DOI: http://dx.doi.org/10.26483/ijarcs.v9i3.6105

ISSN No. 0976-5697

Volume 9, No. 3, May-June 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

EFFICIENT BIG DATA ANALYSIS USING HADOOP FRAMEWORK FOR SENSOR NETWORK DATA

Sahana R
Assistant Professor, Department of Computer Science & Engineering
MVJCE, Bangalore, India

Abstract: In the era of rapid development of the technology, Wireless Sensor Networks is also a field that is been developing and already reached a stage where thousands of nodes in a network will be present and they store and also deliver a vast amount of data. We are living in the age where an explosive amount of data is being generated every day. Majorly the data from sensors, mobile devices, social networking websites, scientific data & enterprises are contributing to a maximum extent for this huge explosion of data. The major problem will be the amount of data collected, its storage and processing. So in an order to store, load and process the large scale sensory data we propose a paper that enable the efficient and effective analysis of the Big Data and also handling the large volumes carried out on the Hadoop platform.

Keywords: Big data, Hadoop, Wireless Sensor Network

INTRODUCTION

In wireless sensor networks, large scale data loading and processing have several issues hence creating a barrier and huge amount of data collected will be lost or not processed properly. Further, effective data handling should go through dynamic services that do not exist in conventional processing mechanisms using normal structured query language, tiny DB and tiny OS.

To overcome the problem of data handling and processing, a novel approach of Hadoop framework can be processed for large scale sensory data that can also be said as the big data.

By using the traditional data management tools, it is difficult to process the huge and complex sets of huge data. The data that is being generated is increasing in a very tedious manner year by year.

So, due to this fact many problems related to the management of data and data processing tasks with respect to time are generated.

The data that is beyond the storage capacity and also beyond the processing power can be said as the big data. For example, in social media like Face book daily 500TB of data is generated daily, in Google 24PB of data is processed, 50million tweets in Twitter, 20 hours of video are uploaded every minute in YouTube. The generation is growing at 40% compound annual rate, reaching 50ZB till 2020. So it may be said as the big data. The interesting point is how these social media's and platforms are able to quickly deal with such large quantities of information.

Essence of Big Data:

Volume - Amount of existing data to be processed. Velocity - Streaming data
Variety - Structured, unstructured, text, multimedia etc.

Google labs developed an algorithm, that allowed for the vast data calculation to be broken down into smaller forms of data and then processed on the computer systems, then the calculation were done be brought back together to produce the resulting data. They called that as the map reduce algorithm which is a major part of the Hadoop platform. Hadoop is the solution for the big data. It is program that provides a framework for distributing and running applications on cluster of servers that was found out from apache software foundation.

The processing of large amount of data from several different heterogeneous sources are done by considering some approaches which are novel. The major problems will include the strategies for searching the data, dissemination of data, analysis of data and even the visualization and the processing procedures. In a traditional manner, the big data deals with the centralized computing resources which are massive, which is also high performance computational centers and large storage systems.[1]

A base station or a high performance network is present, where the ordinary nodes that collect the data is passed on to the sink node. Sink node analyses and processes the data sensed by the sensors.

RELATED WORK

Gurpreet Singh Bedi, Ashima Singh (2014) [1] have proposed that the data is exceedingly large day by day.

The necessity to analyze and process the vast data is there in some organizations. A single machine processing huge data is impossible task. In order to avoid this we have used apache Hadoop distributed file system for storage and analysis. Here some work is done on the map reduce application on health sector dataset. To map and reduce the huge data sets the results of the map reduce application framework can be used. By increasing the dataset size it will become a main problem to check the character of the map reduce.

