A Newfangled IoT Big Data Parallel Preprocessing Frameworkto Facilitate Quality IoT Big Dataanalytics

Mrs. I. Priya Stella Mary
Ph.D Scholar
Department of Computer Science
St. Joseph's College (Autonomous)
Tiruchirappalli – 2, India

A. Jenifer Jothi Mary
Assistant Professor
Department of Computer Science
St. Joseph's College (Autonomous)
Tiruchirappalli – 2, India

Dr. L. Arockiam
Associate Professor
Department of Computer Science
St. Joseph's College (Autonomous)
Tiruchirappalli – 2, India

***Abstract*— Theadvent of the Internet of Things is the driving force behind the recent technological revolution. Internet of things is the ubiquitous collection of internet-connected devices that collect, analyze and transform the massive amount of big data at an unparalleled rate. This enormous amount of data can be transformed into quality information by developing and deploying appropriate preprocessing techniques. The all-new IoT big data technology necessitates changes to be brought to the existing technologies. In this paper, the importance of preprocessing techniques in the IoT big data environment and also recent researches in IoT big data preprocessing techniques have been discussed. Finally, a newfangled IoT Big data parallel preprocessing framework has been proposed to convert the raw data into treasurable information thereby enabling quality IoT big data analytics to attain the full fruition of this emerging technology.**

Keywords*— *Big Data; IoT; preprocessing; analytics; Hadoop map reduce;

I.INTRODUCTION

The internet of things which is being heralded as the next technological revolution is not a new notion. Kevin Ashton, one of the pioneers had already conceived this concept in the early 2000's [3].The Internet of Things (IoT)is an indispensable part of future internet.It refers to the collection of intelligently connected devices and systems possessing unique identities, enabling seamless communication between the real and virtual world entities.The big data is described by four characteristics namely volume, velocity, variety and veracity. It is the massive volume of data,streaming in at an unparalleled rate in a variety of formats.Since it is too big,it cannot be processed using conventional processes or tools [20].IoT and big data convergence will unleash plenty of opportunities that will improve the quality of life of consumers.The internet of things generates huge amount of heterogeneous data such as machine data, sensor data, weather data, geospatial data, satellite data, RFID data, text, image, audio, video data etc [2]. The vast amount of big data produced by the IoT devices contains hidden treasure of information mixed with undesirable data.

Due to the divergent sources, the IoT big data may vary in formats and structures. Data preprocessing will no doubt play a crucial role in dealing with missing data, removing duplicate data, constructing a unified schema by integrating data from multiple sources [1]. Novel preprocessing tools and techniques are needed to prepare this sheer volume of data exploding from IoT devices to garner big insights. The major obstacles deterring the benefits are to be overwhelmed by deploying the suitable preprocessing techniques to realize the full potential of IoT big data.IoT big data preprocessing techniques should function steadfastly to improve data quality and to enable fast and accurate data mining to unravel the hidden knowledge behind the raw data.The IoT big data preprocessing framework built using the big data tools like open source Hadoop, is needed to eliminate invaluable data from the valuable ones.

The objective of this paper is to highlight the prominence of pre-processing techniques in the IoT big data environment; to discuss recent researches in IoT big data preprocessing techniques;to propose a newfangled IoT big data parallel preprocessing framework thatwill be built using Hadoop map reduce framework to preprocess the massive volume of big data generated bythe Internet of Things.

II. IOT BIG DATA PREPROCESSING

IoT big data preprocessing includes data cleaning, data integration, data transformation and data reduction. The big data generated by IoT devices will be in various formats and structures. This data may be extremely affected by undesirable elements such as outliers, noise, missing values, unreliable and

redundant data. So preprocessing becomes mandatory to impute missing values, to eliminate duplicates, to perform data integration with the intention of constructing an integrated schema, to transform the cleaned and integrated data to the standard format before storing in the database. When the preprocessing step has not been carried out properly, it will lead to misleading results. It implies that IoT big data quality is directly proportional to the IoT big data preprocessing.

At large, the IoT big data preprocessing require substantial amount of processing time. The existing techniques and algorithms are not suitable to handle high dimensional IoT big data. Digging out valuable information from the sea of IoT big data and applying appropriate preprocessing tasks that best fit the selected datasets, is a never-ending technical challenge. In view of that, Innovative algorithms and frameworks are needed to tackle this problem. Once all the preprocessing tasks have been completed, the resultant dataset will be sent IoT big data analytics.

