzheng, et al. "Review of classical dimensionality reduction and sample selection methods for large-scale data processing." Neurocomputing (2018).

[18] Song, Li, et al. "A Brief Survey of Dimension Reduction." International Conference on Intelligent Science and Big Data Engineering. Springer, Cham, 2018.

[19] Yan, Shuicheng, et al. "Graph embedding and extensions: A general framework for dimensionality reduction." IEEE transactions on pattern analysis and machine intelligence 29.1 (2007): 40-51.

[20] Sorzano, Carlos Oscar Sánchez, Javier Vargas, and A. Pascual Montano. "A survey of dimensionality reduction techniques." arXiv preprint arXiv:1403.2877 (2014).

[21] Cunningham, John P., and Zoubin Ghahramani. "Linear dimensionality reduction: Survey, insights, and generalizations." The Journal of Machine Learning Research16.1 (2015): 2859-2900.