IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Differential Privacy Based Solution for Protecting Privacy of Big Data

Y. Sowmya[1*], M. NagaRatna[2], C. Shoba Bindhu[3]

[1]Department of Computer Science and Engineering, JNTUA, Anantapuramu, India
[2]Department of Computer Science and Engineering, JNTUH College of Engineering, Hyderabad, India
[3]Department of Computer Science and Engineering, JNTUA College of Engineering, Anantapuramu, India

[*]*Corresponding Author: yalla.soumya.reddy1980@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 11/Jun/2018, Published: 30/Jun/2018

***Abstract*--**With the emergence of distributed programming frameworks like Hadoop and cloud computing technology, big data and its analytics became a reality. As big data needs huge amount of storage and computing resources, cloud has given solution to the needs of big data. However, it is important to protect big data from privacy attacks. Disclosure of identity of an entity or organization or a person in the big data is an example for loss of privacy. In other words, non-disclosure of privacy of certain sensitive attributes is nothing but preserving privacy of big data. As traditional computing is replaced by Internet based computing, it became essential to deal with privacy of big data. Many techniques came into existence to protect big data. In this paper, we considered a specific case where an adversary launches attack to know the presence or absence of an entity in the big data. We proposed an algorithm based on differential privacy to withstand the aforementioned privacy attack on big data workload in MapReduce programming paradigm. We built a prototype application and deployed it in Elastic MapReduce (EMR) of Amazon Elastic Compute Cloud (EC2). The experimental results revealed the utility of the proposed algorithm and showed proof of the concept.

***Index Terms* –** Big data, big data privacy, differential privacy, Elastic MapReduce (EMR)

I. INTRODUCTION

Big data has been around for some years now. It is the data that reflects voluminous, continuously streaming data with different kinds of data present. The big data needs support from cloud computing for storage and computing needs and programming frameworks like Hadoop and Amazon Elastic MapReduce (EMR) to mention few. Big data processing is made when it is outsourced to public cloud. The data owners are concerned with the privacy of big data when it is outsourced to third party of public cloud for mining purposes of publication purposes. In such cases, the privacy of big data becomes crucial. Many methods are found in the literature [4]-[9] to deal with big data, big data processing and security. Privacy of big data is still more important and that is the main focus of this paper.

We proposed an algorithm for protecting privacy of big data. Our algorithm is based on the differentia privacy technique. The algorithm is named as Big Data Privacy Protection (BDPP) Algorithm. It makes use of the output produced by reducers and then adds noise to the output. Thus it can prevent privacy attacks. Especially we made experiments to prevent an attack in which an adversary tries to find the presence or absence of an entity in the MapReduce workload. The algorithm prevents this kind of attack with the noise addition as per the proposed algorithm. The following are the contributions of this paper.

- We proposed an algorithm named Big Data Privacy Protection (BDPP) Algorithm for protecting privacy of big data while it is being processed using MapReduce programming paradigm.
- We built a prototype application to demonstrate proof of the concept. The application has both map and reduces coding that facilitates the presence or absence of an attacker to know the efficiency of the proposed algorithm.
- We made experiments with Amazon EC2's Elastic MapReduce (EMR) in order to evaluate the utility of the proposed algorithm. The results showed the importance of the algorithm in protecting privacy of big data.

The remainder of the paper is structured as follows. Section 2 presents review of literature on big data, security issues and the methods found. Section 3 presents cloud and big data eco system. Section 4 presents proposed algorithm. Section 5 provides experimental results while section 6 concludes the paper and gives directions for future work.

II. RELATED WORK

This section provides review of literature on the big data privacy in the cloud-big data eco system and distributed programming frameworks like Hadoop. Green cloud computing is explored in [1] to have computing with energy efficiency and environmental friendly. With this technique they cold measure Water Usage Effectiveness (WUE) and Data Centre Productivity (DCP). Knowledge transfer and the means of e-Learning to transfer knowledge are explored in [2]. Resource allocation based on trust for effective resource utilization is studied in [3]. Leveraging a real world business with cloud computing case is explored in [4]. Workflow scheduling for performance improvement in cloud computing is made in [5].

