 CMC, vol.59, no.2, pp.361-374, 2019

Security and Privacy Frameworks for Access Control Big Data Systems

Paolina Centonze[1,*]

Abstract: In the security and privacy fields, Access Control (AC) systems are viewed as the fundamental aspects of networking security mechanisms. Enforcing AC becomes even more challenging when researchers and data analysts have to analyze complex and distributed Big Data (BD) processing cluster frameworks, which are adopted to manage yottabyte of unstructured sensitive data. For instance, Big Data systems' privacy and security restrictions are most likely to failure due to the malformed AC policy configurations. Furthermore, BD systems were initially developed toped to take care of some of the DB issues to address BD challenges and many of these dealt with the "three Vs" (Velocity, Volume, and Variety) attributes, without planning security consideration, which are considered to be patch work. Some of the BD "three Vs" characteristics, such as distributed computing, fragment, redundant data and node-to node communication, each with its own security challenges, complicate even more the applicability of AC in BD.
This paper gives an overview of the latest security and privacy challenges in BD AC systems. Furthermore, it analyzes and compares some of the latest AC research frameworks to reduce privacy and security issues in distributed BD systems, which very few enforce AC in a cost-effective and in a timely manner. Moreover, this work discusses some of the future research methodologies and improvements for BD AC systems. This study is valuable asset for Artificial Intelligence (AI) researchers, DB developers and DB analysts who need the latest AC security and privacy research perspective before using and/or improving a current BD AC framework.

Keywords: Big data, access control, distributed systems, security, privacy.

1 Introduction

An article published by an on-line News [MeriTalk (2018)], stated that every two days, people produce about five-exabytes of data, which is almost all the data created until 2003. A journal published in the [Data IQ News (2018)] estimated that the universal data will get close to one billion terabytes by 2020. Definitely, the huge amount of data generated is impacting how data is stored and created to obtain efficient computing speed or for functionalities and for organizing data testing.

The terminology of Big Data (BD) defined by Hu et al. [Hu, Grance, Ferraiolo et al. (2014)] to a high quantity of data that which are incredibly challenging to analyze them with conventional methodologies and technologies. Moreover, the populated data are in

[1] Iona College, Iona College, New Rochelle NY 10801, USA.

[*] Corresponding Author: Paolina Centonze. Email: pcentonze@iona.edu.

different structure and they mostly include network data sharing, videos, sensor data, computing-server log files. These data are created both statically and on the fly by people due to the sophisticated features of computing. Consequently, this exponential increase of data volume introduces even more challenges to our current IT data systems, such as scaling, data storage, deep data analysis, and most of all data security. Big Data are generated in very large volume data sets, in high speed and have different data formats. So, BD technology must be able to process, collect and analyze huge data sets and be able to deal with todays' current data management, analysis, and security frameworks. Currently, dedicated systems for DB processing are used as basic solutions. However, to completely manage BD scalability and data speed, many of the current BD processing frameworks use enormous parallel software executing on computers for distributed frameworks, which could use columnar databases and of BD management results. In 2016, BD analysts predicted that using technology for BD can lead to decrease IT expenses by 49%. In fact, BD systems is slowly redeveloping current data technologies and practices. Nowadays, 65% of universal trademarks, are ready to invest in Bog Data systems to keep up with the BD computing field. In fact, In the last few years, BD computing are experiencing a pick in demand in both academia and industry sectors. Yet, BD computing still does not completely enforce policies and solutions to guarantee users' data from mis-use and abuse. One of the security principals, as a prevention from attacks, which could be put it in development and adopted to secure data-sharing method, is Access Control (AC). In fact, institutions and corporations that need to gather data from many different venues, may put themselves in a very jeopardy situation due to the possibility of leaking user's sensitive data. For example, until 2016 in Hadoop DB system, it was only feasible to do file-level AC policy, which provided users to gain inside to data, which were based on attributes in data-set or the user's role, which is intricate, since the massive data quantity and many different data formats, such as structured, unstructured, and semi-structured). However, in 2017 Ashwin et al. [Ashwin, Hong, Thomas et al. (2017)] came up with a new AC framework named Content Sensitivity Based Access Control (CSBAC) methodology, which ensures AC policy on the fly based on the sensitivity of the data. Detecting a BD misappropriate and exploit when already has occurred, there are very high chances that nothing else can be done to protect user's data, so an AC enforcement may have had already by-passed by malicious intruders. These are still some of the main challenges in current AC BD frameworks since they don't have good measurements to be installed to guarantee hacker's attacks from miss appropriately accessing and manipulating user's data. The AC BD achievements of CSBAC framework, as well as other AC BD frameworks, are explained more in detail in Section 3 in of this paper.