III. LIERATURE REVIEW

Shawn R. Jeffery et al. [19] have presented an ESP (Extensible Sensor stream Processing) framework for constructing sensor data cleansing infrastructures based on spatial and temporal features to clean sensor data. The framework has also been deployed in a smart hometo prove its ability in cleaning data.Saul Gill et al. [18] have built distributed data cleaning system to clean huge amounts of streaming data. First the cleansing functions have been performed using Streams-Esper, and then the filtered data has been transformed into prediction models using R. Finally the developed models have been integrated in to the DCS to detect and remove outliers.

Mervat Abu-Elkheir et al. [1] have surveyed data management solutions for internet of things. Design primitives comprising data gathering, data management and processing of data management solutions have been discussed. The lifespan of data within an IoT system starting from data generation, collection, transmission, preprocessing to storage, has also been explained. Finally the framework for IoT data management has been proposed. Katarina Grolinger et al. [17] have summarizedthe issues and challengesin handling big data using MapReduce to aid in building better big data projects.Mugdha Jain et al. [16] have proposed an innovative approximate algorithm based on k-means to overcome the disadvantage of traditional k-means algorithm which had unknown number of iterations. The proposed algorithm has determined the number of iterations without compromising the precision thereby enabling high speed and accurate big data analytics. But the proposed algorithm can handle categorical data only after its equivalent conversion to numerical data.

Michael Shindler et al. [15] have proposed fast and precise k-means algorithm for large datasets. But better performance could be achieved through this algorithm only when the memory size is large. AlfioFerrarai et al. [12] have given a complete survey on various methods of data linking to generate a group of mappings to link object descriptions in heterogeneous data sources. Also the data linking methods have been categorized based on granularity, type and source of the evidence. Charu C. Aggarwal et al. [9] have explored the internet of things from data centric view point. Data cleaning and mining issues related to IoT big data phenomenon have also been discussed.

S.Gopal Krishna Patro et al. [7] have proposed integer scaling normalization technique by referring existing normalization techniques such as Min-Max normalization, Z-score normalization and applied it to various datasets to prove its efficiency. Ruay-Shiung Chang et al. [8] have proposed a dynamic deduplication approachto increase the storage utilization of a data center using the HDFS system. Through the application of deduplication strategy, duplicates have been eliminated to increase storage space.But the data reliability could not be ensured when the storage space has attained certain threshold.

IV. PROBLEM DEFINITION

Existing preprocessing techniques have, fine grained constraints on valid data or well-defined error models [4]. Nevertheless, for the emerging IoT big data domain, there exist no pre-defined constraints or well-defined error models. IoT big data is often voluminous, noisy, clumsy, heterogeneous, unstructured and unreliable. In fact preprocessing methods needed to handle IoT big data are fundamentally different from the existing preprocessing techniques which had been applied on small samples.Efficiency of advanced forms of data mining algorithms is directly proportional to the proficiency of upgraded preprocessing techniques.

By deploying the proposed IoT big data parallel preprocessing framework built on Hadoop map reduce framework, preprocessing tasks such as data cleaning, data integration, data transformation and data reduction will be carried out to reduce the size of the dataset to improve the processing speed, to considerably minimize the demand for data storage space and to transform the raw IoT big data in to quality data to facilitate quality analytics.

V.METHODOLOGY

The proposed research work will preprocess the selected IoT big dataset to enable quality analytics to

confront the preprocessing related issues.Existing preprocessing tools are not suitable to handleIoT big datasets so the proposed preprocessing framework will deploy renovated preprocessing techniques to handle IoT big data to do quality research.

The proposed parallel big data pre-processing framework shown in Fig.1, will be built using Hadoop map reduce framework.

