An Efficient Data Stream Analytics Model for Real Time Internet of Things (Iot) Applications

K. Kranthi Kumar, E Ramaraj

Abstract*: *Internet of Things (IoT), data analytics is supporting multiple applications. These numerous applications try to gather data from different environments, here the gathered data may be homogeneous or heterogeneous, but most of the data collected from multiple environments were heterogeneous, the task of gathering, processing, storing and the analysis that is being performed on data are still challenging. Providing security to all these things is also a challenging task due to untrusted networks and big data. Big data management in the ever-expanding network may rise several non-trivial concerns on data collection, data-efficient processing, analytics, and security. However, the above said scenarios depends on large scale sensor deployed. Sensors continuously transmit data to clouds for real time use, which can raise the issue of privacy disclosure because IoT devices may gather data including a kind of sensitive private information. In this context, we propose a two-layer system or model for analyzing IoT data, collected from multiple applications. The first layer is mainly used for gathering data from multiple environments and acts as a service-oriented interface to ingest data. The second layer is responsible for storing and analyses data securely. The Proposed solutions are implemented by the use of open source components.

Keywords: **Data, Data Stream, Spark, Analytics, IoT.**

I. INTRODUCTION

The Internet of Things is an advanced communication technology that visions near future, in which the things of everyday life are furnished with software and hardware components such as microcontrollers, transceivers, sensors and specially developed protocols to provide digital communication, among nodes of various networks with the users, becoming integral part of the internet of things. The fundamental idea of this technology is to connect all the things around the world and it is a rapidly developing technology in the scenario of modern wireless telecommunications [1]. The Concept of Internet of Things made internet even more attractive and universal. The Development of smart applications and providing ease of access to these applications along with the wide variety of interaction of devices across the world may made IoT as effective communication technology. The Applications developed by IoT may make use of enormous amount of data and generates variety of data streams to provide services to people, industries, public administrations and all other domains. Indeed, the Concept of Internet of Things discovers applications in different domains called the heterogeneous domains such as industry automation, health care automation, home automation, and many others [2]. There exist different types of computing paradigms but cloud computing is an advanced paradigm that enables users to use shared resource pool of cloud resources such as storage, accessing, processor and applications in an on –demand manner. Integration of cloud with IoT and cloud computing capabilities such as data gathering, data processing data storage, data analysis and security over all of these along with data retransmission facility offered by cloud computing for instance, IoT Sensors first gather data and transmit it to the gateways which then transmit to the cloud for store, process and analysis and it then transmit the data to user on demand. During the entire data transmission process if data transmission is failed the data is retransmitted until they are successfully delivered to the specified users. This facility of cloud computing is fascinating the attention of both academia and industry [3]. Internet of Things many applications enable smart initiatives by connecting with big data and analytics, most of the IoT applications may not only concentrate on managing diverse things but also on mining the data collected from IoT devices. Data collection tools of IoT devices are often sensor-fitted traditional protocols like MQTT-Message Queue Telemetry Transport Protocol, XMPP and all other. Survey on Internet of Things shows that the number of things or devices Inter-connected to IoT is expected to research 50 billion by 2020 as shown in Fig. 1, with Iot lot of opportunities have created that can help to increase revenue, reduce cost, huge amount of data generation and all other.[4]

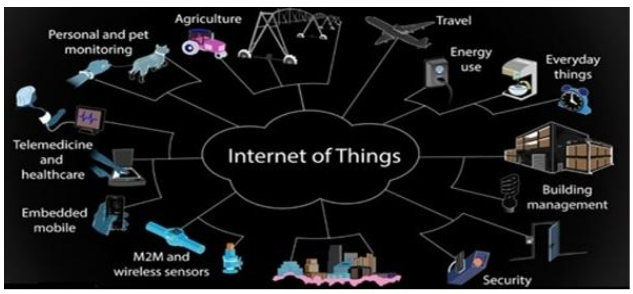


Fig. 1. Services of IoT

To get benefits from IoT Industries must create a platform this can be used to do all the things like collecting, managing and analyzing big data. Big data made by this platform is generated by the sensors used by the system, can be efficient and cost –effective manner and must be scalable.

