A Novel Framework for Big Data Security Infrastructure Components

Manpreet Kaur
M.C.A Department
Saravajanik College of Engineering and Technology
Surat, India
manpreet.kaur@scet.ac.in

Alpa Shah
M.C.A Department
Saravajanik College of Engineering and Technology
Surat, India
alpa.shah@scet.ac.in

Abstract—Big data encompasses enormous data and management of huge data collected from various sources like online social media contents, log files, sensor records, surveys and online transactions. It is essential to provide new security models, concerns and efficient security designs and approaches for confronting security and privacy aspects of the same. This paper intends to provide initial analysis of the security challenges in Big Data. The paper introduces the basic concepts of Big Data and its enormous growth rate in terms of pita and zettabytes. A model framework for Big Data Infrastructure Security Components Framework (BDAF) is proposed that includes components like Security Life Cycle, Fine-grained data-centric access control policies, the Dynamic Infrastructure Trust Bootstrap Protocol (DITBP). The framework allows deploying trusted remote virtualised data processing environment and federated access control and identity management.

Keywords- *Big Data, Security challenges, Big data security life cycle, Framework*

*****

I. INTRODUCTION

Amount of data is growing rapidly every year, this is because of arrival of new technologies trend, devices, and network communication means like social networking sites. The total amount of data in the world was 4.4 zettabytes in 2013. But if we consider on daily basis, the data produce in bytes are 2.5 quintillion. This would fill 10 million blu ray discs. Big data describes the collection of complex and large data sets such that it becomes difficult to capture, process, store, search and analyze using conventional database systems. Its uses are shaping the world around us, offering more qualitative insights into our everyday lives [18]. This rate is still emergent an assortment. Despite the fact that all this information produced is having an important effect and can be helpful when processed, it is being neglected. The term Big Data is now used almost all over in our daily life. The term Big Data came around 2005 which refers to a wide range of huge data sets almost impossible to manage and process using traditional data management tools – due to their volume, but also their complexity [17].

Security is becoming more critical, as more and more companies deploy big data technologies, including Apache Hadoop, Cassandra and other related technologies. As big data technologies turn out to be major stream, it is critical that they are deployed with the same safeguards, auditing and protection capabilities inherent in existing IT platforms such as BI tools, RDBMS platforms and data storage platforms. Because of the relative new-ness of big data platforms, the security community is running rapidly to create the essential capabilities for seamless integration into existing security frameworks [19]. Big data security issues are related not only to the volume or the variety of data, but also to data quality, data privacy, and data security. Our work focuses to provide a brief insight on privacy and security aspects of Big Data. Organizations and government need to address more regulations and concerns towards analysis and storage of data. Most important barrier for spreading of new technologies is security of Big Data; required level of trust will not get achieved without adequate security guarantees. Big Data brings big responsibility [20].

Organisation of the paper

We have described our paper as follows:

- In section II we have given Big Data definition and discuss the features for security that impact it the most.
- Section III provides the level of security and privacy challenges for Big Data.
- Section IV presents a framework for Big Data infrastructure Security components and implementation.
- Section V we have given our conclusions and directions for future work.

II. BIG DATA

Big data brings big significance. Big Data is the word used to describe immense volumes of structured and unstructured data that are so huge that it is very difficult to process this data using traditional databases and software technologies. The big data refers to massive amounts of digital information collected by companies and government about people and their surroundings. Variety of users and devices generates voluminous data, and this data has to be stored and processed at powerful data centers

The main terms that signify Big Data have the following properties describe in figure 1.

A. Volume

Many factors contribute towards increasing Volume - storing transaction data, live streaming data and data collected from sensors etc. Value and potential of data can be determined only by considering the size of the data and whether this data can be considered as Big Data or not. The name “Big Data” itself describe a term which is related to volume and hence is the only characteristic required. Number of factors is involved in the data volume expansion.