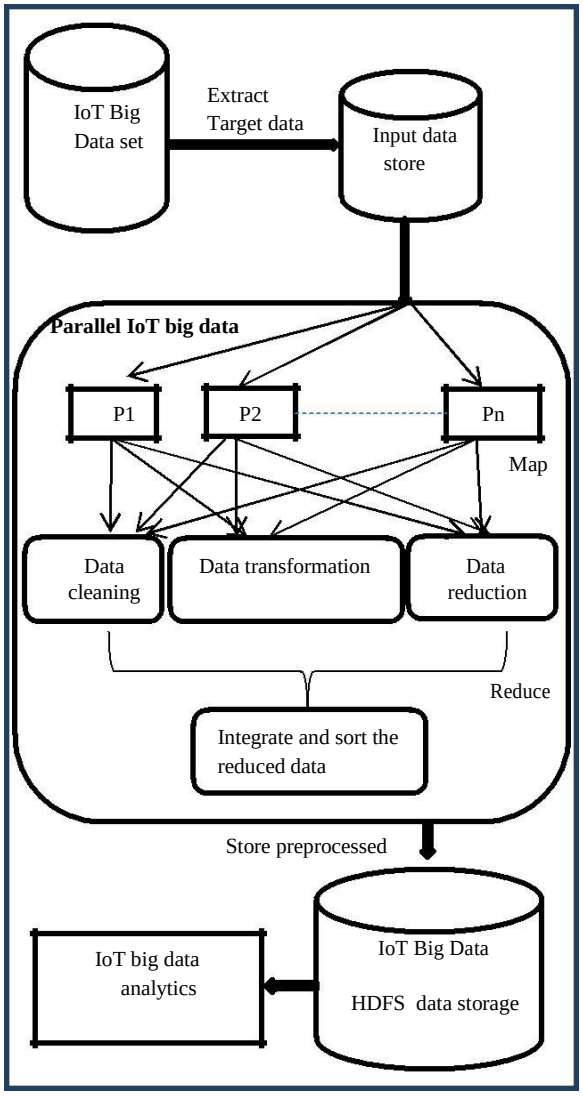

Fig.1 Parallel IoT big data preprocessing framework

The methodology to attain the objective of this research will be carried out as follows.For experimentation, weather observationsIoT datasets of Aarhus city in Denmark has been taken from city impulse dataset collections. The chosen database will be divided into 80% training set to train the proposed frameworkand 20% test set to test the accuracy of it.

From the IoT big data set, the target data will be extracted and split into small datasets and mapped to parallel preprocessors that will perform the preprocessing tasks such as data cleaning, data transformation and data reduction simultaneously and then the reduced results are integrated and stored in the HDFS data storage system for doing IoT big data analytics.

A. IoT big Data Cleaning

Data cleaning will be performed on the selected IoT big dataset to eliminate imprecise, illogical data.

*1) Filling missing values:*Missing data can occurdue tointerrupt in net connection, IoT device errors and softwareflaws [21][22].These missingdata valuesare to be handledappropriately to deterincorrect analytics particularly in case of time-series analysis. The missing values will be estimated from other existingvalues to produce a complete dataset by applying linear interpolation technique built on the big data technology, namely the Hadoop map reduce framework.The pseudo code for filling the missing values is given in Fig.2.

Pseudo code for filling missing values

```
//IDS-input data store
Step 1: Start the procedure d_clean_missing_val Step
2: read target data from IDS
Step 3: For each record in IDS
        if missing values found
                then perform linear interpolation; fill in
                missing values;
    end if ; End for
Step 4: End the procedure
```

Fig.2: Pseudo Code for filling missing values

*2) Spotting and getting rid of outliers:*Outliers hamper the task of developing precise models [23]. Some outliers will cause serious problems so that the

―garbage in, garbage out‖ rule should be followed to discard the unclean data, whereas some outliers enable to find a needle (an innovative pattern discovery) in the haystack(selected dataset) [24].Outlier detection in the IoT generated big data becomes tedious due to the growing volume of data.Because when the size of the big data increases, the amount of outliers also increases.There exists no single method to spot and eliminate outliers. Subsequently, suitable outlier detection technique for the selected dataset needs to be chosen.For this work, k-means clustering basedoutlier detection

techniquebuilt on Hadoop will be used to detect outliers inthe selected high dimensional IoT big dataset to eradicateserious data quality issues.The pseudo code for spotting and removing outliers is given in Fig.3.

Pseudo code for detecting and removing outliers

```
//k-clusters; IDS- input dataset

Step 1: Start the procedure d_clean_outliers Step 2:
select k clusters from IDS
Step 3: Assign each object to the cluster that has the closest centroid
Step 4: After all objects are assigned to the clusters, recalculate centroid.
Step 5: repeat step2 and step 3 until convergence conditions met.
Step 6: remove the objects that don't belong to any clusters
Step 7: end the procedure
```

Fig. 3: Pseudo Code for detecting and removing outliers

B. IoT big data integration

The big data integration differs from traditional data integration in terms of volume, velocity, variety and veracity [25]. The IoT devices produce and exchange massive amount of big data among other interconnected objects which makes data integration complicated. Data integration will be accomplished on the selected IoT big dataset, to consolidate it into consumable information.