Load balancing issues and security aspects of cloud is studied in [6]. Cloud computing infrastructure is analysed and a method is proposed in [7] for effective knowledge management. In [9] analysis of top 10 security threats is analysed in cloud computing. The threats they identified include abuse, Denial of Service (DoS), technology vulnerabilities, insufficient due diligence, malicious insiders, data loss, insecure interfaces and data breaches. Cloud computing with skyline queries is the main focus in [10] for better performance and meeting user needs. The trends in the big data processing are investigated in [11] while the issues related to big data storage and processing is found in [12].

The focus of [13] is to understand an emerging framework known as Hama for processing big data and running big data applications. The programming mode it follows is known as bulk synchronous parallel programming model. Performance of MapReduce is studied with different platforms like OpenNebula, KVM and OpenVZ in [14] to have comparative analysis. They found KVM to provide better performance with given I/O benchmarks. Big data security with respect to Hadoop is studied in [15]. They found different solutions for security issues like encryption, authentication, and authorization to be provided besides security to big data which is in transit and at rest. They evaluated these security issues in MapReduce, Pig, HBase, Pig, OOzie, Hue and Zookeeper. MapReduce programming and its performance with various data centres is found in [16]. They employed two big data case studies known as High Energy Physics (HEP) and Large Hordon Collidor (LHC). In fact, they improved Hadoop and named the product as G-Hadoop to achieve this. Optimisation Hadoop clusters and the improvement of big data analytics with Elastic MapReduce (EMR) with Amazon 3 storage is investigated in [17].

Big data processing is studied in [18] for understanding MapReduce paradigm. They found different challenges and issues related to big data processing. They studied big data projects related to Government, private sector, and Big Science. They identified certain challenges like storage, significance, best practices, volume and technical challenges related to big data analytics. Different virtualization systems that are used with the framework MapReduce are investigated to understand the utility of computing resources. Container based virtualization is understood to prevail in the industry [19].

As explored in [20] big data can help in analysing and producing business intelligence. Resource navigator known as YARN is studied in [21] that are associated with Apache Hadoop. YARN divides the functionality of resource management from the programming part and improves performance. Big data analytics for mobile platforms is made in [22]. They explored it to know the utility of RESTful web services in the context. They built a prototype application to investigate it. The functioning of G-Hadoop with SSL is explored in [23] with respect to managing big data and security across data centres. It also focused on the authentication procedures and job scheduling.

K-Means, a data mining algorithm, is executed in [24] to understand the functionality of Hadoop MapReduce. The relationship between big data and cloud computing is explored in [25] while the challenges in big data processing are investigated in [26]. The trends and security challenges in big data process is studied in [27]. The trends are related to hardware platforms, virtualization techniques, data analytics, and emerging applications. In [28] MapReduce programming is understood in terms of processing spatial big data. Motion imagery is used as big data to know the performance of Hadoop and HDFS. They employed a technique known as Cloud-Enabled WAMI Exploitation (CAWE). Open source frameworks being used for big data analytics are studied in [29]. They investigated different frameworks such as Apache Spark, Pregel, YARN, GPS, and Apache Hadoop. They also examined the performance of Scala-based frameworks like Scalation, Samza, Kafka and Spark. The utility of big data and its usage is studied in [30] when layered architecture is used. Caching mechanism with data-awareness is investigated in [31].

Scheduling algorithms with Hadoop MapReduce are studied in [32]. They used different schedulers to compare performance. They include fair scheduling, FIFO, self-adaptive MapReduce, Maestro, delay scheduling, and capacity scheduler. Apache Spark, an engine for processing big data, is examined in [33] and found that the framework has potential capabilities to process big data. In [34], MapReduce framework is studied to ascertain Internet traffic and the scalability of the framework. They found that caching mechanism improved query performance. In [35] security of the frameworks like YARN and Hadoop in processing big data is explored with respect to threats like DoS and provided

useful insights. Secure data processing with MapReduce [40] and deduplication mechanism for security [41] are other contributions found in the literature. From the review of literature it is found that big data security and privacy is an issue to be resolved. In this paper we proposed an algorithm to secure privacy of big data in the confines of MapReduce programming paradigm.

III. BIG DATA – CLOUD ECO SYSTEM

Big data is the data with characteristics like volume, velocity and variety. Huge amount of data that cannot be processed and stored in the local machines is big data. This way big data can be understood with ease. It is measured in peta bytes and needs computing resources of cloud. As shown in Figure 1, it is evident that the big data processing provides comprehensive business intelligence. If big data is not considered, it leads to biased conclusions that harm businesses when decisions are made on the biased outcomes.