Revised Manuscript Received on February 07, 2020.

* Correspondence Author

Mr. Kranthi Kumar, Research Scholar, Department of Computer Science, Alagappa University, Karaikudi, Tamil Nadu 630003, India

Dr. E.Ramaraj, Professor and Head, Department of Computer Science, Alagappa University, Karaikudi, Tamil Nadu 630003, Indi

In this connection supporting big data can generate massive volume of data, utilizing this massive volume of data, integration of heterogeneous data becomes very important.
Specially developed data analytics tools can be used by these industries to get above said benefits from IoT, and this integration of heterogeneous fields may create vision on lots of research also this paper mainly focuses on big data collection and analytics performed on it. Big data management in the ever-expanding network may rise several non-trivial concerns on data collection, data-efficient processing, analytics, and security.
The contribution of this paper are as follows:

- Review on current literature in IoT.
- Providing two layered architectures for collecting and analysing data.
- Use of Open source components for Big data management in IoT.
- Integration of IoT and Big Data
- Providing key requirements for integrating big data and IoT
- List out the open research challenges and the vision of big data analytics in IoT as future research areas.

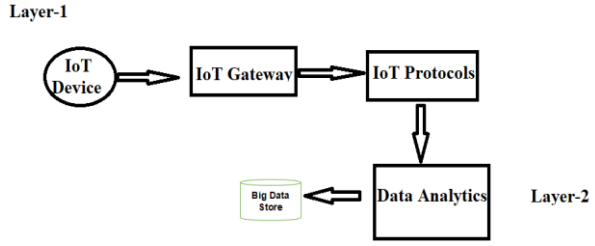


Fig. 2. Proposed Architecture

In this connection we propose a two layered architecture as shown in figure 2. The first layer data collecting is mainly responsible for the collection data from multiple sources, and transmit it to the second layer analytics where data is analyzed in order to extract knowledge data from it that will be useful for real time IoT applications.

The remainder of this paper is organized as follows:

Section II Related Work aims to focus on research work in big data with IoT. Section III demonstrates the first layer of data collecting. Section IV explains about analytics of our proposed architecture. Next section V describes the implementation of our proposed architecture. Finally, we mention conclusion and future work as section VI and VII.
Some of the terms used in this paper are

Heterogeneous Data: Data in various types and formats.

MapReduce: A programming model and connected implementation for processing and generating large volume of data with an equivalent, disseminated procedure on group. Its framework provides filtering, sorting and summary with two functions Map and Reduce.

Authentication: Authentication is the act of verifying a claim, such as the identity of a computer system customer. In distinction with proof of identity, the act of representing a person or thing's characteristics, authentication is the process of verifying that uniqueness.

Hadoop: Apache Hadoop could be a mixture of ASCII text file software package utilities that facilitate employing a network of the many computers to unravel issues involving large amounts of information and computation. It provides a software framework for distributed storage and process huge data using MapReduce programming model.

Spark: Apache Spark is an open-source distributed general-purpose cluster-computing framework. Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance.

II. RELATED WORK

I. MapReduce: MapReduce is one of the most popular data analytics frameworks created by Google to analyse Google's own big data. In order to enhance system performance and measurability MapReduce and its open source components such as Hadoop has been used widely to support massive calculations over huge information sets or data sets. MapReduce is the basis for other open source data analytic framework like Hadoop. Social media and ecommerce often use MapReduce to analyses large data [5].