There are many laws, such as the Sarbanes-Oxley Act [Investopedia (2018)], which protects corporate financial information and the Health Insurance Portability and Accountability Act (HIPPA) [Edemekong and Haydel (2018)], that enforces policies and restrictions of health patients' data when data are shared across data systems. Until 2017, obeying to these complicated level data sharing in Hadoop implantation was impossible. However, especially in the past couple of years, the AC BD research community has invested more efforts to add in Hadoop a fine-grained AC mechanism, which allows perfect data distribution and not conceding sensitive information. But, developing a fine-grained AC methodology for BD is very complex. For instance, traditional

methodologies, such as Access Control Lists (ACLs) and Role Based Access Control (RBAC) are not appropriate for BD systems, since they will be clumsy to administer since the data and users raise exponentially.

This paper aims to contribute in the field of Access Control (AC) for Big Data. Specifically, this work has the following contributions: 1) an overview of some of the latest (2014-2018) state-of-the-art literature dealing with security and privacy issues, and challenges for BD AC systems, 2) a study of the latest research solution frameworks to reduce privacy and security issues for AC BD systems, 3) an overview of possible improvements and extensions AC BD frameworks. Moreover, this study is also a great asset for Artificial Intelligent (AI) researchers, DB developers and BD analysts before they need to use and/or improve a current BD AC framework. To our best knowledge, there is no such latest study that includes all the three contributions mentioned above, but rather they discuss mostly older AC BD frameworks, without analyzing some of the frameworks' restrictions and their future directions.

2 Security and privacy challenges in access control big data

The paper of Gupta et al. [Gupta, Pandhi, Bindu, et. al (2016)] reveals that security is the greatest challenge in deploying BD, and that the major struggle dealing with and using technologies for BD data is to ensure data security. AC policy is one of the essential techniques for management enforcement, which allows organizations to protect their BD to address security and privacy permissions. But, the "three Vs" of data are very complex in current BD systems, which were neither intended nor made with AC capabilities. For instance, most of them donot properly manage the creation, use, and dissemination of BD process. Thus, they either lead conflict in working with immensely rigid policies, or are a very huge risk for loosing data by over permitting data access. Although, some security features were added in the latest BD systems, an essential security component, such as Access Control (AC) to enforce BD protection for users and from inside network attack still residues a possible problem. Hu et al. [Hu, Grance, Ferraiolo et al. (2014)] developed a general-purpose AC scheme for distributed BD processing clusters. The scheme extends the general Big Data (BD) model. More of this scheme's achievements are discussed in the following section.

BD AC has not only to mandate Access Control policies on data exit the Master System (MS) [Hu, Grance, Ferraiolo et al. (2014)], but it must also oversee access to the Cooperating Systems (CSs) [Hu, Grance, Ferraiolo et al. (2014)]' resources. Contingent on the sensitivity of the data, it must also check that BD applications, the Master System (MS), and the Cooperating Systems (CSs) have authority to get into the data that they are examining, and work with the access to the distributed BD process and data from local users. In fact, the features of BD distributed computing model belong to an exclusive group of difficulties for BD AC, which need a special group of notions and attentions. Specifically, handling the BD's "three Vs" characteristics introduces more challenges for a system including Access Control (AC) implementation, because the problems are usually controlled by distributed computing, fragmented and redundant data and node-to-node communication, each with its security challenges [Hu, Grance, Ferraiolo et al. (2014)]. For instance, distributed systems are more vulnerable to attacks than centralized

data repositories, since BD data is administered anywhere, assets are offered allowing enormously parallel computation between Master System (MS) and Cooperating Systems (CSs). In addition, fragmented data increases complexity to the data integrity and confidentiality. Moreover, Master System (MS) and Cooperating Systems (CS) in node-to-node communication use unsecure protocols, such as Remote Procedure Call (RPC) over TCP/IP.

Unfortunately, as of today, there are still very few tools and solutions to that enforce security BD process restrictions and policies in a reasonable time frame and cost. Traditional solutions that still relay on perimeter security (i.e., firewalls, intrusion detection, and prevention technologies) are unable to adequately secure BD cluster. In addition, although the Hadoop community added more robust security controls, such as Kerberos, firewalls, and basic HDFS (Hadoop Distributed File System) permission, yet these additions did not solve completely all the AC challenges, since these technologies are not suitable for DB systems by allowing possible hackers to by-pass AC restrictions [Hu, Grance, Ferraiolo et al. (2014)]. For example, even though, Apache Accumulo and HBase added some AC based on user and group permission by ACLs (AC Lists), however ACLs solutions are demonstrated to be limited, because many institutions use adaptable and changeable policies on security specifications for users and business systems. Moreover, as of today, only file-level access control is possible in Hadoop system by granting user's data accessibility established on the attributes in a dataset or the user's role, which is very complex due to the infinite data and different data formats. Ashwin et al. [Ashwin, Hong, Thomas et al. (2017)] proposed an AC solution, that inflicts on the fly AC guidelines depending on the sensitivity of the data. The solutions of these frameworks are explained in the following section.