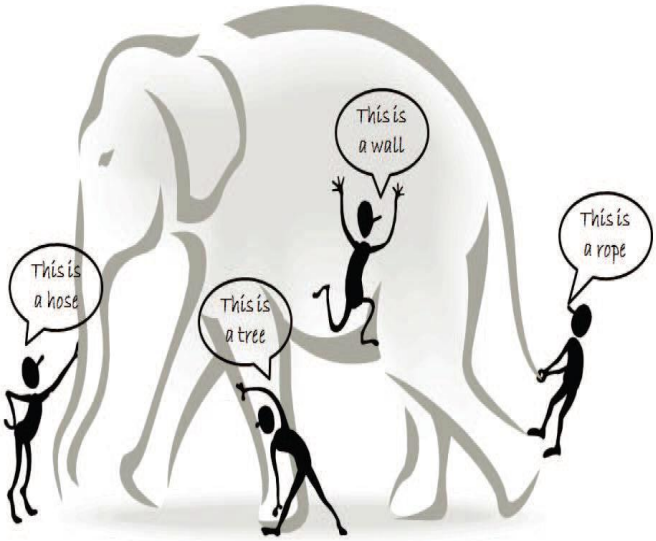

Figure 1: Shows how bias conclusions are made [36]

There are many distributed programming frameworks or platforms that can be used to process big data. They are used for real time processing and batch processing. The frameworks used are Storm, S4, Splunk, Apache Kafka, SAP Hana, and SQL-Stream server. Other frameworks include Tablueau, Skytree server, Dryad, Apache Mahout, Apache Hadoop, Jaspersoft BI Suite, Talend Open Studio, Karmasphere Studio and Analyst, and Pentaho Business Analytics.

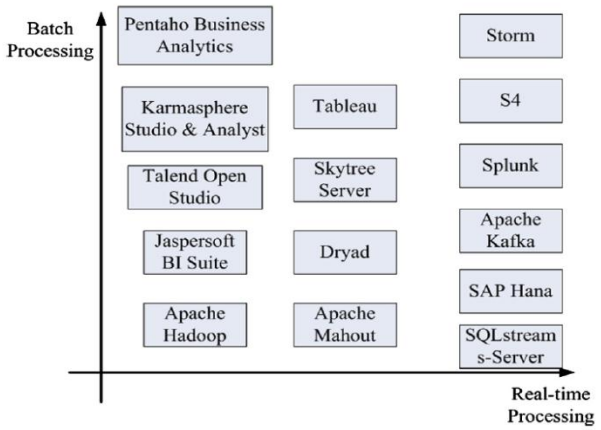

Figure 2: Big data platforms for batch and real-time processing [37]

As presented in Figure 2, the real time and batch processing frameworks are provided. They are used to process big data. Then big data requires MapReduce programming framework which is used to process large volumes of data. In Hadoop kind of distributed programming framework, MapReduce architecture is as shown in Figure 3. Client program submits a job to Job Tracker. Then job tracker allocates it to task tracker. The inputs and outputs are maintained in a distributed file system known as Hadoop Distributed File System (HDFS). The input is split into number of parts and each part is given to map task. Once map task is completed, the intermediate result is given to reduce tasks to produce final output.

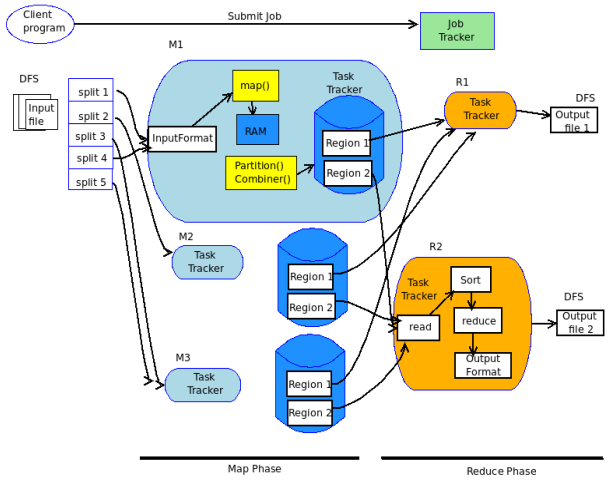

Figure 3: MapReduce programming paradigm in Hadoop

As presented in Figure 3, both map and reduce tasks are illustrated. There are many bench marks available to run in MapReduce model. They include WordCount and Terasort to mention few. The MapReduce functionality with WordCount benchmark is illustrated in Figure 4.