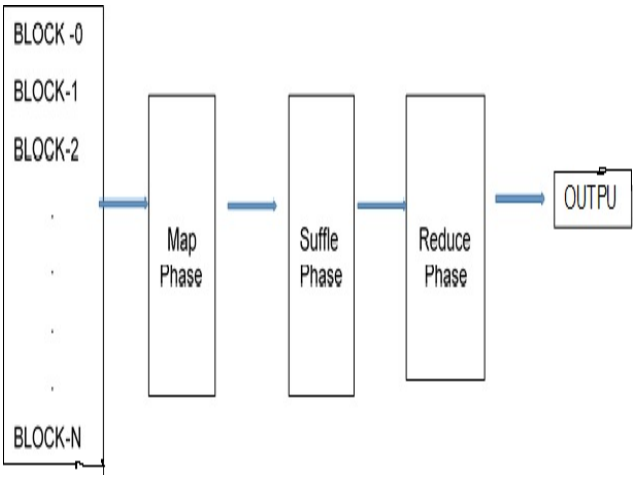


Fig. 3. MapReduce

Designers specify calculation using two functions Map and Reduce, Shuffle phase is specially used for sorting data, MapReduce is a high-level data analytics model for clustering all the data generated by its phases. MapReduce got lots of popularity for its easy programming interface, outstanding performance. When it comes to the behavior of the framework MapReduce is a distributed processing system, google uses an open source component Hadoop for implementing MapReduce [6]. The above figure 3 shows the implementation of MapReduce as it shows there are three phases, one is Map Phase it takes input from blocks for mapping and send it to the second phase for Shuffle and it can again send it to the third phase Reducer, finally it generates output. the description and pictorial representation of Hadoop is shown in fig 4

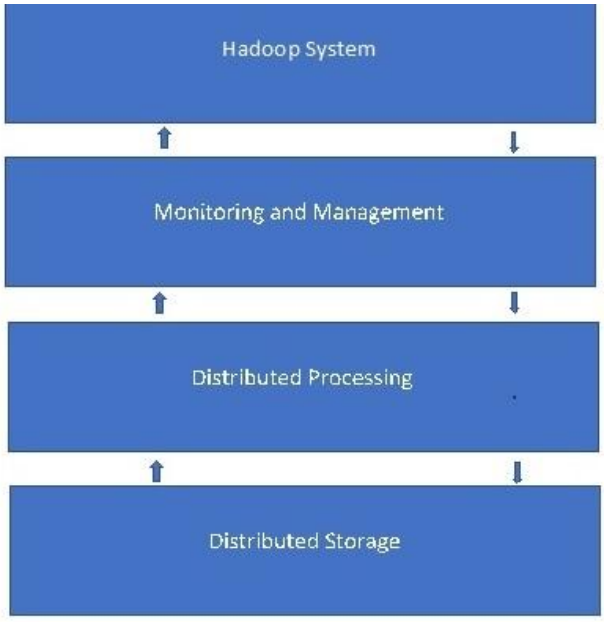


Fig. 4. Hadoop

II.MapReduce Authentication: Authentication in MapReduce is a major challenge in the perspective of data analytics scenario, there are some situation as mentioned here MapReduce applications are accessed via internet, Data in MapReduce are divided and stored on a set of distributed and common nodes, once executed job may be executed multiple times. [7]

III. MAPREDUCE FRAMEWORK FOR IOT:

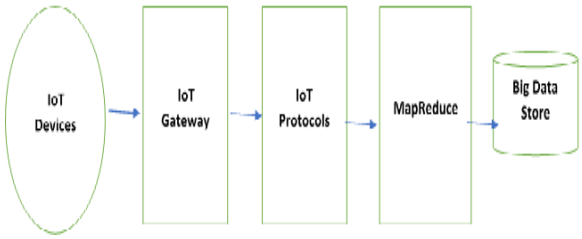


Fig. 5. Analytics Framework for IoT

MapReduce with Hadoop has given excellent performance for data analytics. Even though MapReduce is good at data analytics for large volume of data but it has some limitations like It is not suitable for running machine learning algorithms, every iteration needs creation of new Mapper and Reducer, and reading same disk again and again. There are two broad classifications of data analytics in IoT one is Batch processing and the other is Event processing, batch processing is the technique used by MapReduce. Data analytics in Internet of Things specifies data inspecting and studying, the outcome of this level produce knowledge data that is useful to public administrations, healthcare, people, industries, companies and others. Hadoop is an Apache open source system written in java that permits disseminated preparing of enormous datasets crosswise over bunches of PCs utilizing straightforward programming models. The Hadoop system application works in a domain that gives dispersed capacity and calculation crosswise over bunches of PCs. Hadoop is intended to scale up from single server to a large number of machines, each offering nearby calculation and capacity.[8] In this