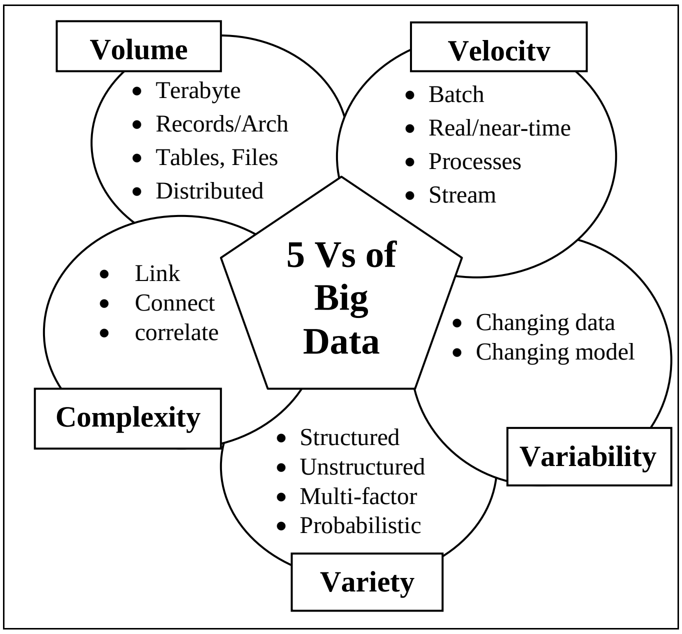

Figure 1. 5 Vs of Big Data and Security Related Properties of Veracity, Variety, and Volume.

Sources such as mobile and online transactions, social media traffic and GPS coordinates helping employees, contractors, partners and suppliers using social networking sites, intranets, extranets, corporate wikis and financial transaction-based data, increasing number of sensors and machine-to-machine data have contributed hugely in collection of Big Data.

B. *Variety*

The next characteristic of Big Data is its variety. Data analysts need to know about the category to which big data belongs. This helps the people, who are closely analyzing the data and are related with it, to effectively use the data to their benefit and thus keeping the importance of the Big Data. In number of formats data is available like- from traditional databases, text documents, emails, video, audio, and transactions.

C. *Velocity*

The term "velocity" in this perspective refers that how speedily the data is being generated or how fast the data is processed to meet the difficulty and the challenges which lie in front of growth and development. A crucial challenge for most of the organizations is to react quickly enough to deal with data velocity.

D. *Variability*

Along with the Velocity, the data flows can be highly inconsistent with periodic peaks. For analyzing the data variability can be one factor which can create problem. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively. By rising velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. It is difficult to manage daily, seasonal and event-triggered peak data loads. It is even more challenging when unstructured data gets involved [17].

E. *Complexity*

When data comes from multiple sources it becomes more complex. To convey the information by these data it is necessary to link, connect and correlate all this data. This situation, is therefore, termed as the "complexity" of Big Data. Today's data comes from multiple sources. And it is still a task to link, match, cleanse and transform data across systems. The data must be linked, matched, cleansed and transformed into required formats before actual processing [21].

Technologies today not only support the collection of large amounts of data, but also help in utilizing such data effectively [22]. One high-profile example is the Target data breach of late 2013, in which cybercriminals stole account data for millions of Target's retail shoppers. At least 40 million credit cards were compromised, according to the retail chain's estimates, and the thieves also stole personal information, including names, addresses, email addresses, and phone numbers of as many as 110 million customers.

Target's stock dropped dramatically following the breach, and the company faced numerous lawsuits, including one from a group of financial institutions claiming tens of millions of dollars in damages. In the wake of the breach, the company's CEO and CIO, who had both worked at Target for decades, resigned [23].

III. Security Challenges in Big Data

Big data security comes with unique challenges. Big Data security is basically different from traditional data security. Level of the challenges related to privacy and security get increased by big data as they are addressed in traditional security management, but also create new ones that require to be approached in a new manner. Organizations and Governments have more concern towards storing and analyzing more data, which requires more policy and regulations concerns. Big data security now becomes a big barrier for achieving security that could slow down the spread of technology. To achieve level of trust big data must ensure adequate security guarantees [20]. The differences between Big Data environments and traditional data environments include:

- The data collected, aggregated, and analyzed for big data analysis
- The infrastructure used to store and house big data
- The technologies applied to analyze structured and unstructured big data [24].

3.1 *Security and Privacy Challenges*

According to Cloud Security Alliance Big Data Security Working Group the following security and privacy challenges are identified to overcome Big Data [25].

1. Secure computations in distributed programming frameworks.
2. Security best practices for non-relational data stores.
3. Secure data storage and transactions logs.
4. End-point input validation/filtering.
5. Real-time security monitoring.
6. Scalable privacy-preserving data mining and analytics.
7. Cryptographically enforced data centric security.
8. Granular access control.
9. Granular audits.
10. Data provenance.

The above challenges are grouped into four broad components by the Cloud Security Alliance. They are describing in figure 2.

