SafeHadoop: Reinventing Data Security for Distributed Environments

Sunitha T[1], Sarala DV[2], Monisha G B[2], Supreeth S[2],

[1]Assistant Professor, Dept. of Computer Applications,
B.M.S. College of Engineering

[2]Assistant Professor, Dept. of Computer Science and Engineering,
B.M.S. College of Engineering

Abstract: ***Big Data is a widespread technology: a large amount data could be quickly generated from various sources such as phones, sensors, and social media. Conventional computer systems are not competent to store and process large datasets hence big data frameworks are required. Mainly there is a lack of control over the data, and data privacy for data owners, while cloud service providers may offer security entities such of protection of data in rest and at motion, end usage people have no control whatsoever over those mechanisms. To the best of our knowledge, none of these current solutions provide big data users a guarantee regarding the integrity and processing of the big data. For example in cloud environment end usage entity store the data into the cloud platform and believe on the cloud server to share their data to other users. To share end users data to authenticated and authorized end users, it is necessary to apply access control mechanisms based on the requirements of end users. When data owners upload data onto cloud or big data nodes they lose control of their data physically, hence there is a need for cyber security tool at client workstation to provide data security with encryption***
and provide users control over their data hence enhancing information security features of the platform.

Keywords: CP-ABE, AES, HDFS, VMware

1. Introduction

Cyber Security is a modern paradigm to provide protection services for systems, networks, and computer programs from attacks from digital world through the Internet. Dealing with all aspects of security is key part of cyber security in computing era, which involves the concept of cyber attacks that are usually developed and have the objective to accessing, changing, or destroying sensitive information, such that the users are lead to face money extortions. To effectively implement cyber security and its measures has become particularly challenging issue these days, one of these issues is there are more devices present in the world compared to number of people. With more devices arises more data generated every second hence big data platforms are used, hence by combining cyber security with big data will provide us with a secure platform for information security domain.

Information Security is considered to be a sub domain under the bigger branch cyber security, this project mainly comes under information security which is commonly abbreviated as infosec. The work that goes into information security deployment in any organization or institution can be thought of as preventing unauthorized use or access of any data, source code, or application. This can be further stated as preventing or blockage of processes done by cyber criminals and their espionage activities like disclosure of data and intellectual property rights, disruption of work flow of any organizations work environment, modification of sensitive data, data biasing, inspection and recording of information based on their suit of needs. Information of users upon deployed to the cloud or big data servers for processing and access can be categorized under data at flow and data in motion, and information security focuses on both data at motion and at rest.

Moving along the information security domain, security managers have been following the CIA triad which is very popular amongst information security community of people, which is confidentiality, integrity, and availability. Protecting confidentiality of information remains one of the top priority under information security procedure, by enabling confidential protocols the information will be accessed only from the desired community of users and not by others, confidentiality is more important because if it fails the intellectual property, the work code and ethics will be in ruins, and the competitive organizations can gain strategic advantage over others. Protecting data integrity is also one of the top priority under information security procedure, to make sure whether the data sent is same as the data received and to ensure no changes have been made is tricky task to accomplish because there are lot of things that could go wrong during information transmission, like network errors such as node failing, server down, ISP failure, corruption of checksum bits etc., so it is the duty of security programmers to distinguish between the event where data integrity is lost due to natural causes and one where data integrity is lost by manipulation of unauthorized community of users. So information security programmers strive to achieve data integrity which ensures data accuracy to its core and completeness in all layers including the metadata level. Despite of having confidentiality and integrity of data in check doesn't guarantee the always availability phenomena of data when needed, hence information security guideline upholds the concept of availability which is information

should be readily available whenever deemed necessary by application, end user nodes, and community. The common enemy for successful flow of availability is the denial of service caused by distributed denial of service attack which is an old trick used by the cyber criminals but works effectively, hence information security managers take measures to overcome distributed denial of service attacks.

Apart from the CIA triad there are many more security guidelines to be followed and non repudiation is one that comes up more often. The principle of non-repudiation clearly states as to the parties involved in service with each other cannot deny of receiving a message, or a sort of transaction that took place between them in order to maintain ingenuity. Another factor is authentication where only authenticated users are allowed to access information or applications by providing proper credentials such as user names, passwords, hash keys, one time passwords, biometric data, and other commonly used credential authentication factors used based on the need of company.