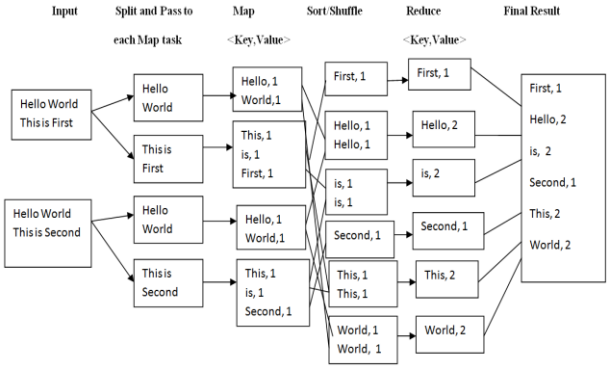

Figure 4: Illustrates MapReduce functionality with word cloud application

As illustrated in Figure 4, plenty of text documents (big data) are given as input to the system. Then the text is divided into number of parts. Each part is given to a mapper (worker nodes in the cluster). Then the mapper produces intermediate output. The reduce needs to produce final output by taking intermediate output produced by mappers after sorting. The final output is then stored in HDFS.

IV. PROPOSED ALGORITHM

This section provides details of the proposed algorithm. The algorithm takes big data as in put in the form of key/value pairs and returns output that has been subjected to differential privacy. In other words enough noise is added to the data so as to prevent such privacy attacks that focus on finding the presence or absence of chosen entity in the MapReduce workload. The dataset used for experiments is known as EDGAR dataset [39]. It is of big data where IP address is the sensitive attributes. The adversary targets to know the presence or absence of an IP address in the MapReduce workload. The algorithm is based on the differential privacy explored in [38]. Noise is added to the reducer output in order to protect the big data from privacy attacks.

Algorithm 1: Big Data Privacy Protection (BDPP) Algorithm

```
1    Privacy_BigData(key k, values vector V)
2    Initialise output vector V'
3    For every value v  in V do
4    For k in 1 to n
5    If v is unique and found only once Then
6    v is considered unique
7    Add v to V'
8    end if
9    end for
10   finalOutput = ReduceFunction(V') x (1+R(ε))
11   return finalOutput
```

Algorithm 1: Big data privacy protection algorithm based on differential privacy

As presented in Algorithm 1, it is evident that the privacy of big data is protected by using noise addition. The reducer output is subjected to noise addition so as to ensure that the attacker will not be able to find the presence of absence of an IP address in the MapReduce workload. As the IP addresses are sensitive the attacker might wish to know the presence of an IP address. In order to avoid malicious attacks, the key and value lists are separately used in the process. When the adversary tries to fix a value in Map and find in Reduce phase, the output subjected to differential privacy defeats the purpose of the attacker as he fails in finding the presence of absence of particular entity.

V. EXPERIMENTAL RESULTS

Experiments are made with Amazon Elastic Compute Cloud (EC2) in Elastic MapReduce (EMR) where the MapReduce application is executed in presence of an attacker an in the absence of attacker. The absence of attacker showed genuine count (number of occurrences) of an IP address. In presence of attacker the proposed algorithm is employed to employ differential privacy on big data. EDGAR dataset is used for experiments. Here the attacker wants to know the presence or absence of an IP address by setting target value to count of occurrences. The purpose of the attacker in fixing a value and finding the presence of absence of an IP address in the dataset is defeated using the proposed methodology.

Table 1: Result of MapReduce in absence of attacker

IP	**Genuine Count**
104.35.98.11	33123
104.35.98.12	7456
104.35.98.13	9265
104.35.98.14	12432
104.35.98.15	14567
104.35.98.16	7652
104.35.98.17	8142
104.35.98.18	5839
104.35.98.19	15464

As shown in Table 1, it is evident that the IP addresses and the count of occurrences of each IP address are presented