IV. DATA COLLECTION AT FIRST LAYER

There are numerous sensors available to collect data. Sensors play vital role in the scenario of gathering raw data. The vivid articles are considered as the structure squares of IoT. Diverse kind of items where sensors are installed, create and share a lot of information. This ongoing spilling information accumulated utilizing sensors [9]. This information is utilized for constant basic leadership and offline information investigative. A class of uses that produce information that have worth regardless of whether the handling doesn't happen progressively. This natural crude information required AI or sign preparing calculation that wires information from various brilliant items. Likewise expect that we want to execute examination calculation that wires the information from remote sensors to the handling focus situated in the cloud. Heterogeneous sensor information gives a few administrations to the individual. We have sensors on a fleet of the truck of our client and gather client acoustic information from the motors. Any looming issue will cause the sound info change, when you can complete a preventive upkeep. In this way, heterogeneous versatile sensors information gathering is one of the most significant research difficulties. In the present period, distributed computing is picking up heaps of enthusiasm for a few areas by handling enormous information [10].

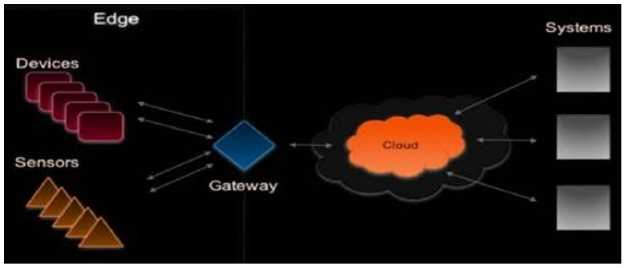


Fig.6. IOT Gateway

Where information is gathered from a few sources, for example, sensor systems, interpersonal organizations, and vehicles. There is still extension to address security worry of information gathered from above sources to cloud server farm. There is need a typical design to help the information gathering of information from sensors to cloud. System conventions are the foundation of any correspondence framework which pursues certain Quality of Service (QoS) for every correspondence application. With the brisk improvement of introduced advancement, remote framework system turns out to be dynamically fit, with its topological structure and correspondence ending up being increasingly unusual. As another remote advancement, ZigBee mastermind has the relative issue to the others. It is a remote development, which is an equipment of negligible exertion remote framework. It is a remote framework show made by the ZigBee Alliance in light of IEEE 802.15.4 Standard. In the headway strategy of ZigBee remote framework, ZigBee in perspective on Stack show accept a piece of the skeleton, and its portions and different levelled structure give an extraordinary foundation to the improvement of a compelling application structure.

Regardless, standard headway procedure has not been gathering the essentials, so it is believed that a suitable remote framework improvement system in light of embedded structure can shorten the structure improvement cycle, just as reduction the improvement costs and addition structure quality.[11]

The Gateway shown in figure 6 describes about data collection instance, here some major components like sensors, devices, gateway, cloud, local DB, and Systems make the system of IoT Gateway. The task of each component is explained below ,first sensors provide raw data it seems as values by taking particular action from the devices, so both sensors and devices interact with gateway ,it acts as a bridge between devices and cloud .The cloud play major role here it takes data from gateway, process it and transform it to systems upon the request made by systems ,it also send it to the local dB for local storage .