2. Related Work

Kan Yang and Qi Han, et al., have proposed the best idea for the development of attitude scheme for big data access [1]. In here they describe that the current encryption policies are not completely doing the job encrypting the data properly, they leave out the job of encrypting the policy regarding the access of the big data. Hence they propose as to encrypt everything so that we may achieve complete privacy regarding these issues with respect big data.

Leyou Zhang, et al., have made the analysis of cipher text "policy attribute based encryption with fast decryption for personal health record system" [2]. It is similar to the previous paper but here complete encryption is not done, instead the part about the file which contains rules about who can access the file is not encrypted, the paper suggests that these access file of health data will leak privacy if we do not encrypt this completely.

Nikunj Joshi and Bintu Kadhiwala, et al., have given a detailed analysis of big data security and privacy issues – a survey [3] where they describe the amount of data that is being generated every single day and how we are lacking the sight of security and data privacy with respect to big data platforms.

Denglong Lv, et al., have given an analysis of review of big data security and privacy protection technology [4]. It is shown in their work that they have concerns related to issues in security of big data and privacy protection and other goals of security that are required to be analyzed from different vantage points.

L. Vijay and C. S. Chandra, et al., have conducted thorough research on "new secretive policy cp-abe reason for big data access control with privacy-preserving policy in cloud computing" [5]. It is stated that they have proposed a new policy for maintaining privacy of data in cloud environment and the necessary steps that we need to take in that regard.

Qi Li and Youliang Tian, et al., have given a comparative analysis of "efficient privacy preserving access of mobile multimedia data in cloud computing" [6]. It is regarded in their work "that cloud based on computing is a new trend emerging in IT environment with huge requirements of infrastructure and resources, and that load balancing is an important aspect of cloud computing environment", and also the privacy of data in cloud.

Alphonse, et al., has discussed an efficient "cipher text policy and encryption for big data access control in cloud computing" [8]. It is shown in work that cloud computing is a known model to structured model that provides services, where resources and data retrieved from a Cloud Service Provider (CSP) with help from a platform based internet web tool or applications, is also stated that multiple users of the cloud generate multiple requests for resources present on the cloud, which may cause a deadlock.

Mingjian, et al., have conducted research on big data for cyber security vulnerability disclosure trends dependencies [9]. It is shown in work that they have proposed an approach where they are integrating the domain of cyber security with that of big data enabling many good attributes from each other.

3. Implementation Details

Firstly, the environment is set up by installing virtual machine workstation, VMware workstation player 12 compatible to host big data platform Hadoop. This followed by setting up of cygwin software, which provides a large collection of GNU and open source tools which helps bring Linux functions to windows 10. This is followed by the set-up of the Hadoop appliance environment inside VMware workstation player. This is followed by setting up of MySQL database with authentication credentials for data privacy. The user data sets are encrypted using AES algorithm and blowfish key, which is in turn stored in Hadoop distributed file system(HDFS), where map reduce algorithm runs to process the bulk data sets.

The user data sets can be shared only to genuine users which can be verified under the cloud ID generated to every authenticated users, and control over the data sets is given completely to the data owners by implementing policies which helps them decide whether to share data with others or not. Since the Hadoop instance runs on VMware workstation player it helps in porting of project to multiple computers and configurations are easily transferred. The architecture for information security tool is shown in Figure 1 where the users like data owner and data consumer are present and perform their operations. Use of workstation player adds an extra layer of security.