Understanding the performance of the map reduce is the major analysis factor. Increase in the execution time linearly with the size of dataset is expected, but our analysis says sometimes the execution time varies non-linearly with the increase in the dataset size. The experimental result shows that time will be distinguished by the scaling of the datasets. The aim of the project was to analyze the huge volume of hospital data. The data about all the patients are collected and is examined. Next to this process, by using he map reduce the ratings of the hospital is been computed. By using these ratings, to check the behavior of map reduce application a graph is been drawn. Firstly, to check the compatibility between the scalability of the data and the execution time, an experiment is carried out. To check the execution time of map reduce by using single node and also multimode cluster of hadoop an experiment is carried out. Non-linear relationship in between the size of the dataset and the time is observed.

*Bi*ng Tang and Yu Wang (2012) [2] have proposed a large-scale wireless sensor networks (WSNs), will be having a necessity for high performance of the sensory data processing, where commonly there will be having limited processing power and storage capacity of sensor nodes. This specifies the connection between the wifi sensor group and cloud-based storage devices and even some other processing facilities. The idea suggests the thinking behind allocated databases to retail store sensory files and also map reduce selection product with regard to large-scale sensory files parallel processing. In this prototype of huge sensory files processing technique, Hadoop allocated file technique and also hbase utilized with regard to sensory files storage devices, and also Hadoop map reduce is used with regard to files processing program delivery structure. The style and also execution of the technique are usually referred to in depth. This simulation involving surroundings heat range surveillance request is needed for you to examine your feasibility along with reasonableness of the system, that also shows which it substantially helps the info running ease of WSNs. The technologies involved in the data management are data aggregation, storing, querying and accessing. Particularly there are three main data storing strategies, they are as follows,

Centralized storing: The data that is being collected by the sensor nodes are stored in a single centralized storage system called the base station, where storing, accessing and the processing of the data takes place.

Distributed storing and indexing: The data is stored in a distributed manner in a network. And also for having highly efficient query a data index is built.

Locally storing: The data is stored in sensor nodes locally, having lower query efficiency and lower communication overhead. It implements the grid processing technique. Grid processing produces the virtual supercomputers by using extra processing sources geographically sent out with web, and also processing sources are separate processing groups which might be not really in a single domain. This processing haves the internet calculation and even the storage devices. This products and services architectural mastery is designed in order to define the latest common and also regular architectural mastery pertaining to grid-based program. Various pursuits around the globe understand the interconnection associated with sensor nodes together with g infrastructure of grid processing. Sensor grid is actually a real cross architectural mastery which in turn integrates instant sensor cpa networks together with the infrastructures of grid to have real-time physical information variety and the also the sharing associated with computing and also storage devices sources pertaining to physical information control and also managing**.**

Han hui ,Yong gang, Tat-seng chual, xuelongli (2013)[3] have proposed that normally the data explosion takes place from the fields such as hospitality and health care, scientific sensors, user- generated data through Internet and mobiles and the financial organizations in past several years. For this evolving trend, a term called big data was coined to get and understand the deep meaning of generation and handling of huge data. For the data acquisition, transmission, storage and further processing, anew system architecture was introduced. First, the definition and the challenges of the big data is defined. In the second stage, a detailed work to describe systems of big data into some modules. A big data chain is constructed which is followed by some approaches and procedures from both research and industry communities detailed survey. In an additional manner, for addressing big data challenges the Hadoop framework is presented. Finally, some evaluation benchmarks are directed for the big data framework. The big data and the big data value chain technologies are focused and are presented, to cover the entire big data life cycle. There are four phases in the big data value chain data generation, data acquisition, data storage, and data analysis. In the big data generation phase and further we have some efficiently rich big data sources which can be the data attributes. And here the programming models and data storage are coupled together.