IoT Protocols: Various protocols were designed for implementing internet of things applications. The classification of protocols based on its application and usage is taken in consideration as follows

Messaging Protocols:

MQTT-Message Que telemetry transport, CoAP-Constrained Application Protocol, AMQP-Advanced Message Que protocol, HTTP-Hyper Text Transfer Protocol

Communication Protocols:

6LOWPAN, ZigBee, Bluetooth LE, RFID, NFC, SigFox.

Security Protocols: IP Security, Wireless Hart etc

V. DATA ANALYTICS AT SECOND LAYER

Google MapReduce is a programming model and a product structure for huge scale appropriated processing on a lot of information. Figure 3 outlines the elevated level work flow of a MapReduce work. Application designers determine the calculation as far as a guide and a lessen work, and the fundamental MapReduce employment planning framework naturally parallelizes the calculation over a bunch of machines. MapReduce gains notoriety for its basic programming interface and incredible execution when actualizing a huge range of utilizations. Actually, for data analytics many research articles use MapReduce but it has some disadvantages as we explained in the above. MapReduce gives an institutionalized system to executing enormous scale circulated calculation, to be specific, the huge information applications. Be that as it may, there is a confinement of the framework, i.e., the inefficiency in gradual preparing. Gradual handling alludes to the applications that steadily develop the information and persistently apply calculations on the contribution to request to produce yield. There are potential copy calculations being performed in this procedure. Be that as it may, MapReduce doesn't have the system to distinguish such copy calculations and quicken work execution. Inspired by this perception, use of Hadoop gave an excellent benefit from it. Distributed computing technologies may make many advanced innovations by solving any distributed problems using various algorithms, MapReduce is used for Distributed computing [12] Spark Streaming, which may be a big data platform which will efficiently process an enormous amount of data in order that we will monitor the network status in real time and is robust enough so as to suffer a failure without aborting the whole monitoring process. Big data platforms, like Hadoop and Spark, provide an efficient way of processing an enormous amount of knowledge. for instance, the MapReduce model and its open-source version, Hadoop [13], are widely adopted by the large data analytics community thanks to their simplicity and simple programming [14]. However, the intermediate data of Hadoop are stored on disk (which usually has poor I/O performance); therefore, there'll be dramatic production degradation for algorithms requiring many iterations. to enhance the performance of Hadoop, in-memory computing methods, such as Apache Spark [15], have been proposed. The intermediate data in Spark are stored in Resilient Distributed Datasets (RDD), which are cached in memory; therefore, data are often processed much faster as compared with Hadoop [16]. Some offline Internet monitoring systems have used big data platforms to enhance their processing efficiency. However, only a couple of studies have focused on online network monitoring. Both Hadoop and Spark are supported execution, which is suitable for offline data analysis. execution is applied to process large datasets, where operations on multiple data items can be batched for efficiency [17].

Tools and Technology Used

Sno	Tools and Technology Used	
1	Spark	Open source distributed general purpose cluster framework
2	Cassandra	Database
3	Kafka	Used for real time stream analysis
4	JDK	The Java Development Kit
5	Maven	Automation tool
6	Zookeeper	A service for DS
7	Spring Boot	Stand-alone App Creator

This needs input file to be readily accessible when the calculation begins in order that all of the info is often simultaneously processed. Online Internet traffic monitoring resembles a stream analytics problem, where the input is an unbounded sequence of knowledge. Although MapReduce doesn't support stream processing, it can partially handle streams employing a technique referred to as micro-batching. Here the stream is treated as a sequence of small batch data chunks. At short intervals, the incoming stream is packed to a piece of knowledge and is delivered to the batch system for processing [18]. Spark, for instance, has provided the Spark Streaming library to support this system. additionally, other platforms exist and are inherently designed for giant data streams, like Apache Storm and S4, where data are processed through several computing nodes. Each node can process one or more input stream(s) and generate a group of output streams. Data are going to be processed as soon as they arrive.