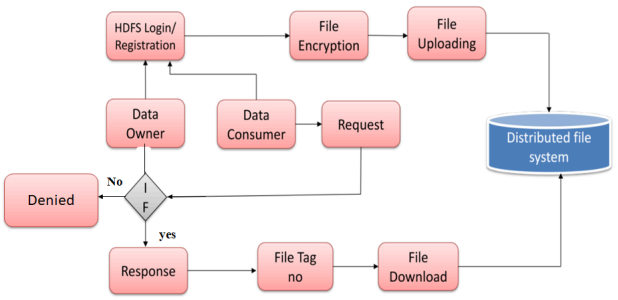

Figure 1: System Architecture of Information Security Tool

Windows 10 operating system will not to run Linux based commands, but it is necessary for the project to have some Linux functionality to manage Hadoop files and other services. Cygwin helps the project by providing the Linux utilities to run on windows operating system. It has large collection of tools that are available in GNU operating systems. POSIX stands for portable operating system interface which is provided by cygwin by utilizing cygwin1.dll which is a dynamic link library that provides source code and necessary data that can be subjected to reuse which helps bring up efficient use of memory and other application programming interfaces that are necessary for the project.

VMware workstation player is free for personal use and it is used in this project to run Hadoop instance. Since Hadoop instance is setup on VMware workstation player it is easily portable and deployed on other systems.

Hadoop instance is deployed to run on VMware workstation player, the Hadoop project by Apache provides reliable solutions to applications in the domain of the computing with respect to scalable and distributive metrics as primary ones in programming model.

Eclipse IDE is where our work space and all configuration files are present; it also provides us with plug in options for connecting to MySQL with JDBC drivers.

4. Experimentation

4.1 Generation of Profile

The generation of profile module is shown in Figure 2 which is a common module for both data owner and consumers. In this module, users have to provide personal information such as name, phone number, address along with creating login credentials like username and password. Upon successful registration a unique ID is generated for users, which is also another login credential require for further use. Both data owner and data consumer can login using those credentials that are username, password and Unique ID.

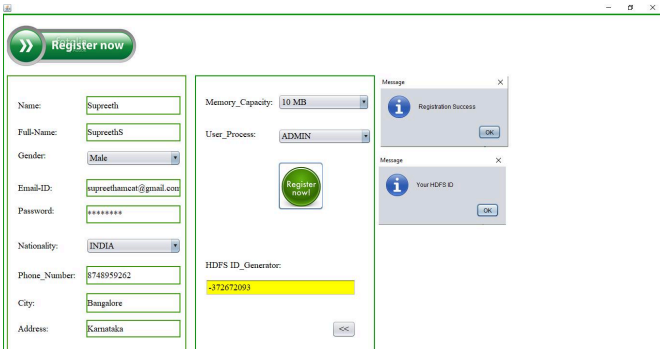

Figure 2: Generation of Profile

4.2 Data Owner Dataset Upload

In this module the data owner who is willing to share his data sets can select one of his data set to be uploaded. Only the data owner can upload data set with proper login credentials, data consumer is not allowed to upload anything to the Hadoop distributed file system, they can only request for the dataset. Control of dataset is hence given to data owner which he can either choose to proceed or not. Before uploading the data set the data owner needs to encrypt the dataset using a symmetric key encryption algorithm called AES also using a blow fish key. After encrypting the data set, it is uploaded to the hdfs file system this is shown in Figure 3.

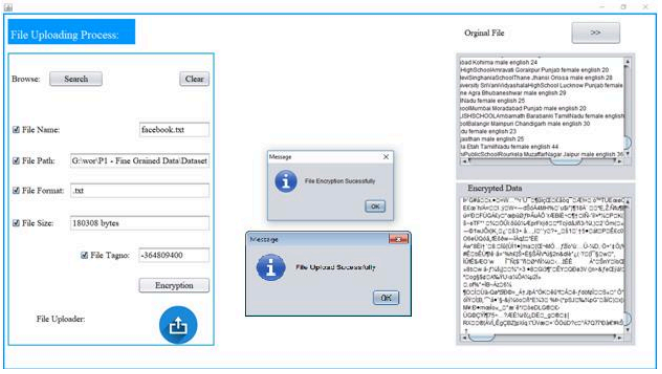

Figure 3: Dataset Encryption and Upload

4.3 MapReduce Process

After data set is uploaded to Hadoop distributed file system, mapper and reducer process starts to process the data set based on key value pairs. The name node acts as master and data node acts as slave, the job tracker allocates the map and reduce task to the task tracker by allocating blocks. All these processes and services are in constant communication during the execution of map reduce process.