Mandar Mokashi, Dr.A.S.Alvi (2013)[4] have proposed that the greatest challenging thing is to store and manage the huge amount of data that is generated and along with that there is also challenge to analyze the data. There are many processes for the collection, storage, processing and analyze big data. A collection or a group or a network of distributed sensors brought together to keep monitoring the physical or environmental conditions, like temperature, pressure, sound etc., can be said as the wireless sensor networks. By using them the sensory values or the sensed data can be sent to the sink node by using networks. The motivation for the development of the wireless sensor networks is done by the military applications such as the war surveillance. And now-a-days such technologies are used in almost all the fields. Commonly the sensors are viewed as a distributed database that collects the data that it senses from both the physical and environmental conditions. By using a group of sensor nodes, a stream of data is generated. Commonly, a small node is included in each sensor with sensing, computing, and communication abilities. We may have to think that we can have something new concept for technological improvement which can be used in the future management and analysis of the data. Based on new standards, networks are introduced and also low power systems are developed. Even in these days we can also see the widespread deployment of distributed databases in wireless sensor networks. Sensor nodes are basic in terms of components and their interfaces. In the current trend of technology, we can observe a wide scope and various implementations of distributed database management in wireless sensing devices.

PraveenKumar, Dr Vijay Singh Rathore, (2014) [5] have proposed that we are living in an age where an explosive amount of data is being generated every day. For the huge explosion of the data sensors, mobile devices, social networking websites, scientific data & enterprises contribute. This sudden bombardment can be grasped by the fact that we have created a vast volume of data in the last few years. Big data- as these large chunks of data is generally called and this technology has become one of the major field for research. Research says that tapping the potential of this data can be a good benefit for the businesses, scientific fields and public sectors. Keeping in mind the current challenges related with the analysis, scalability, time and privacy, there is a need for the development of the effective systems that can have a good potential. An evolution of the centralized architecture of data processing systems to the distributed systems has been done. The enterprises will be having problem associated with the processing of huge data, and none of the present centralized systems can efficiently process the huge data. So the distributed architecture is been used. Many solutions to the big data problem have got that includes the map reduce in as Hadoop distributed processing.

Hadoop with its efficient DFS & programming framework based on concept of mapped reduction, is a powerful tool to manage large data sets. With its map-reduce programming paradigms, overall architecture, ecosystem, fault- tolerance techniques and distributed processing, Hadoop offers a complete infrastructure to handle big data. Users must leverage the benefits of big data by adopting Hadoop infrastructure for data processing. However, the issues such as lack of flexible resource management, application deployment support, and multiple data source support pose a challenge to Hadoop adoption. Proper skill training is also needed for achieving large scale data analysis. These challenges must be overcome so that we can tap the full potential of Hadoop data management power.

PROPOSED SYSTEM

The Design is divided into following modules:

- Data Acquisition
- Data Storage
- Data Analysis

DATA AQUISITION

This phase is the data acquisition phase where the task is to collect the information for the storage and analysis of same data. In the starting stage, the data acquisition stage consists of three sub-steps, data collection, data transmission, and data pre- processing. There is no restriction between data transmission and data pre-processing regarding the order.

A. Data Collection

The retrieval of the raw data from real-world objects is the data collection process. The data **c**ollection can be done in several ways, it can be through sensors or using log files.

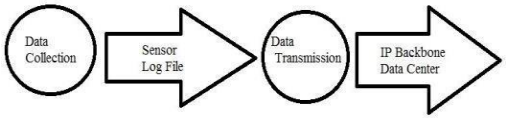

Fig 1: Data Acquisition Stages

B. Data Transmission:

Once the collection of the raw data is done, we must transfer that raw data into a data storage point, preferably in a data center and then further processed. The data transmission concept can be divided into two steps, one is the IP backbone transmission and data center transmission.[7]

- The IP backbone transmission is done at the Internet scale, where high-capacity line is being established to transfer the big data to a data center from the origin.
- When massive amount of data is transferred into the data center, it will be transited within the data center for storage

and processing. This transmission procedure is referred to as data center transmission.[9]

DATA STORAGE

- Storage infrastructure for persistent and reliable storage.
- A scalable access interface for querying and analyzing huge quantity of data.

The physically collected information is stored in the hardware infrastructure and the storage devices can be classified based on the specific technology.

Hadoop: A computing tool that is used for the distributed processing of the data can be said as Hadoop.