*1) Removing redundancies and discrepancies:*Redundant and inconsistent data deter the ability to quicklyrespond to new requestswith in time as well as escalates the demand for data storage space [27]. To ensure accuracy and consistency, removal of duplicate information and merging of different data representations should be done.In this work, offline block-level data deduplication will be performed using HDFS and map reduce.

C. Data transformation

Data transformation, which is one of the constituting features of preprocessing, greatly enhances the subsequent analytics by converting all the values to a common form.

*1) Normalization:*Data normalization will be performed on the data that has been represented in different units [7]. In this work, simple linear transformation will be performed

for the crucial attributes on the selected IoT big dataset through the application of Min-Max normalization technique developed in Hadoop map reduce environment to transmute the data to fall within a common interval and also individual element scaling will be performed.The pseudo code for normalization is given in Fig.4.

Pseudo code for data transformation-normalization

```
/*minv-minimum attribute value; maxv-maximum attribute value; IB-dataset */
Step 1: Start the procedure dt_norm
Step 2: Select the attributes from IB to normalize; Step 3: find minv;
Step 4: find maxv; Step 5:
set range;
Step 6; for each attribute value
        do min-max normalization ;
end for;
step7: if pre-defined boundary modification needed go to
        step 5;
    end if; step8 :
End.
```

Fig. 4: Pseudo Code for normalization

D. Data reduction

Massive volume of data entails enormousprocessing time.Data reduction considerably curtailsthe processing time by reducing the dataset without impacting analytical results.

1) Dimensionality reduction: Data columns with too many missing values are improbable to possess valuable information. In this work, dimensionality reduction will be implemented by eliminating data columns with lots of missing values.This can be performed by removing the data columns with number of missing values greater than a calculated threshold for the selected dataset [26]. The pseudo code for dimensionality reduction is given in Fig.5.

Data scientists have articulated that 80% of data work will be accomplished by performing efficient data preprocessing tasks.Hadoop is the best platform for executing all the pre-processing tasks efficiently and in a distributed manner over big datasets, using map-reduce [17]. Through the implementation of this innovative framework, the processing speed will be improved, demand for data storage space will be considerably reduced and the raw IoT big data will be

Pseudo code for data transformation (dimensionality reduction)

```
/* thr_val -threshold value mis_ratio -
missing value ratio */
Step 1: Start the procedure dt_reduction
Step 2: Select the column(s) with too many missing values;
Step 3: Count the number of missing values; Step
4: Calculate missing value ratio;
Step 5: Set the threshold;
Step 6: if mis_ratio>thr_val then remove
            the column(s);
        else
            retain the column(s); Step
7: End.
```

Fig. 5: Pseudo Code for dimensionality reduction

converted into quality data to facilitate quality analytics.

VI.CONCLUSION

The rapidly evolving Internet of Things has spurred the tremendous influx of big data,which is in an unrefined state. Driving meaningful insights out of it is a herculean task. Consequently, efficient preprocessing tools and techniques are needed to filter the unwanted data and deliver valuable information. In this paper, a newfangled IoT big data parallel preprocessing framework has been proposed to preprocess the enormous amount of big data emanating from the Internet of Things. This framework will greatly enhance the processing speed; significantly limit the demand for data storage space and transform the raw IoT big data in to quality data to facilitate quality analytics so as to tap the full potential of the IoT big data technology.

REFERENCES

[1] Abu-Elkheir, Mervat, Mohammad Hayajneh, and Najah Abu Ali. "Data management for the internet of things: Design primitives and solution." *Sensors*, Vol 13, No. 11, 2013, pp. 15582-15612.

[2] Zhang, Chunguang, GuangpingZeng, Hongbo Wang, and Xuyan Tu. "Analysis on Data Mining Model Objected to Internet of Things", *International Journal of Advancements in Computing Technology*, Vol 4, No. 21, 2012.

[3] Mattern, Friedemann, and Christian Floerkemeier. "From the Internet of Computers to the Internet of Things‖, *From active data management to event-based systems and more*, 2010, pp. 242-259.

[4] Labrinidis, Alexandros, and H. V. Jagadish. "Challenges and opportunities with big data", *Proceedings of the VLDB Endowment*, Vol 5, No. 12, 2012, pp. 2032-2033.