Since the BD era, SQL-based databases have been replaced with NoSQL-based systems to be able to handle the voluminous, unstructured, complexity of BD data. These systems are designed to permit data input without the need of a predefined schema. Many BD systems only allow AC at the schema level without a precise granularity that includes attributes of BD. So, NoSQL system must handle the ability to create and administer AC information, for example using NoSQL systems and have above them an application layer. Moreover, security issues, such as data communication, data storage, and data certification in distributed networks continue to be the main focus in the AC BD research community. In order to solve these issues Gupta et al. [Gupta, Pandhi, Bindu et al. (2016)] introduced a conceptual data-space model based on the mixture of Role Based Access Control (RBAC) and Hadoop Distributed File System (HDFS). Their proposed framework is intended for private institutions and allow private organizations to save both sensitive and public data on the same cloud computing. The contributions of this idea are explained more in detail in the following section.

Another adjustable, and fine-grained Access Control used for not secure system is called multi-authority attribute-based encryption. But, this scheme has downsides for example, revocation, which is considered to be one of the top difficulties. It consists, to eliminate users from the system or some of user' attributes to restrict to gain control on user's information. As of today, well used solutions, as time-based and proxy methodologies advise to feature a termination time to users' keys or to trust on a semi-trusted proxy to

withdraw users. However, in the time-based methodologies, the withdrawal is not instantly and the retracted users can still get into the data up to the next key generation phase, while proxy-key frameworks donot get to a fine-grained access and the users are not allowed to gain into their data if the proxy is not online. However, in 2018, Imine et al. [Imine, Lounis and Bouabdallh (2018)] proposed a new and proficient withdrawal framework for dispersed attribute-based systems in order to solve these restrictions. Their proposed framework certifies changeable and fine-grained Access Control and averts security degradations. More of the achievements of this framework is analyzed in the following section.

Also, in many other healthcare sectors, such as medical, biomedical, there is an exponential increase in huge quantity of data that must be managed and analyzed in order to better cure diseases. However, these sectors are still very hesitant in switching to big data technologies, since the security and private issues are still a very crucial concerns for the privacy of patients' data residing on big data technologies. For example, in 2016 CynergisTek [CynergisTek (2016)], published the Redspin's 7th annual breach report: Protected Heath Information (PHI) which stated cyber threats on healthcare providers raised by 320% in 2016, and 81% of record data leaks in 2016 resulted specifically from cyber attacks. Access Control is one of the securities and privacies technologies used in BD health care systems, in fact, Role Based Access Control (RBAC) and Attribute Based Access Control (ABAC) are the most well used frameworks used for Electronic Health Record (EHR). However, RBAC and ABAC frameworks still reports some downsides when these models are deployed alone in healthcare technologies [Huseyin, Kantarcioglu, Pattuk et al. (2114)]. In 2014, Zhou et al. [Zhou and Wen (2014)] proposed a cloud-oriented storage efficient dynamic access control scheme cipher text based on Ciphertext-Policy Attribute-Based Encryption (CP-ABE) and symmetric encryption algorithm, such as Advanced Encryption Standard (AES). Abouelmehdi et al. [Abouelmehdi, Beni-Hssane, Khaloufi et al. (2017)], concluded that to fulfill necessities of fine-grained Access Control, and yet still preserving both security and privacy, it is necessary to adopt these models in with other security methodologies, such as encryption and Access Control techniques. In addition, the paper [Abouelmehdi, Beni-Hssane, Khaloufi et al. (2017)] also discusses that there are some other BD techniques used in healthcare to ensure patients' privacy, such as Hiding a Needle in a Haystack [Jung and Park (2014)], Attribute Based Encryption Access Control, Homomorphic encryption, Storage path encryption, etc., however, security and privacy problems are always imposed even when using all these techniques.

Another challenge that is increasing, even more, in the Big Data ecosystem is the struggle between data mining and data privacy protection. Well used information security methodologies focus on protecting the security of attribute values without semantic association. Meijuan et al. [Meijuan, Jian, Lihong et al. (2018)] proposed a Data Access Control framework for single users based on the semantic integration nature of XML data.

All the security and privacy downsides for Access Control Big Data discussed in this section, which emphasizes even more the crucial necessity for BD analysts to invest even more in big data to protect users' information and fight the increase threats which are present to industries, businesses, governments, and healthcare communities. The

following section also analyzes the latest AC BD frameworks' contributions to reduce some of the security and privacy challenges discussed in this section.