4.4 Data Consumer Request and Access

Once the process of MapReduce is complete, each dataset gets a unique number called File Tag No. which is used by data consumer to request data owner for access. The data owner can verify whether the request is from valid user or not by checking the verification. Upon verifying the request as valid the data owner can choose either to provide access to

the file or deny the request as shown in Figure 4.

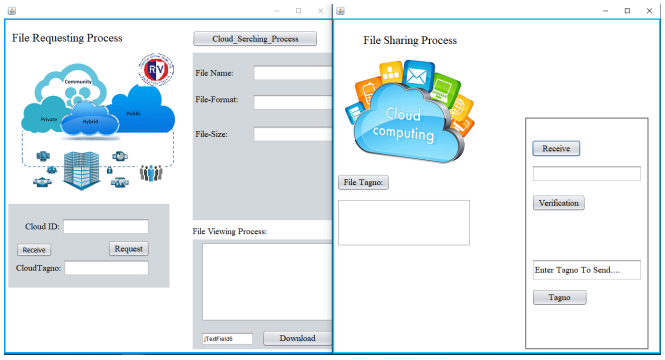

Figure 4: Dataset Consumer Request and Access

5. Result Analysis

Following Metrics are considered in evaluating results of the project, and performance analysis over the results.

Metrics for which lower value means better performance are:

- Mapper-Reducer Runtime
- Dataset Encryption Time
- HDFS Storage Time

Metrics for which higher value means better performance are:

- Storage Array/Memory Unit Size
- RAM Size
- Processor Speed
- Number of nodes in Hadoop Cluster.

Performance Analysis was considered while running modules of both data consumer and data owner. The results are analyzed over various metrics like size, time and other important factors in calculating the performance.
One performance analysis is with respect to time it takes to encrypt the data sets as shown in the Figure 5. The data sets are of two formats, one is text version, .txt, and another is .csv version. Further data sets vary in three sizes in comparison, a group of smaller, medium, and larger sizes. It is noted that data sets of smaller sizes gets encrypted much faster than the data sets of larger sizes. This is due to the fact that we have a slow processor on single node Hadoop that is Intel Celeron running at 1.44GHz frequency.

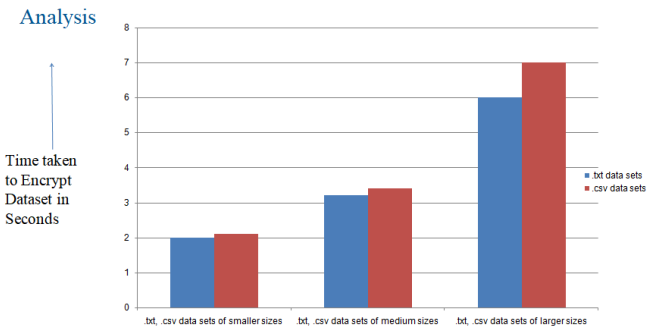

Figure 5: Dataset vs. Encryption Time

Another performance analysis is with respect to time it takes to store the data set in Hadoop distributed file system, where each data set takes different time to get stored on HDFS file system as shown in the Figure 6.

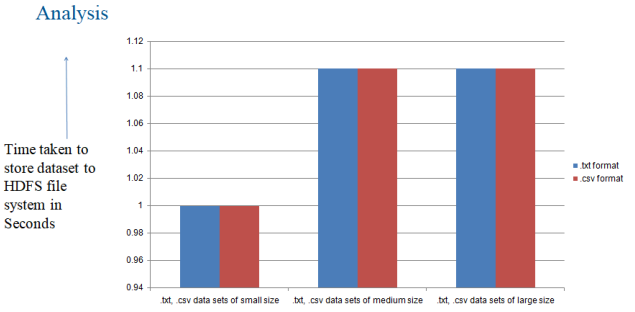

Figure 6: Dataset vs. HDFS storage time

Each data set varies in matter of seconds and milliseconds when it comes to different storage time, as expected data set of smaller sizes gets stored even quicker than the data sets of larger sizes.