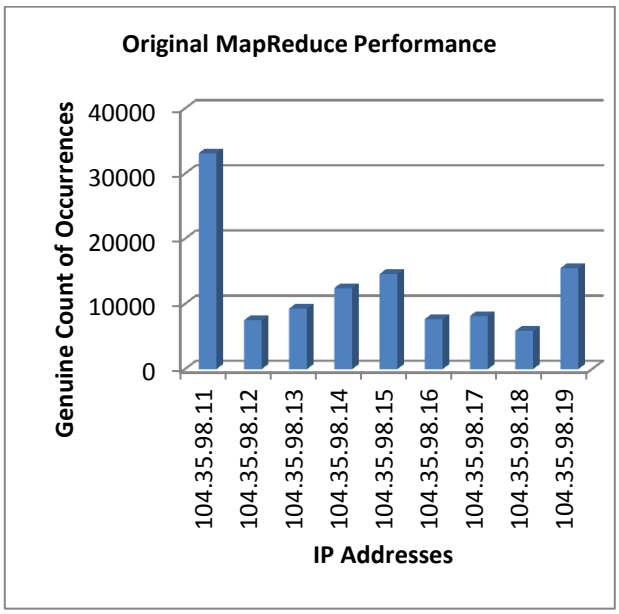

Figure 5: Original MapReduce performance

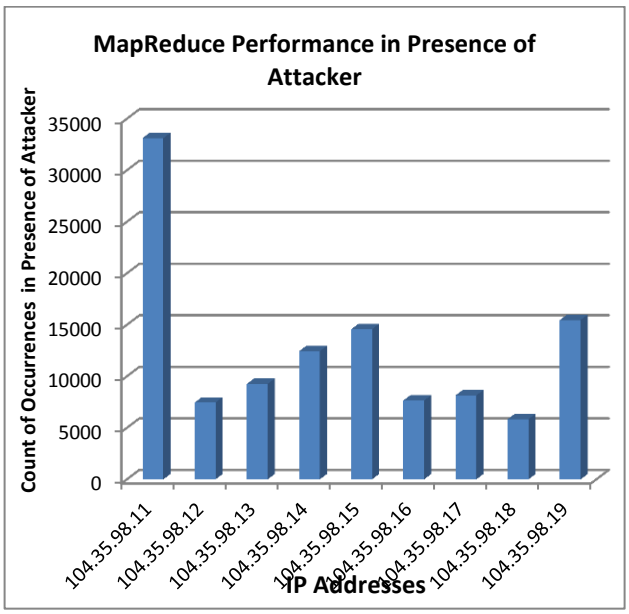

Figure 6: MapReduce performance in presence of attacker

As shown in Figure 5, the IP addresses found in the EDGAR dataset are presented in horizontal axis. The count of occurrences of each IP address is showed in vertical axis. The results showed the genuine count of IP addresses in absence of an attacker. The results will slightly changes in presence of attacker in order to protect privacy of big data. The results in presence of attacker are shown in Table 2.

Table 2: Result of MapReduce in presence of attacker

IP	Count in Presence of Attacker
104.35.98.11	33122
104.35.98.12	7455
104.35.98.13	9261
104.35.98.14	12431
104.35.98.15	14566
104.35.98.16	7651
104.35.98.17	8141
104.35.98.18	5838
104.35.98.19	15463

As presented in Table 2, the IP addresses of EDGAR dataset and the corresponding count of occurrences of each IP address in presence of an attacker are shown.

In presence of an attacker, the proposed methodology employs differential privacy and the target value of the attacker to know the presence of absence of an IP address in the EDGAR database (big data) is changed. Thus the attacker will not be able to find the presence of absence of the target IP address. The attack fails in its intended purpose and thus the differential privacy is able to provide privacy to big data.