3 Frameworks for access control big data

Researchers have been proposing many different solutions providing AC for Hadoop, which as of today, is still one of the most used BD management systems. However, in order to guarantee AC protection, these solutions completely need data and the sharing of the data only by the data's owner. In 2013, the work completed in Cavoukian [Cavoukian (2009)], implemented a Content Based Access Control (CBAC) methodology for Hadoop with function built on data content itself. In 2017, Ashwin et al. [Ashwin, Hong, Thomas et al. (2017)] proposed the Content Sensitivity Based Access Control (CSBAC) framework, that relates to the CBAC solution. However, the main difference between these two frameworks is that the CBAC framework usages the top-*k* matches by comparing the data set similarities and a based set to grant access control. On the contrary, the CSBAC framework in order to grant AC protection relies only on the sensitivity data set without using the base set. In addition, the CSBAC framework has one more advantage in which it takes into consideration if the data change their sensitivity when more data are added, while the CBAC solution is not able to take this very dynamic change into its analysis. The CSBAC framework is a new solution in the context of BD security and privacy since it is able to make AC protection only by depending only on its own sensitivity data. Moreover, the CSBAC solution is also an augmentation of Sensitive Data Detection (SDD) [Ashwin, Liu, Thomas et al. (2015)] framework, which was also created by the same authors who created the CSBAC framework. In the CSBAC system, the authors used data gain to sense data sensitivity in its place of the information value equation, which they applied in the SDD framework. The authors used this approach because approximating coefficients for the information value equation needs a lot of time and it changes for every data set, so it is not consistent. Furthermore, the CSBAC framework fulfills the essential seven principles of Privacy by Design (PbD) [Gupta, Patwa and Sandhu (2018)], and it is a hybrid Access Control methodology, which needs both attributes to understand sensitivity and user role to enforce Access Control decisions. In addition, the CSBAC framework is practical, but not dynamic [Ashwin, Liu, Thomas et al. (2015)] as it prevents hackers to gain access to sensitive data, and privacy is implanted as the foundation of this methodology. Moreover, the data owners or consumers have minimal work to enforce AC mechanisms and security. Finally, the CSBAC framework gives lifecycle protection end-to-end by guaranteeing security on sensitive data, but they have to be in the HDFS. Fig. 1 [Ashwin, Hong, Thomas et al. (2017)] depicts the CSBAC framework. However, one downside of the CSBAC framework is that there is a small overhead which is due to the extra computation for ensuring that sensitive information are not accessed by hackers by enforcing the correct use and not abusing user's data.

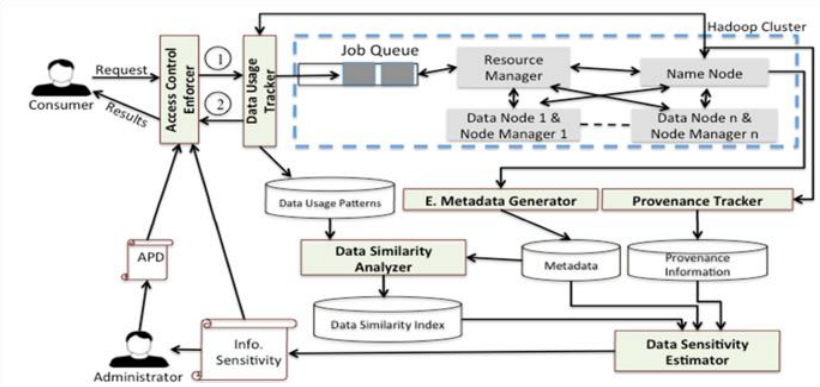

Figure 1: CSBAC framework

Usually, AC consists of methodologies which enforce only authorized users who can gain inside the shared data. In addition, AC mechanisms deal with revoking users from accessing data. The withdrawal can be a feasible solution when all users have the permission to access only one data-sharing space. But, the AC revocation becomes more challenging when users must gain inside to many domains which require different AC rights, or when users need to log off the system more often, or when new AC rights increase. Many well-known AC frameworks must trust the server providers, because the data owners don't have any control on their data, on the contrary server providers have completely control and gain completely access to users' data. So, there is a great need for more dynamic AC solutions to protect data owners form service providers and to give more control to data owners on their own data. For example, the multi-authority attribute-based encryption is a methodology which gives a dispersed, dynamic and fine-grained AC in untrusted data systems. Yet, this framework has some problems in the AC withdrawal, which is one of its main downsides. In 2018, Imine et al. [Imine, Lounis and Bouabdallh (2018)] proposed a new methodology which implements revocation for disperse attribute-scheme for multi-authority systems, such as cloud-based systems in which DB are stored. Specifically, this scheme brings the following advantages: a) Re-keying is not needed because of the secret sharing of the revocation's allocation. So, this new feature enforces the revoked user not to gain access to the original cipher-text, b) The user's revocation happens immediately, by changing the hidden group's attributes so that only permitted users can actually see the new secret, c) The computation factors are very minimal for recreating the secretes of the attributes, d) The feature is very dynamic for situations in which users need to join and leave the group's attributes. e) Data-sharing is feasible in a personal domain. This allows, the data owner to share its own data to an outside server and restrict revocations when needed, f) No need to add new computing resources in the system, since the data sharing and the management of user' revocations are done in a completely dispersed data system. However, the authors of this framework already envisioned an improved solution to reduce the cost of encryption and description of the entities, and yet still maintaining the same security level. More improvements of this frameworks are mentioned in the following section.