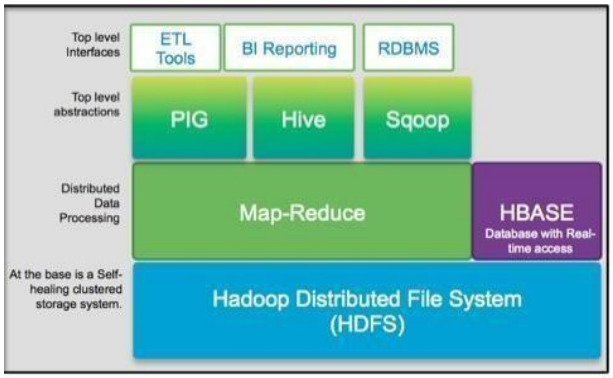

Fig 2: A Hierarchical Architecture of Hadoop Core

The pig, scoop and hive are the data integration tools that makes an easy path for the accomplishment of the data acquisition of the big data value chain and even allows for the effective import and export of data among data stores.

HDFS is a Hadoop distributed file system which is a primary substrate for the data storage in hadoop. Commonly the HDFS runs on the commodity hardware designed which is designed by using the GFS.[10]

Map reduce is the part of Hadoop which acts as the core of computation for vast amount of data analysis. Hadoop map reduce is modeled after the Google's map reduce. The map reduce framework consists of a single master job tracker and one slave task tracker per cluster node. The major responsibility of the master is to schedule the jobs for the slaves, and keep monitoring the slaves and executing the failed tasks once again.[15].

By the usage of the hadoop computing tool, the simultaneous or parallel processing of the data and efficient computation is possible. Hadoop provides distributed processing of large clusters of commodity servers. The Apache Hadoop software library is a vast computing framework consisting of many modules, including HDFS, Hadoop map reduce, hbase, hive, pig, scoop etc,.These modules can be used for examining the big data value chain structure by using which it is easy to have the powerful and efficient solutions for big data applications. [8]

In order to analyze and value extract the information, two features are provided by the data storage subsystem:

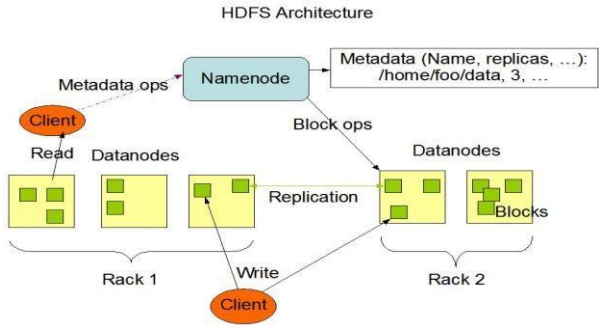

Fig 3: HDFS Architechture

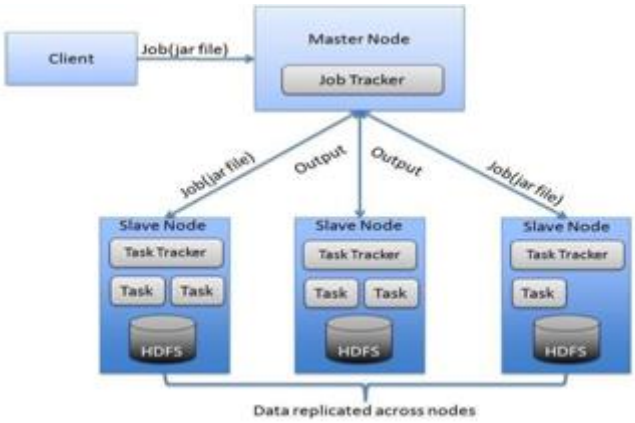

Fig 4: Hadoop with Job Tracker and Task Tracker

The management of huge set of data and its analysis tasks in map reduce programs is expressed by two SQL-like high-level declarative languages named Pig Latin and Hive.[16]

The advantages of using hadoop can be, if hadoop joins the click stream data along with the other data sources, some of the additional data often provides much more complete information about the product. Commonly the click stream data is used to understand what all qualities and constraints users consider while purchasing products. With click stream analysis, marketers can optimize and improve the product. Hadoop makes easier for scaling and also to store several years of data.