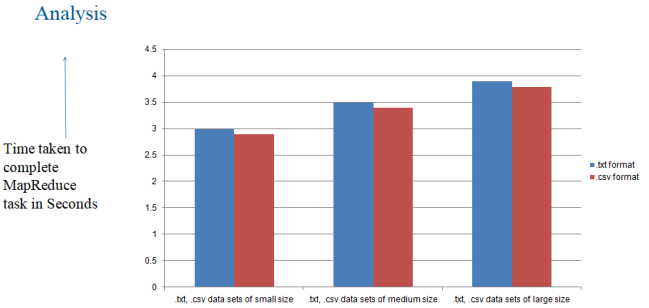

Figure 7: Dataset vs. MapReduce Time

Hadoop is running on a single node. Hence this affects the time taken to map reduce time. More the number of nodes less time it takes to execute mapper reducer tasks. The data sets have similar map reduce time as shown in the Figure 7. This is another performance analysis where dataset is compared against time it takes to map reduce. Moving on we have another performance analysis where the CPU utilization is recorded during the execution of different stages of the project modules as shown in the Figure 8. In here we have comparison between CPU utilization against dataset.

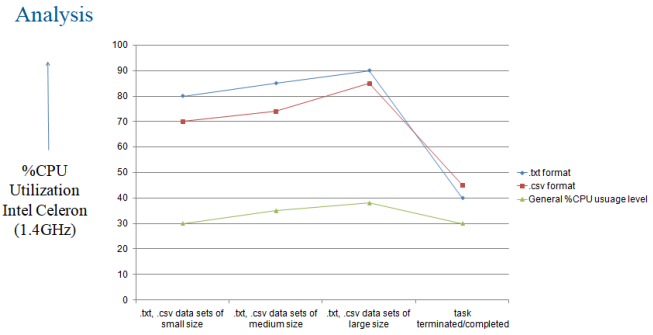

Figure 8: %CPU Utilization vs. Dataset

As we can see, the cpu usage is at 30 to 40 percent when none of the modules are executing. The larger the data set, larger is the cpu utilization taken to process that, we can also see after completion of tasks, the cpu utilization decreases to a 40 to 50 percent. The processor used is very poor in performance; better processor will lead to better throughput and better results. Another performance analysis is with respect to RAM utilization as shown in Figure 9, where data

sets of two types one in .txt format and another in .csv format is compared against the amount of random access memory utilized during the execution of the modules with respective datasets. We can see that RAM utilization is at 25 percent during regular process and it increases to 25.6 percent for smaller data sets and to 26.1 for medium data sets, more than 27 for higher size data sets.

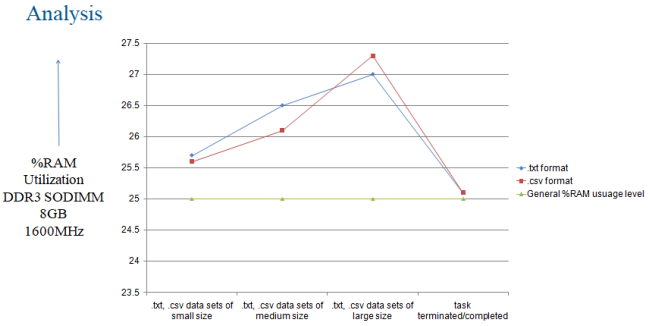

Figure 9: %RAM Utilization vs. Dataset

We can summarize the comparison of data sets of varying sizes and types with measures considered like MapReduce Time vs. Datasets, HDFS Storage Time vs. Datasets and Encryption Time vs. Datasets. Hence we analyzed the performance with varying data set sizes and types by considering different metrics and found that smaller data sets were processed faster due to single node setup of Hadoop and poor hardware specifications.

6. Conclusion

Big Data is a modern paradigm that is used to store high volume data in distributed manner and to process the data using multiple clusters and perform analytics over it. The data privacy of data uploaded to big data servers by data owners is described by service providers where data owners lose physical control over their data. Hence there needs to be a system tool developed by cyber security principles that work on client workstation to enable all these features. Hence this project resulted in providing a safe and secure platform for data owners to share their data while retaining physical control over their datasets on Hadoop platform by means of authentication using login credentials, data encryption using AES algorithm with blow fish symmetric key and extra layer of VMware hosted computer protection. Smaller size data sets were proven to be processed faster than larger one due to single node setup of Hadoop.