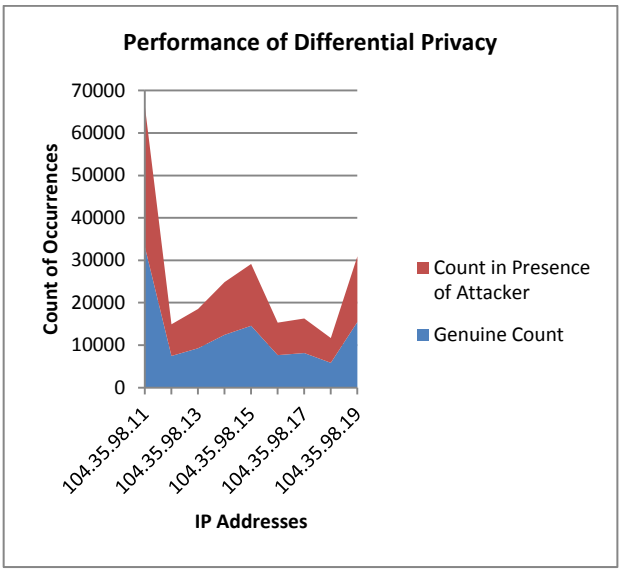

Figure 7: MapReduce performance comparison in presence and absence of attacker

As presented in Figure 7, it is evident that the IP addresses found in EDGAR dataset (big data) are provided in horizontal axis while the vertical axis shows the genuine count of IP addresses and count of occurrences of IP addresses in the presence of an attacker. In the presence of an attacker, the differential privacy is employed and thus the values are changed for each IP address. Thus the target IP address whose presence of absence is to be found by the attacker is not disclosed. Since IP address is the most sensitive attribute in the EDGAR dataset, its presence of absence in the MapReduce job workload reveals its privacy. Thus the attack made by adversary is not successful.

VI. CONCLUSIONS AND FUTURE WORK

In this paper the privacy issues of big data are explored in the context of MapReduce programming frameworks like Amazon Elastic MapReduce (EMR). Privacy of big data is very important as the data is outsourced by data owners to public cloud. Non-disclosure of sensitive data is an important aspect in preserving privacy of big data. In this paper we proposed an algorithm based on differential privacy to add noise to big data in order to protect sensitive attributes from privacy attacks. EDGAR dataset is used to evaluate the proposed algorithm. Map and Reduce tasks are used to evaluate it with and without an attack scenario. When there are no attacks, the normal execution takes place and the MapReduce application provides the count of occurrences of IP addresses in the dataset. In presence of an attacker, the proposed algorithm based on differential privacy is employed and the results showed that the purpose of the attacker to know the presence of absence of an IP address in the MapReduce workload is defeated. A prototype application is built and executed in Amazon EC2 Elastic MapReduce (EMR) environment. The results showed that the proposed algorithm is efficient in withstanding privacy attacks that target to know the presence or absence of certain entity in the MapReduce workload. In future, we plan to work on MapReduce security issues in presence of rogue worker nodes in Haloop clusters.

REFERENCES

[1] Imran Ghani, Naghmeh Niknejad, Seung Ryul Jeong, "Energy Saving In Green Cloud Computing Data Centers: A Review", *Journal Of Theoretical And Applied Information Technology*. 74 (1), P1-16, 2015.

[2] Danny Manongga, Wiranto Herry Utomo, Hendry, "E-Learning Development As Public Infrastructure Of Cloud Computing", *Journal Of Theoretical And Applied Information Technology*. 62 (1), P1-6, 2014.

[3] V.Suresh Kumar, Dr. Aramudhan, "Hybrid Optimized List Scheduling And Trust Based Resource Selection In Cloud Computing", *Journal Of Theoretical And Applied Information Technology*. 69 (3), P1-9, 2014.

[4] Bachtiar H. Simamora, M.Sc., Ph.D., Julirzal Sarmedy, S.Kom, "Improving Services Through Adoption Of Cloud Computing At Pt Xyz In Indonesia", *Journal Of Theoretical And Applied Information Technology*. 73 (3), P1-10, 2015.

[5] P. Kumar And Sheila Anand, "An Approach To Optimize Workflow Scheduling For Cloud Computing Environment", *Journal Of Theoretical And Applied Information Technology*. 57 (3), P1-7, 2013.

[6] Ayman G. Fayoumi, "Performance Evaluation Of A Cloud Based Load Balancer Severing Pareto Traffic", *Journal Of Theoretical And Applied Information Technology*. 32 (1), P1-7, 2011.

[7] Ratna Sari, Yohannes Kurniawan, "Cloud Computing Technology Infrastructure To Support The Knowledge Management Process (A Case Study Approach)", *Journal Of Theoretical And Applied Information Technology*. 73 (3), P1-6, 2015.

[8] S.Sudha, V.Madhu Viswanatham, "Addressing Security And Privacy Issues In Cloud Computing", *Journal Of Theoretical And Applied Information Technology*. 48 (2), P1-13,2013.

[9] M. Lemoudden, N. Ben Bouazza, B. El Ouahidi, D. Bourget, "A Survey Of Cloud Computing Security Overview Of Attack Vectors And Defense Mechanisms", *Journal Of Theoretical And Applied Information Technology*. 54 (2), P1-6, 2013.

[10] Abdellah Idrissi And Manar Abourezq, "Skyline In Cloud Computing", *Journal Of Theoretical And Applied Information Technology*. 60 (3), P1-12, 2015.