[5] Souza, Alberto MC, and Joseé RA Amazonas. "An Outlier Detect Algorithm using Big Data Processing and Internet of Things Architecture." *Procedia Computer Science, Vol* 52, 2015, pp. 1010-1015.

[6] Saleem, Khalid, and ZohraBellahsene. "New Challenges in Data Integration: Large Scale Automatic Schema Matching",2007.

[7] Patro, S., and Kishore Kumar Sahu. "Normalization: A Preprocessing Stage", *arXiv preprint arXiv:1503.06462* , 2015.

[8] Chang, Ruay-Shiung, Chih-Shan Liao, Kuo-Zheng Fan, and Chia-Ming Wu. "Dynamic Deduplication Decision in a Hadoop Distributed File System", *International Journal of Distributed Sensor Networks*, 2014.

[9] Aggarwal, Charu C., Naveen Ashish, and Amit P. Sheth. "The Internet of Things: A Survey from the Data-Centric Perspective", 2013, pp. 383-428.

[10] Jagdale, Ashish R., Kavita V. Sonawane, and Shamsuddin S. Khan. "Data Mining and Data Pre-processing for Big Data"

[11] Barnaghi, Payam, Wei Wang, Cory Henson, and Kerry Taylor. "Semantics for the Internet of Things: early progress and back to the future." *International Journal on Semantic Web and Information Systems (IJSWIS)*, Vol 8, No. 1, 2012, pp. 1-21.

[12] Hasan, Souleiman, and Edward Curry. "Approximate semantic matching of events for the Internet of Things", *ACM Transactions on Internet Technology (TOIT)*, Vol 14, No. 1, 2014.

[13] Ding, Guoru, Long Wang, and Qihui Wu. "Big Data Analytics in Future Internet of Things." *arXiv preprint arXiv:1311.4112*, 2013.

[14] Che, Dunren, MejdlSafran, and ZhiyongPeng. "From big data to big data mining: challenges, issues, and opportunities", *In Database Systems for Advanced Applications*, 2013, pp. 1-15.

[15] Zhao, Weizhong, Huifang Ma, and Qing He. "Parallel k-means clustering based on mapreduce" *Cloud Computing*, 2009, pp. 674-679.

[16] Shindler, Michael, Alex Wong, and Adam W. Meyerson. "Fast and accurate k-means for large datasets‖, *Advances in neural information processing systems*, 2011, pp. 2375-2383.

[17] Grolinger, Katarina, Michael Hayes, Wilson Akio Higashino, Alexandra L'Heureux, David S. Allison, and Miriam Capretz. "Challenges for mapreduce in big data", *Services (SERVICES), 2014 IEEE World Congress*, 2014, pp. 182-189.

[18] Gill, Saul, and Brian Lee. "A Framework for Distributed Cleaning of Data Streams." *Procedia Computer Science*, Vol 52, 2015, pp. 1186-1191.

[19] Jeffery, Shawn R., Gustavo Alonso, Michael J. Franklin, Wei Hong, and Jennifer Widom," Declarative support for sensor data cleaning", Pervasive Computing, 2006, pp. 83-100.

[20] Cai, Li, and Yangyong Zhu. "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era." *Data Science Journal*, Vol 14,2015.

[21] Yan, Xiaobo, WeiqingXiong, Liang Hu, Feng Wang, and Kuo Zhao "Missing Value Imputation Based on Gaussian Mixture Model for the Internet of Things", *Mathematical Problems in Engineering*, 2015.

[22] Kaiser, Jiri. "Dealing with Missing Values in Data",*Journal of Systems Integration*, Vol 5, No. 1, 2014, pp. 42-51.

[23] Liu, Boyuan, Wenhui Fan, and Tianyuan Xiao. "A Fast Outlier Detection Method for Big Data", *AsiaSim*2013, pp. 379-384.

[24] Patel, Vaishali R., and Rupa G. Mehta "Impact of outlier removal and normalization approach in modified k-means clustering algorithm" , *IJCSI International Journal of Computer Science*, Vol 8, no. 5, 2011.

[25] Dong, Xin Luna, and DiveshSrivastava "Big data integration", Data *Engineering (ICDE)*, 2013, pp. 1245-1248.

[26] Rosaria Silipo.(2015, March 30) Dimensionality Reduction: Removing Data Columns with too many Missing Values [Online]. Available: http://www.knime.org/blog/seven-techniques-for-data-dimensionality-reduction.

[27] Mishra, Deepak, and Sanjeev Sharma, "Comprehensive study of data de-duplication‖.