References

[1] Kan Yang,Qi Han,Hui Li,Kan Zheng,Zhou Su,Xuemin Shen, “An Efficient and Fine –grained Big Data Access Control Scheme with Privacy-Preserving Policy”, IEEE Internet of Things Journal, vol. 3, April 2019.

[2] Leyou Zhang,Gongcheng Hu, Yi Mu,Fatemeh Razaeibagha, “Hidden Ciphertext Policy Attribute –Based Encryption with Fast Decryption for Personal Health Record System” IEEE Access, vol. 7 March 2019.

[3] Qi Li, Youliang Tian, Yinghui Zhang, Limin Shen “Efficient privacy-access control of mobile multimedia data in cloud computing” IEEE Access, vol. 5 September 2019.

[4] Sucharita Khuntia, Syam Kumar “New Hidden Policy CP-ABE for Big Data Access Control with privacy-preserving policy in cloud computing” 9 th ICCNT, vol. 9Jan 2018.

[5] Praveen Kumar, Syam Kumar, Alphonse, “Atribute based encryption in cloud computing: A survey, gap analysis, and future directions”, Journal of Network and Computer Applications, vol. 6 Feb 2018.

[6] Dong Zheng, Axin Wu,Yinghui Zhang, Quinglan Zhao, “Efficient and privacy-preserving data sharing in Internet of Things with limited computing power” IEEE Access, vol. 1 Jan 2018.

[7] Nikunj Joshi, Bintu Khadiwala, “Big Data Security and Privacy Issues – A Survey”, International Conference on Innovations in Power and Advanced Computing Technologies, vol. 7 Mar 2017.

[8] MingJian Tang, Mamoun Alazab, Yuxiu Luo, “Big Data for Cybersecurity: Vulnerability Disclosure Trends and Dependencies”, IEEE Access, vol. 5 May 2018.

[9] Denglong Lv, Shibing Zhu, Xuazheng Xu, Ran Liu, “A Review of Big DataSecurity and Privacy Protection Technology” IEEE International Conference on Communication Technology, vol. 3 May 2018.

[10] Chamikara, Bertok, D Liu, Camtepe, Khalil, “An efficient and scalable privacy preserving algorithm for big data and data streams”, Computers & Security Elsevier, vol. 4, June 2018.

[11] Youyang Qu , Shui Yu, Jingwen Zhang, Huynh Thi Thanh Binh, Longxiang Gao, and Wanlei Zhou “GAN-DP: Generative Adversarial Net Driven Differentially Privacy-Preserving Big Data”, IEEE Access, vol. 8, August 2019

[12] Rachad Atat, Lingjia Liu, Jinsong Wu, Guangyu Li, Chunxuan Ye,Yang Yi, “Big Data Meet Cyber-Physical Systems: A Panoramic Survey”, IEEE Access, Dec 2016

[13] P. Mell and T. Grance, “The NIST definition of cloud computing,” Recommendations of the National Institute of Standards and Technology – Special Publication 800-145], 2011.

[14] R. Lu, H. Zhu, X. Liu, J. K. Liu, and J. Shao, “Toward efficient and privacy preserving computing in big data era,” IEEE Network, vol. 28, no. 4, pp. 46–50, 2014.

[15] K. Yang and X. Jia, “Expressive, efficient, and revocable data access control for multi-authority cloud storage,” IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 7, pp.1735–1744, July 2014.

[16] H. Li, D. Liu, K. Alharbi, S. Zhang, and X. Lin, “Enabling fine-grained access control with efficient attribute revocation and policy updating in smart grid,” KSII Transactions on Internet and Information Systems (TIIS), vol. 9, no. 4, pp. 1404–1423, 2015.