Since the exponential growth of cloud computing, BD has become pervasive and

essential for many application domains, which leads to many significant problems dealing with security, reliance, threat, eco-efficiency, price and lawful downsides. Security issues exist for Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) cloud systems. The security of cloud storage becomes weak to transmission security, AC, database, and data certification. In fact, because the users' data reside in a cloud space with no access control, the user has no control to oversee their own data. In 2015, Gupta et al. [Gupta, Pandhi, Bindu et al. (2016)] proposed a new conceptual framework which is an augmentation of Hadoop to solve some of the security problems, such as communication, data allocations, and data authentication in disseminated networks, which are greatly used by BD systems. Their solution is for private cloud of an institution and permits institutions to save confidential and public data in the same cloud storage. Moreover, this methodology for cloud space isolates data centered on roles assigned in the grading of the institution which are promoted access to the data. Furthermore, it considers how worthen is the data for the institution and checks when data have been accessed during a specific time frame. The solution integrates both Role Based Access Control (RBAC) and Hadoop Distributed File System (HDFS), it also uses a standardization feature to use the assets cost-effectively by equally allocating data through unlike server systems. The final experiments done on this model concluded the following outcomes: 1) The standardization ensures an equal dissemination of documents on many diverse security servers, allowing the usage of means in a very effective way. 2) In situation of large data volumes, the normalization function is called iteratively, therefore resulting in re-evaluation for file' storage, so providing more security for the file system. 3) It provides more security to the file systems, since the file storage dynamically modifies as the number of times increases while file is accessed with one day. Since the authors of this scheme proposed only a conceptual model, the next step that they would like to do is to implement the actual model in the Hadoop Framework.

In 2018, Gubta et al. [Gupta, Patwa, and Sandhu (2017)] proposed a fine-grained Attribute-Based Access Control model (HeABAC), which satisfies the security and privacy necessities of multi-tenant Hadoop environment. The scheme is an extension to the existing Hadoop Access Control model (HeAC), which includes the authorization capabilities of core Hadoop (2.x), two important security projects, such as Apache Ranger (version 0.6), Sentry (version 1.7.0), and RBAC extension object-tagged role-based access control (OT-RBAC) model [Wenrong, Yang and Luo (2013)], a previously proposed work done by the same authors. In addition, in Gupta et al. [Gupta, Patwa and Sandhu (2017)] the authors proposed an implementation approach for HeABAC model, as depicted below in Fig. 2. [Gupta, Patwa and Sandhu (2017)], using open source Apache Ranger, context enricher and condition evaluators. Furthermore, the authors of this framework proposed that context enricher will not only be used for enriching user information, but also for services and objects in the access request. As shown below in Fig. 2, the security administrator will add text files for different users, objects and services into the system with their relevant attributes. These files will then be used by context enricher implemented, which will add attributes of users, services and objects in the access request. Similarly, condition evaluators also need to be extended to incorporate the attributes of objects and services in policies, which will be also evaluated when the

enriched access request with attributes is checked against a defined security policy. In this model, the administration of policies is done through the central policy server, while the decisions and enforcements are made by Apache Ranger security plugins attached with the individual services as shown in Fig. 2.