Fig.7.Traffic Scenario

Hours and Respective Value:

SNO	Hours	Value
1	0h	5.176531
2	1h	0.347362
3	2h	0.99331
4	3h	-0.31845
5	4h	-0.28967
6	5h	0.595623
7	6h	0.683265
8	7h	0.812636
9	8h	0.523157
10	9h	0.689745
11	10h	0.435689
12	11h	0.235479
13	12h	0.785217
14	13h	0.325416
15	14h	0.892313
16	15h	0.987163
17	16h	0.123659
18	17h	0.589623

VI. IMPLEMENTATION

For data analytics purpose in this system we have taken real time traffic scenario. Traffic Monitoring is one of the big problems very often faced in many developed countries. As it is mentioned in previous section data analytics layer here Apache Spark has been taken to analyses traffic data.

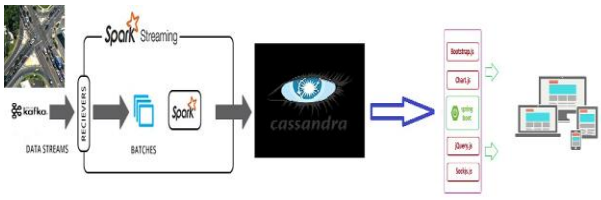


Fig.8.Spark Data Analytics

Data processing in Apache Spark is really different when it is compared with other data analytics tools and that to it supports data input as batch, at first data is organized as batches and then it is converted to stream. The stream of data is transformed to Cassandra for storage. Cassandra acts as a Database for storing data, the overall scenario is represented in the below figure.

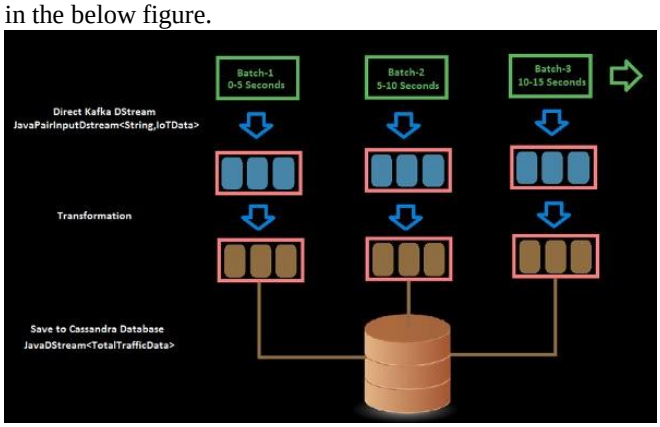
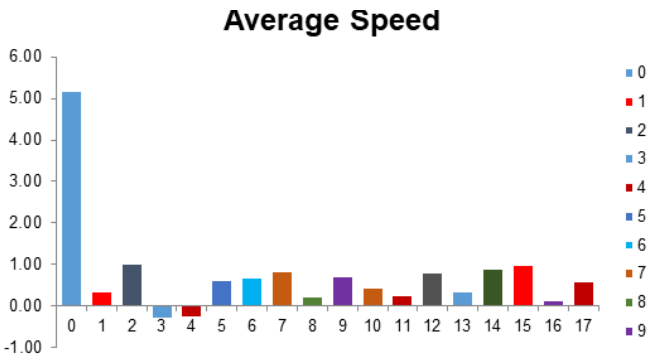


Fig.9. Data Processing and Storage

We have taken results from spark as shown in the below table it shows hours and its respective vehicle speed is mentioned in hours and the graph is drawn for displaying average vehicle speed.

Results:



VII. CONCLUSION

In this paper, we have shown and employed a solution for ingesting, analyzing and connecting heterogeneous data streams in order to provide an efficient, ascendable and consistent solution for real time traffic management. A two layered methodology has been taken here one is meant for data collection and other is for analytics. We have used Apache Spark for this.

FUTURE WORK

In future we have a goal to use many real time IoT application for analyzing and showing their performance toward making smart world.