Data Analysis

The purpose, importance and the classification metric of data analytics is done in this phase. The application evolution for various data sources is reviewed and the summarization of the most relevant areas is done.[6]

The aim of this phase is to extract as much information is suitable for consideration conditions. The following points lists some potential purposes:

- Extrapolation and Interpretation of the data and its usage procedure.
- To check whether the data is legitimate.

- Assist for the decision-making.
- To find solutions for the faults.
- To predict what will occur in the future.

SYSTEM ARCHITECTURE:

The below shown diagram representation represents the system architecture

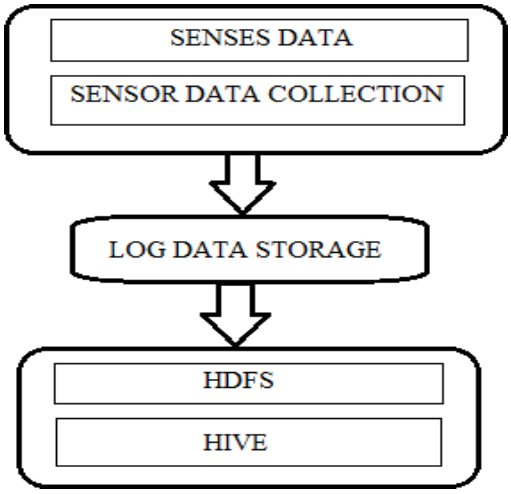

Fig 5: System Architecture

IMPLEMENTATION

Implementation steps of our project mainly involves 5 steps, they are as follows, [12]

- Generating sensor data files in a required format.
- Load the sensor data into Hadoop HDFS file system.
- To refine the sensor data run Hive scripts.
- Refined sensor data is accessed from visualization tool.
- Sensor data is visualized

RESULTS AND DISCUSSION

Generating sensor data files: The sample sensor data is collected from SensorFiles.zip and then extract the files.[13]

Load the sensor data into Hadoop HDFS file system:

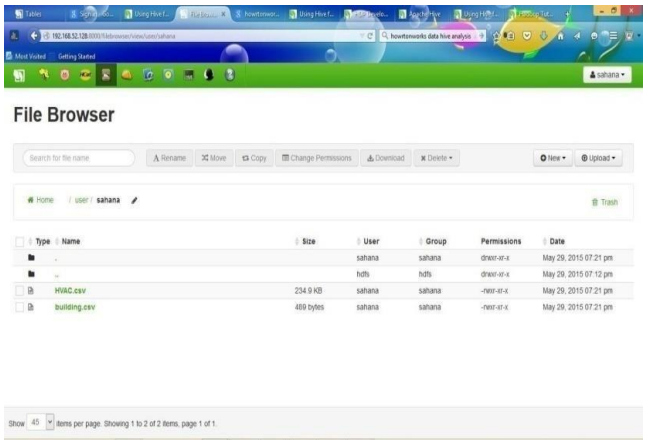

Fig 6: Creating User Account and Login, after which should upload the files

To refine the sensor data:

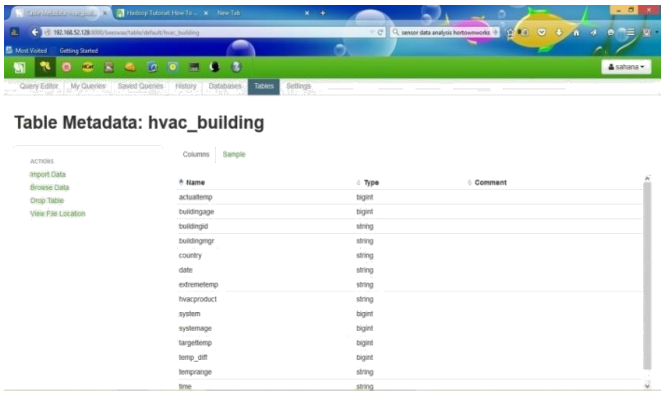

Fig 7: The sensor data that is being collected is refined and table is been created.