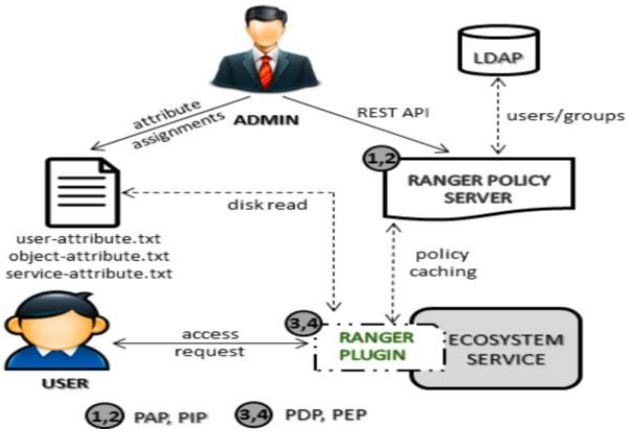

Figure 2: HeABAC Implementation

Big Data AC lacks security protection on data cooperation among processing systems, which depend on components under disperse AC controlling. In 2014, Hu et al. [Hu, Grance, Ferraiolo et al. (2014)], developed a general-purpose AC scheme for dispensed BD cluster systems. Authentication is most of the time done by Master System (MS) and Cooperating Systems (CSs) independently for both BD and non-BD systems. Yet, the first reason is that BD is analyzed by allocating its resources and information from Master System (MS) and Cooperating Systems (CSs), and the second reason is that data has no standard scheme database management, so DB AC requires more AC features than non-BD systems. A BD cluster is a component of an enterprise computing that needs Master System's AC feature to be imbedded along Cooperating System's. This model is build up the general BD framework explained in detailed in their paper [Hu, Grance, Ferraiolo et al. (2014)]. In addition, their framework contains AC functionalities to enforce the BD AC requirements, such as Security Agreement (SA), Trust CS List (TCSL), Master Systems AC Policy (MSP), Cooperating System AC Policy (CSP), and Federated Attribute Definitions (FAD), which are all explained in their paper [Hu, Grance, Ferraiolo et al. (2014)]. Fig. 3 [Hu, Grance, Ferraiolo et al. (2014)] depicts the authors' generic BD AC architecture. Fig. 4 [Hu, Grance, Ferraiolo et al. (2014)] depicts their BD AC control/mange domain where Master System (MS) AC component is imposed cooperatively by data in the Service Agreement (SA), Trust CS List (TCSL), Master Systems AC Policy (MSP), and Federated Attribute Definitions (FAD) records, which are all aligned and directed by Master System (MS). Cooperating System (CS) AC component is imposed cooperatively by data in Master Systems AC Policy (MSP), Federated Attribute Definitions (FAD), and Cooperating System AC Policy (CSP) entries; however only Cooperating System AC Policy (CSP) entries are configured and managed

by Cooperating System (CS). In addition, the authors discussed effective concerns that needs to be solved also for many BD AC systems, such as confidence establishing, content attributes, AC auditing, and combining them with cloud environment.

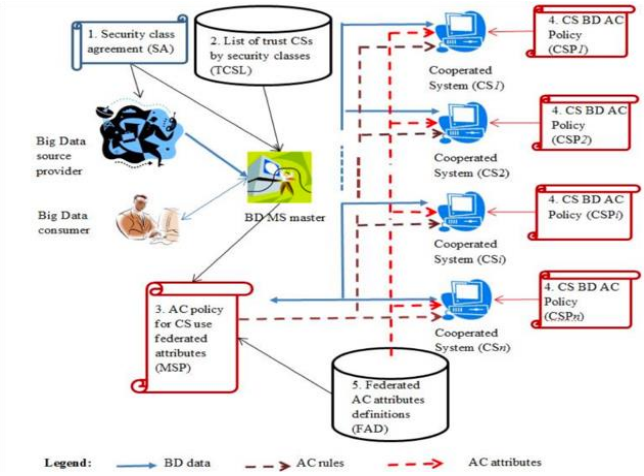

Figure 3: Generic BD AC architecture

Their model is designed so that it can be utilized by disperse system-based BD AC, and prevents from the foundational AC model which depends on components, such as subjects, objects, actions, and ecosystem situation, which are attributes for constructing blocks for Attribute-Based Access Control (ABAC) models, to affirm access permission. However, this scheme has few working and development problems for applicable applications. These problems are knotted to the BD system, and neesa be conducted based on the BD security requirements. In summary, this BD AC framework is built on confidence among BD providers and BD Master System (MS), and thus, among Master System and Cooperating Systems (CSs). The scheme demonstrates that no unauthorized privileges, either from Master System, or Cooperating Systems are possible.

In the Big era DDoS is also one of the major securities and privacy issues. Jieren et al. [Jieren, Ruomeng, Xiangyan et al. (2018)] proposed an unusual network flow characteristics sequence estimate framework, which can be applied as a DDoS attack indicator in Big Data systems. The authors came-up with a network flow abnormal index as Participatory Disaster Risk Assessment (PDRA) with the proportion of old IP addresses, the increase of the new IP addresses, the percentage of new IP addresses to the old IP addresses, and the average accessing rate of each new IP address. They designed an IP address database utilizing sequential storage example which has a constant time computation. The Auto-Regressive Integrated Moving Average (ARIMA) forecast module will be run only if, the quantity of constant PDRA sequence answer, which all surpass an PDRA Abnormal Threshold (PAT), gets a specific preset limit. Soon after that, compute the likelihood, that is, the ratio of forecasting PDRA sequence value, which surpass the PAT. Lastly, the authors identified the DDoS threat built on the irregular likelihood of the predicting PDRA sequence. In their work, the theory and the

experimental results demostrate that the solution that they developed can successfully decrease the computed assets used, detecting DDoS threat during its very beginning stage with a greater discovery rate and less false alarm rate.