REFERENCES

1. L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," Comput. Netw., vol. 54, no. 15, pp. 2787–2805, 2010.
2. Andrea Zanella,Nicola Bui,Angelo Castellani,Lorenzo Vangelista and Michele Zorzi " Internet of things for Smart Cities,"., vol. 1, no. 1, pp. 2327–4662, 2014.
3. Halah Mohammed al-kadhim and hamed s.al-raweshidy "Energy Efficient and Reliable Transport of Data in Cloud-Based IoT,"., vol. 7, pp. 2169–3536, 2019.

4. E. Ahmed et al., ''The role of big data analytics in Internet of Things,'' Comput. Netw., vol. 129, pp. 459–471, Dec. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128617302591
5. Nan Zhu, Xue Liu*, Jie Liu, and Yu Hua, "Towards a Cost-Efficient MapReduce: Mitigating Power Peaks for Hadoop Clusters,"., vol. 19, no. 1, pp.24-32 1007–0214, 2014.
6. Yaxiong Zhao' Jie Wu, and Cong Liu, "Dache: A Data Aware Caching for Big-Data Applications Using the MapReduce Framework,"., vol. 19, no. 1, pp.39-50 1007–0214, 2014.
7. Ibrahim lahmer and ning zhang, "Towards a virtual Domain Based Authentication on MapReduce," vol. 4, 2016, 2558456.
8. Available Online: www.wikipedia.org/wiki/Apache_Hadoop.
9. "Fuzzy Assisted Event Driven Data Collection from Sensor Nodes in Sensor-Cloud Infrastructure," Cluster, Cloud and Grid Computing (CCGrid), S. S. Bhunia, J. Pal and N. Mukherjee," 2014 14th IEEE/ACM International Symposium on, Chicago, IL, 2014, pp. 635-640
10. "A Dynamic Key-Length-Based Efficient Real-Time Security Verification Model for Big Data Stream." D. Puthal, S. Nepal, R. Ranjan, and J. Chen." DLSeF: ACM Transactions on Embedded Computing Systems (TECS), Vol. 16, no. 2, pp. 51, 2016.
11. " Sensors Data Collection Architecture in the Internet of Mobile Things as a Service (IoMTaaS) Platform", Prasenjit Maiti, Bibhudatta Sahoo,Ashok Kumar , Suchismita Satpathy, Conference Paper · February 2017 DOI: 10.1109/I-SMAC.2017.8058245.
12. "Spark-Based Large-Scale Matrix Inversion for Big Data Processing: JUN LIU1, (Member, IEEE), YANG LIANG1, AND NIRWAN ANSARI2, (Fellow, IEEE)- Digital Object Identifier 10.1109/ACCESS.2016.2546544.
13. http://hadoop.apache.org/
14. "Trends in big data analytics"- vol. 74, no. 7, pp. 2561–2573, 2014. K. Kambatla, G. Kollias, V. Kumar, and A. Grama, J. Parallel Distrib. Comput.
15. Apache Spark, http://spark.apache.org/
16. " Cluster computing with working sets, in Proc.2nd USENIXConf. HotTopicsinCloudComputing, "Boston, MA, USA, 2010, M.Zaharia,M.Chowdhury,M.J.Franklin,S.Shenker,and I. Stoica, Spark.
17. "Data-intensiveapplications, challenges, techniques and technologies: A survey on big data "C.L.P.ChenandC.Y.Zhang, , Inf. Sci., vol. 275, pp. 314–347, 2014.
18. "Towardsreal-time and streaming big data, Computers, vol. 3, no. 4, pp. 117– 129, 2014. S.Shahrivari,Beyondbatchprocessing

AUTHORS PROFILE

Mr.K. Kranthi Kumar, is currently Rsearch Scholar of Computer Science Department Alagappa University Karaikudi,Tamil Nadu He has 7 years of Teaching Experience his Research interests are Data mining,Network Security,Big Data and Internet of Things.

Dr. E.Ramaraj, Professor and Head of The Computer Science Department Alagappa University karaikudi Tamil Nadu .He Has 32 years of Teaching Experience and 17 years of Research Experience,his Research interests are Data mining,Network Security,Big Data and Internet of Things.