[17] K. Yang, Z. Liu, X. Jia, and X. S. Shen, “Time-domain attribute-based access control for cloud-based video content sharing: A cryptographic

approach," IEEE Trans. on Multimedia (to appear), February 2016.
[18] B. Waters, "Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization," in Proc. of PKC'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 53–70.
[19] H. Lin, Z. Cao, X. Liang, and J. Shao, "Secure threshold multi authority attribute based encryption without a central authority," in Proc. Of INDOCRYPT'08. Springer, 2008, pp. 426–436.
[20] T. Nishide, K. Yoneyama, and K. Ohta, "Attribute-based encryption with partially hidden encryptor-specified access structures," in Applied cryptography and network security. Springer, 2008, pp. 111–129.
[21] J. Li, K. Ren, B. Zhu, and Z. Wan, "Privacy-aware attribute-based encryption with user accountability," in Information Security. Springer, 2009, pp. 347–362.
[22] D. Boneh and B. Waters, "Conjunctive, subset, and range queries on encrypted data," in Theory of cryptography. Springer, 2007, pp. 535–554.
[23] J. Katz, A. Sahai, and B. Waters, "Predicate encryption supporting disjunctions, polynomial equations, and inner products," in Advances in Cryptology–EUROCRYPT'08. Springer, 2008, pp. 146–162.
[24] J. Lai, R. H. Deng, and Y. Li, "Fully secure cipertext-policy hiding cp-abe," in Information Security Practice and Experience. Springer, 2011, pp. 24–39.
[25] L. Lei, Z. Zhong, K. Zheng, J. Chen, and H. Meng, "Challenges on wireless heterogeneous networks for mobile cloud computing," IEEE Wireless Communications, vol. 20, no. 3, pp. 34–44, 2013.
[26] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, "Big data-driven optimization for mobile networks toward 5g," IEEE Network, vol. 30, no.1, pp. 44–51, 2016.
[27] Z. Su, Q. Xu, and Q. Qi, "Big data in mobile social networks: a qoe-oriented framework," IEEE Network, vol. 30, no. 1, pp. 52–57, 2016.
[28] H. Li, D. Liu, Y. Dai, and T. H. Luan, "Engineering searchable encryption of mobile cloud networks: when qoe meets qop," IEEE Wireless Communications, vol. 22, no. 4, pp. 74–80, 2015.
[29] H. Li, Y. Yang, T. Luan, X. Liang, L. Zhou, and X. Shen, "Enabling fine-grained multi – keyword search supporting classified sub-dictionaries over encrypted cloud data,", IEEE Transactions on Dependable and Secure Computing [DOI: 10.1109/TDSC.2015.2406704], 2015.
[30] K. Frikken, M. Atallah, and J. Li, "Attribute-based access control with hidden policies and hidden credentials," IEEE Trans. on Computers, vol. 55, no. 10, pp. 1259–1270, 2006.
[31] S. Yu, K. Ren, and W. Lou, "Attribute-based content distribution with hidden policy," in Secure Network Protocols (NPSec'08 Workshop). IEEE, 2008, pp. 39–44.
[32] J. Lai, R. H. Deng, and Y. Li, "Expressive cp-abe with partially hidden access structures," in Proc. of ASIACCS'12. ACM, 2012, pp. 18–19.
[33] J. Hur, "Attribute-based secure data sharing with hidden policies in smart grid," IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 11, pp. 2171–2180, 2013.
[34] A. Beimel, "Secure schemes for secret sharing and key distribution," Ph.D. dissertation, Israel Institute of Technology, Technion, Haifa, Israel, 1996.
[35] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," Communications of the ACM, vol. 13, no. 7, pp. 422–426, 1970.
[36] K. Yang, X. Jia, and K. Ren, "Secure and verifiable policy update outsourcing for big data access control in the cloud," IEEE Trans. Parallel Distrib. Syst., vol. 26, no. 12, pp. 3461–3470, Dec 2015.
[37] C. Dong, L. Chen, and Z. Wen, "When private set intersection meets big data: an efficient and scalable protocol," in Proc. of CCS'13. ACM, 2013, pp. 789–800.
[38] A. Sahai and B. Waters, "Fuzzy identity-based encryption," in Advances in Cryptology – EUROCRYPT 2005, R. Cramer, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 457–473.
[39] J.Bethencourt, A.Sahai, and B. Waters, "Ciphertext - policy attribute-based encryption," in Security and Privacy, 2007. SP '07. IEEE Symposium on, may 2007, pp. 321 –334.