The challenge among data mining and data privacy security it is even more prominent in the Big Data ecosystem. Established data security emphases on guarding the security of attribute values without semantic connotation. The data privacy of Big Data is mostly limited in the efficient usage of information, and not leaking any user's sensitive data, taking into account the semantic connotation, decent security restrictions for privacy are mandatory. Semi-structured and self-descriptive XML (eXtensible Markup Language) is a standard form of data structure and in Big Data environments. Meijuan et al. [Meijuan, Jian, Lihong et al. (2018)] proposed a Data Access Control solution for distinct users, based on the semantic integration nature of XML data, and across the semantic reliance among data and the incorporation method from lowest to highest, and the comprehensive graphic scope of upturned XML structure is gathered. Their analysis outcomes prove that their framework enforces the privacy with an access complexity very efficiently.

The following section summarizes the contributions of our paper and some future improvements of the BD AC schemes and frameworks discussed in this paper.

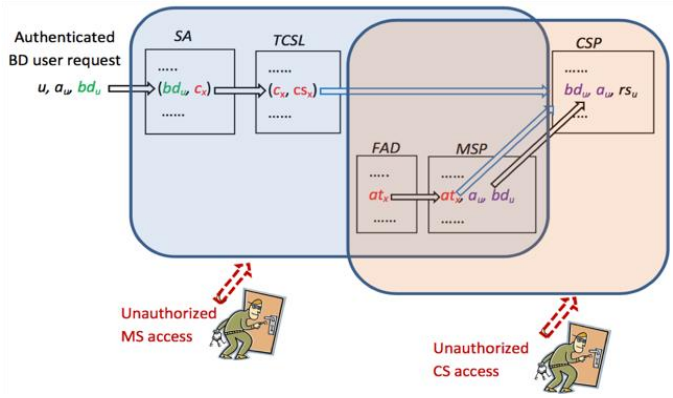

Figure 4: BD AC control/mage domain

4 Summary and future directions

In this paper, we aimed to contribute in the field of Access Control (AC) for Big Data. Specifically, this paper has the following contributions: 1) an overview of some of the latest (2014-2018) state-of-the-art literature dealing with security and privacy issues, and challenges for BD AC systems, 2) a study of the latest research solution frameworks to reduce privacy and security issues for AC BD systems, 3) an overview of possible improvements and extensions of the frameworks discussed in this paper. In addition, we see that this study is a great asset to Artificial Intelligent (AI) researchers, DB developers and BD analysts before they need to use and/or improve a current BD AC framework. To our best knowledge, there is no such latest study that includes all the three contributions above, but rather related studies refer to mostly older AC BD frameworks, without analyzing some of the frameworks' restrictions and their future improvement

directions. Specifically, we described and analyzed the following AC BD frameworks in this specific order: Content Based Access Control (CBAC) [Cavoukian (2009)], Content Sensitivity Based Access Control (CSBAC) [Ashwin, Hong, Thomas et al. (2017)], Role and Access Based Data Segregator [Gupta, Pandhi, Bindu et al. (2016)], Revocable Attribute-based AC in multi-authority systems, Fine-grained Attribute-Based Access Control model (HeABAC) [Gupta, Patwa and Sandhu (2017)], and a general-purpose AC scheme for BD framework [Hu, Grance, Ferraiolo et al. (2014)].

As we briefly mentioned in Section 3 of this paper, most of these frameworks are subject to future improvements. For instance, the CSBAC [Ashwin, Hong, Thomas et al. (2017)] framework has a minimal overhead in order to prevent unauthorized users to misuse and abuse users' data, however the authors of this framework are planning to make the following two important enhancements in their framework: 1) advancing the theoretical methodology for assessing data sensitivity and 2) a deeper analysis performance of individual components. Moreover, the authors of the Revocable Attribute-based AC [Imine, Lounis and Bouabdallh (2018)] framework, would like to improve the cost efficiency for encryption and description on the components with very few resources, for instance they would like to experiment the feasibility of adding proxy servers in the system. Furthermore, since the Role and Access Based Data Segregator [Gupta, Pandhi, Bindu et al. (2016)] is a conceptual model, the authors are planning to implement this model for Hadoop framework. In addition, in this framework [Gupta, Pandhi, Bindu et al. (2016)] the repeated normalization of storage critically adds an overhead. To improve the overhead, the limit size of the buffer must be changed to a value that can still perform a smallest amount of normalizations. In addition, the authors of this framework [Gupta, Pandhi, Bindu et al. (2016)] would consider for optimization of MapReduce work to to eliminate the extra time. Moreover, their improved model will be able to identify additional features for assessing storage as to better construct the formulation and be more suitable. Finally, the authors of the Fine-grained Attribute-Based Access Control (HeABAC) [Gupta, Patwa and Sandhu (2017)] model, are planning an extension of this scheme, which may involve data ingestion security at HDFS data nodes level. We also discussed the increasing struggle between data mining and data privacy protection and looked at the new methodology proposed by Meijuan et al. [Meijuan, Jian, Lihong et al. (2018)], which developed a Data Access Control mechanism for specific users based on the semantic integration nature of XML data.