[11] Marcos D. Assuncaoa, Rodrigo N. Calheirosb, Silvia Bianchic, Marco A. S. Nettoc And Rajkumar Buyyab, "Big Data Computing And Clouds: Trends And Future Directions", *Acm*. P1-44, 2014.

[12] Arpit Gupta,Rajiv Pandey, And Komal Verma, "Analysing Distributed Big Data Through Hadoop Map Reduce", *Ieee*. 129, P1-7, 2015.

[13] Kamran Siddique, Zahid Akhtar, Edward J. Yoon, Young-Sik Jeong, I, Dipankar Dasgupta, And Yangwoo Kim, "Apache Hama: An Emerging Bulk Synchronous Parallel Computing Framework For Big Data Applications", *Ieee*. 4 ,P1-9, 2016.

[14] Pedro Roger Magalhaes Vasconcelos And Gisele Azevedo De Araujo Freitas, "Performance Analysis Of Hadoop Mapreduce On An Opennebula Cloud With Kvm And Openvz Virtualizations", *Icitst*. P1-7, 2014.

[15] Priya P. Sharma And Chandrakant P. Navdeti, "Securing Big Data Hadoop: A Review Of Security Issues, Threats And Solution", *Ijcsit*. 5, P1-6, 2014.

[16] Lizhe Wanga, Jie Taoc, Rajiv Ranjan D, Holger Martenc, Achim Streit C, Jingying Chene And Dan Chena, "G-Hadoop: Mapreduce Across Distributed Data Centres For Data-Intensive Computing", *Ieee*, P1-14, 2013.

[17] Yanish Pradhananga,Shridevi Karande And Chandraprakash Karande, "High Performance Analytics Of Bigdata With Dynamic And Optimized Hadoop Cluster", *Isbn*, P1-7, 2016.

[18] Avita Katal, Mohammad Wazid And R H Goudar, "Big Data: Issues, Challenges, Tools And Good Practice", *Ieee*, P1-6, 2104.

[19] Miguel G. Xavier, Marcelo V. Neves And Cesar A. F. De Rose, "A Performance Comparison Of Container-Based Virtualization Systems For Mapreduce Clusters", *Acm*, P1-9, 2014.

[20] Alberto Fernandez, Sara Del Rio, Victoria Lopez, Abdullah Bawakid, Maria J. Del Jesus, Jose M. Benítez And Francisco Herrera, "Big Data With Cloud Computing: An Insight On The Computing Environment, Mapreduce, And Programming Frameworks", *Acm*, P1-31,2014.

[21] Vinod Kumar Vavilapallih, Arun C Murthyh, Chris Douglasm, Sharad Agarwali ,Mahadev Konarh, Robert Evansy, Thomas Gravesy, Jason Lowey, Hitesh Shahh, Siddharth Sethh ,Bikas Sahah, Carlo Curinom And Owen O'malleyh San, "Apache Hadoop Yarn: Yet Another Resource Negotiator", *Acm*, P1-P16, 2013.

[22] Ngu Wah Win And Thandar Thein, "An Efficient Big Data Analytics Platform For Mobile Devices", *Ijcsis*. P1-5, 2015.