Access the refined sensor data:

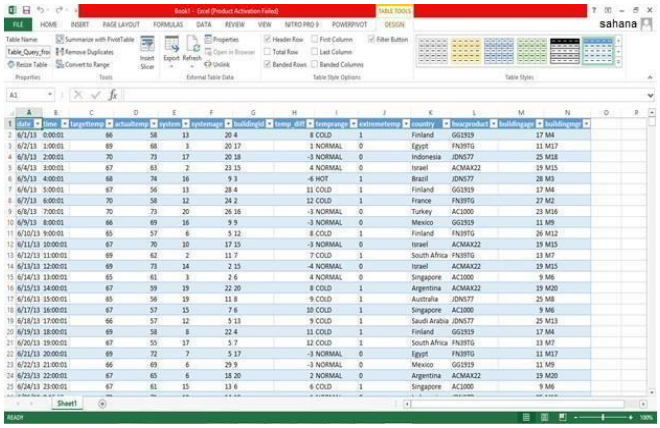

Fig 8: The data is been filtered by using filter data screen and the refined data is further processed.

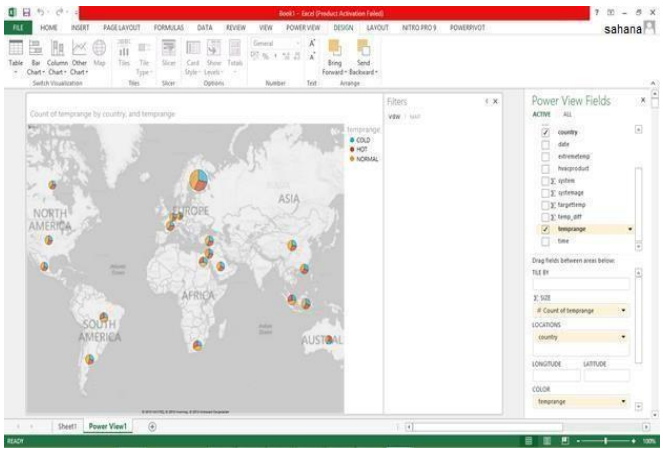

Fig 9:Data in the Table is represented in the form of Map

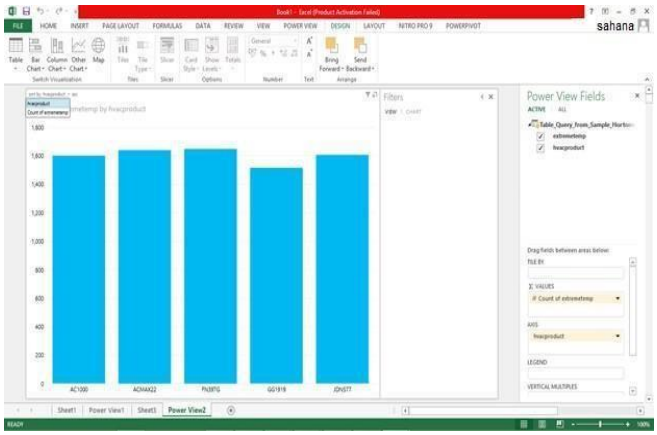

Fig 10: Difference between the models that regulate temperature

CONCLUSION AND FUTURE ENHANCEMENT:

A new survey of present research reputation upon substantial scale WSN files examination, processing along with managing, that papers offers a new WSN sensory files process using rising Hadoop tactic[11]. Powerful files processing center acquires a strategy for monitoring the lack of sensors files storage along with the files processing capacity. This project also covers the actual challenges from the big data files. The actual big data worth chain involves some stages of development: files generation, files order, files storage, along with files examination. The actual story a mix of both architecture permits the actual series, stocking along with processing of a lot of sensory files. The look along with put into action from the sensory files processing process can also be in depth. By using the easiest evaluation technique for the performance analysis, procedure of data query and data approximation, a Hadoop based sensory data processing and data analysis is a very effective solution for huge sensor data processing. The rapid growth in the data volume will be a problem for the real time requirements and other batch processing to adopt that. Some solutions are need to be designed for real-time processing model or for the data analysis mechanism. Commonly more time is wasted for the data transmission, storage, and processing.[14]

REFERENCES

[1] Gurpreet Singh Bedi, Ashima Singh, *Big Data Analysis with Dataset Scaling in Yet another Resource Negotiator (YARN)*, International Journal of Computer Applications (0975–8887) Volume 92 –No.5, April2014

[2] *B*ing Tang and Yu Wang ,*Design of Large-Scale Sensory Data Processing System Based on Cloud Computing,* , School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, Hunan, 411201, China, College of Computer and Information Engineering, Honai University, Research Journal of Applied Sciences, Engineering and Technology ISSN: 2040- 7467, © Maxwell Scientific Organization,2012

[3] Han hui ,yong gang wen, tat- seng chual , and xuelongli, *Toward Scalable Systems for Big Data Analytics: A Technology Tutoria*l, October 2013.

[4] Mandar Mokashi, Dr.A.S.Alvi ,*Data Management in Wireless Sensor Network: A Survey*- IJARCCE, Vol. 2, Issue 3, March 2013.

[5] PraveenKumar, Dr Vijay Singh Rathore *Efficient Capabilities of Processing of Big Data using Hadoop Map Reduce*-, IJARCCE, Vol. 3, Issue 6, June 2014

[6] X. Wu, X. Zhu, G. Wu, and W. Ding, Transactions onKnowledge and Data Engineering, *Data mining with big data*, vol. 99, June2013.

[7] Vodel and W. Hardt, *Data aggregation in resource-limited wireless communication environments - differences between theory and praxis*, M. in Proceedings of the International Conference on Control, Automation and Information Sciences (ICCAIS2012). Ho Chi Minh City, Vietnam: IEEE Computer Society, November 2012, pp.282–287.

[8] D. Laney, Gartner, *The Importance of 'Big Data': A Definition.* 2012.

[9] S.Madden, R. Szewczyk, M. J. Franklin, and D. Culler ,*Supporting Aggregate Queries Over Ad-Hoc Wireless Sensor Networks*, ,in Proceedings Fourth IEEE Workshop on Mobile Computin Systems and Applications, 2002.

[10] M. J. Millerand N. H. Vaidya,*Minimizing energy consumption in sensor networks using a wakeup radio*, in Proceedings of the Wireless Communications and Networking Conference. WCNC. 2004 IEEE, vol. a, March 2004, pp.2335– 2340.

[11] Chair of Computer Engineering, Technische Universit¨at Chemnitz, "Planet – platform for ambient networking," www.ce.informatik.tuchemnitz.de/forschung/demonst ratoren/planet/, September2013.

[12] M. Vodel, R. Bergelt, and W. Hardt, *A generic data processing framework for heterogeneous sensor-actor-networks*, , International Journal On Advances in Intelligent Systems, vol. Vol. 4, December 2012.

[13] hortonworks.com/products/hortonworks-sandbox-URL

[14] M. Ceriotti ,*Monitoring heritage buildings with wireless sensor networks: The Torre Aquila deployment*,., in Proc. Int. Conf. Inform. Process. Sensor Netw., Apr. 2009,pp. 277–288.

[15] F. Gallagher,*The Big Data Value Chain [Online]*,.Available:http://fraysen.blogspot.sg/2012/06/big-data-value-chain.html(2013)

[16] M. Sevilla.*Big Data Vendors and Technologies, the list! [Online]*, Available: http://www.capgemini.com/blog/capping-off/2012/09/big-data-vendors-a%nd-technologies. (2012).