We would like to extend this study by conducting a deep comparative analysis of the frameworks discussed in this paper to come up with classifications and ranking of these frameworks based on: their security and privacy achievements, their time cost efficiency, their limitations, and provide suggestions to users, such as which of these frameworks are more suitable to be used, based on the specific security and privacy users' achievements and goals.

References

Abouelmehdi, K.; Beni-Hssane, A.; Khaloufi, H.; Saadi, M. (2017): Big data security and privacy in healthcare: a review. *Proceedings of the 8th International Conference on*

Emerging Ubiquitous Systems and Pervasive Networks. Procedia Computer Science, vol. 113, no. 2, pp. 73-30.

Ashwin Kumar, T. K.; Hong, L.; Thomas, J. P.; Hou, X. (2017): Content sensitivity-based access control framework for hadoop. *Digital Communications and Networks*, vol. 3, no. 4, pp. 213-225.

Ashwin Kumar, T. K.; Liu, H.; Thomas, J. P.; Mylavarapu, G. (2015): Identifying sensitive data items within hadoop. *Proceedings of the IEEE 17th International Conference on High Performance Computing and Communications*.

Cavoukian, A. (2009). Privacy by design: the 7 foundational principles. *Proceedings of the Information and Privacy Commissioner of Ontario*.

CynergisTek (2016): Redspin: breach report 2016: Protected Health Information *(PHI)*. https://www.redspin.com/resources/download/breach-report-2016-protected-health-information-phi/.

Data IQ News (2018): Big Data to turn 'mega' as capacity will hot 44 zettabytes by 2020. https://www.dataiq.co.uk/news/20140410/big-data-turn-mega-capacity-will-hit-44-zettabytes-2020.

Edemekong, P. F.; Haydel, M. J. (2018): The Health Insurance Portability and Accountability Act (HIPPA). https://www.ncbi.nlm.nih.gov/books/NBK500019/.

Gupta, A.; Pandhi, K.; Bindu, P.; Thilagam, P. (2016): Role and access-based data segregator for security of big data. *Procedia Technology*, vol. 24, no. 7, pp. 1550-1557.

Gupta, M.; Patwa, F.; Sandhu, R. (2017). Object-tagged RBAC model for the hadoop ecosystem. *Proceedings of the 31st Annual IFIP WG 11.3 Conference*.

Gupta, M.; Patwa, F.; Sandhu, R. (2018): An attribute-based access control model for escure big data processing in hadoop ecosystem. *Proceedings of the 3rd ACM Workshop on Attribute-Based Access Control*.

Hadoop (2018): HDFS permission guide. http://archive.cloudera.com/cdh5/cdh/5/hadoop-2.2.0-cdh5.0.0-beta-1/hadoop-project-dist/hadoop-hdfs/HdfsPermissionsGuide.html.

Hu, V. C.; Grance, T.; Ferraiolo, D. F.; Kuhn, D. R. (2014): An access control scheme for big Data processing. *Proceedings of the 10th IEEE International Conference on Collaborative Computing: Networking, Applications and Work-sharing*, vol. 3, no. 9, pp. 1-7.

Huseyin, U.; Kantarcioglu, M.; Pattuk, E.; Hamlen, K. (2114): Fine-grained access control for map-reduce systems. *Proceedings of the 2014 IEEE International Congress on Big Data*.

Imine, Y.; Lounis, A.; Bouabdallh, A. (2018): Revocable attribute access control in multi-authority systems. *Network and Computer Applications*, vol. 122, no. 8, pp. 61-76.

Investopedia (2018): *Sarbanes-Oxley Act of 2002-SOX*. https://www.investopedia.com/terms/s/sarbanesoxleyact.asp

Jung, K.; Park, S. (2014): Hiding a needle in a haystack. *Proceedings of the Privacy Preserving Apriori Algorithm in MapReduce Framework*.

Meijuan, W.; Jian, W.; Lihong, G.; Lein, H. (2018): Inverted XML access control model based on ontology semantic dependency. *Computers, Materials & Continua*, vol. 55, no. 3, pp. 465-482.

MeriTalk (2018): The big data cure.

https://www.meritalk.com/study/the-big-data-cure/.

Wenrong, Z.; Yang, Y.; Luo, B. (2013): Access control for big data using data content. *Proceedings of the Big Data IEEE International Conference*.

Zhou, H.; Wen, Q. (2014): Data security accessing for HDFS based on attribute-Group in Cloud Computing. *Proceedings of the International Conference on Logistics Engineering, Management and Computer Science*.