[23] Jiaqi Zhaoa, Lizhe Wangb, Jie Taoc, Jinjun Chend, Weiye Sunc, Rajiv Ranjane, Joanna Kołodziejf, Achim Streitc And Dimitrios Georgakopoulose, "A Security Framework In G-Hadoop For Big Data Computing Across Distributed Cloud Data Centres" *Journal Of Computer And System Sciences*, P1-14, 2014.
[24] Amresh Kumar,Kiran M.,Saikat Mukherjee And Ravi Prakash G, "Verification And Validation Of Mapreduce Program Model For Parallel K-Means Algorithm On Hadoop Cluster", *International Journal Of Computer Applications*. 72, P1-P8, 2013.
[25] Mythreyee S,Poornima Purohit And Apoorva D.R, "A Study On Use Of Big Data In Cloud Computing Environment", *Ijariit*. P1-7, 2017.
[26] Katarina Grolinger, Michael Hayes, Wilson A. Higashino, Alexandra L'heureux, David S. Allison And Miriam A.M. Capretz, "Challenges For Mapreduce In Big Data", *IEEE*, P1-P10, 2014.
[27] Karthik Kambatlaa, Giorgos Kollias B, Vipin Kumarc And Ananth Gramaa, "Trends In Big Data Analytics", *IEEE*, P1-13, 2014.
[28] Erkang Chenga, Liya Maa, Adam Blaissea, Erik Blaschb, Carolyn Sheaffb, Genshe Chenc, Jie Wua And Haibin Linga, "Efficient Feature Extraction Fromwide Area Motion Imagery By Mapreduce In Hadoop", *Acm*. P1-9, 2015.
[29] John A. Miller, Casey Bowman, Vishnu Gowda Harish And Shannon Quinn, "Open Source Big Data Analytics Frameworks Written In Scala", *IEEE*. 1-5, 2016.
[30] Harshawardhan S. Bhosale, Prof. Devendra And P. Gadekar, "A Review Paper On Big Data And Hadoop", *Ijsrp*, P1-7, 2014.
[31] Yaxiong Zhao, Jie Wu, And Cong Liu, "Dache: A Data Aware Caching For Big-Data Applications Using The Mapreduce Framework", *Tsinghua Science And Technology*. P1-12, 2014.
[32] Seyed Reza Pakize, "A Comprehensive View Of Hadoop Mapreduce Scheduling Algorithms" *Ijcncs*. P1-10, 2014.
[33] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez And Scot, "Apache Spark: A Unified Engine For Big Data Processing", *Acm*. 59, P1-10, 2016.
[34] Yeonhee Lee And Youngseok Lee, "Toward Scalable Internet Traffic Measurement And Analysis With Hadoop", *Acm*. P1-8, 2013.
[35] Jingwei Huang, David M. Nicol, And Roy H. Campbell, "Denial-Of-Service Threat To Hadoop/Yarn Clusters With Multi-Tenancy", *Ieee*. P1-8, 2014.
[36] Xindong Wu,Xingquan Zhu,Gong-Qing Wu, "Data Mining With Big Data", Ieee. 26 (1), P.97-107, 2014).
[37] C.L. Philip Chen , Chun-Yang Zhang, "Data-Intensive Applications, Challenges, Techniques And Technologies: A Survey On Big Data", Elsevier. P.32-44, 2014.
[38] R. Agrawal And R. Srikant , "Privacy-Preserving Data Mining", In Proceedings Of The Acm Sigmod Conference On Management Of Data. Dallas, Pp.439-450. 2000.
[39] Securities And Exchange Commission, Edgar Log File Data Set. Available: Https://Www.Sec.Gov/Data/Edgar-Log-File-Data-Set. Last Accessed 10 November 2016.
[40] H. Kousar and B.R.P. Babu, "Efficient Map/Reduce secure data using Multiagent System," International Journal of Computer Sciences and Engineering. 6 (5), p1-5, 2018.
[41] M. Murugesan and A. Kalaiyarasi, "An Efficient Deduplication Mechanism for Big Data Analysis in Cloud Environments," International Journal of Computer Sciences and Engineering. 6 (4), p1-7,2018.

Authors Profile

Y. Sowmya pursued M.Sc,M.Tech in Computer Science and Engineering ,is pursuing Ph.D from JNTUA Anantapuramu from .She is a research scholar of JNTUA in Computer Science and Engineering.She has published 3 international journals and 1 international conference .Her mainresearch focuses on Big Data Mining.She has 12 years of teaching experience.

Dr.M. Nagaratna pursued B.E,M.Tech,Ph.D in computer science & Engineering from JNTUH.She is currently working as an Associate Professor in Computer Science &Engineering since 2003 in JNTUH .Chaired a Technical Session in the 1st International Conference on Computational Intelligence and Informatics (ICCII-2016) held during May 28-30, 2016 at JNTUH.She has published more than 20 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it's also available online.Her main research work focuses on Computer Networks, Design & Analysis of Algorithms, Advanced Data Structures, Mobile Computing.She has 15 years of experience.

Dr.C Shoba Bindhu pursued B.Tech,M.Tech,Ph.D in computer Science engineering from JNTUA Anantapuramu .She is currently working as a Professor in Computer Science Engineering in JNTUA.She is a coordinator of MSIT JNTU College of Engineering Anantapur from February 2016,NMEICT Nodal officer JNTUA College of Engineering, Anantapur .Received gold medal in B.Tech and University Topper in M.Tech..She has published more than 50 international journals ,4 national,31 international conferences,19 national conferences including IEEE..Her main research work focuses on Networking Security, Wireless Networks, Cloud Computing.She has 15 years of experience.