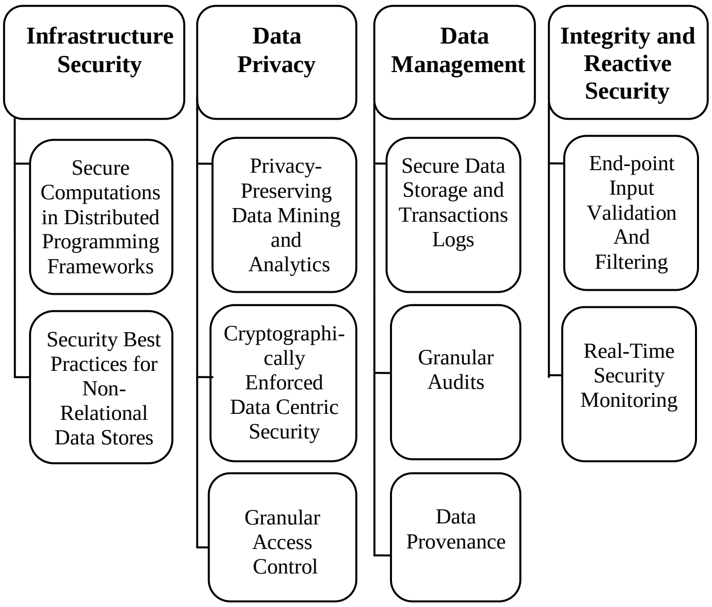
Infrastructure Security
Secure Computations in Distributed Programming Frameworks
Security Best Practices for Non-Relational Data Stores
Data Privacy
Privacy-Preserving Data Mining and Analytics
Cryptographi-cally Enforced Data Centric Security
Granular Access Control
Data Management
Secure Data Storage and Transactions Logs
Granular Audits
Data Provenance
Integrity and Reactive Security
End-point Input Validation And Filtering
Real-Time Security Monitoring

Figure 2. CSA- Classification of the Top 10 Challenges

3.2 Security Controls Approaching Data Centric Security

Organizations need to adopt a data-centric approach to security in this era of big data. Essentially, they need to employ three key types of security controls:

A. Preventive

Encryption of data at rest and in motion, redaction of data in applications, and use of identity and access management, these are some scenarios for securing the data itself with controls.

B. Detective

Looking for anomalous behavior by, for instance, auditing operating systems, Hadoop services, data activity, and monitoring systems throughout the big data environment, and providing compliance reports or alerts about potential problems.

C. Administrative

Implementing tools that enable the processes and procedures for security, such as sensitive data discovery, privileged user analysis, configuration management, and encryption key management capabilities [23].

IV. Proposed Framework for Big Data Infrastructure Security Components

Having discussed about Big Data in previous sections, we now introduce the Security Architecture Framework for Big Data. We have identified core components of the Framework and in this section we will discuss it in length. Figure 3 highlights our proposed Framework. The framework includes following components [3]:

- Security lifecycle
- Data Centric Access Control
- Trusted environment
- FADI for cooperation and services integration

4.1 Big Data Security Lifecycle

In this section, we present components of our proposed Big Data Framework. We try to address Big Data from user role viewpoint; four types of users' role are specified for Big Data environment: data provider, data collector, data miner, and decision maker. Big Data framework consists of four phases- data collection phase, data storage phase, data processing and analysis, and knowledge creation [16].

A. Data Collection Phase

In data collection phase, data comes with different formats from different sources: like structured, semi-structured, and unstructured. From the first phase of the lifecycle security of big data technology should start from a security perspective. Before gathering the data we must be ensure about the trusted sources and make sure that this phase is secured and protected. In fact, we need to take some security actions in order to keep data from being released. Some security measures can be used in this phase like restricted access control (for those who receive data from data provider) and encrypting some data fields (personal information identifier).

B. Data Storage Phase

In data storage phase, the collected data is stored and prepared for being used in the next phase (data analytics phase).it is necessary to take adequate precautions, during data storing as the collected data may contain sensitive information. In order to assurance the safety of the collected data, some security actions can be used like data anonymization approach, permutation, and data partitioning (Vertically or Horizontally).

C. Data Analysis Phase

To construct useful information data processing analysis is performed. In this phase, data mining methods such as clustering, classification, and association rule mining are used. It is essential to provide secure processing environment. In fact, data miners use powerful data mining algorithms that can extract sensitive data. Thus, a security violation may happen. Therefore, data mining process and its output must be protected against data mining based attacks.

D. Knowledge Creation Phase

Finally, decision makers can use valued knowledge from the valuable information which is extracted from data analytics phase. The created knowledge is considered as sensitive information especially in a competition environment. Further, organizations should be aware of their sensitive data (e.g. client personal data) not to be publicly released.

4.2 Data centric Access control

Consistent data centric security and access control will require solving the following problems:

- Fine-granular access control policies.
- Encryption enforced attribute based access control

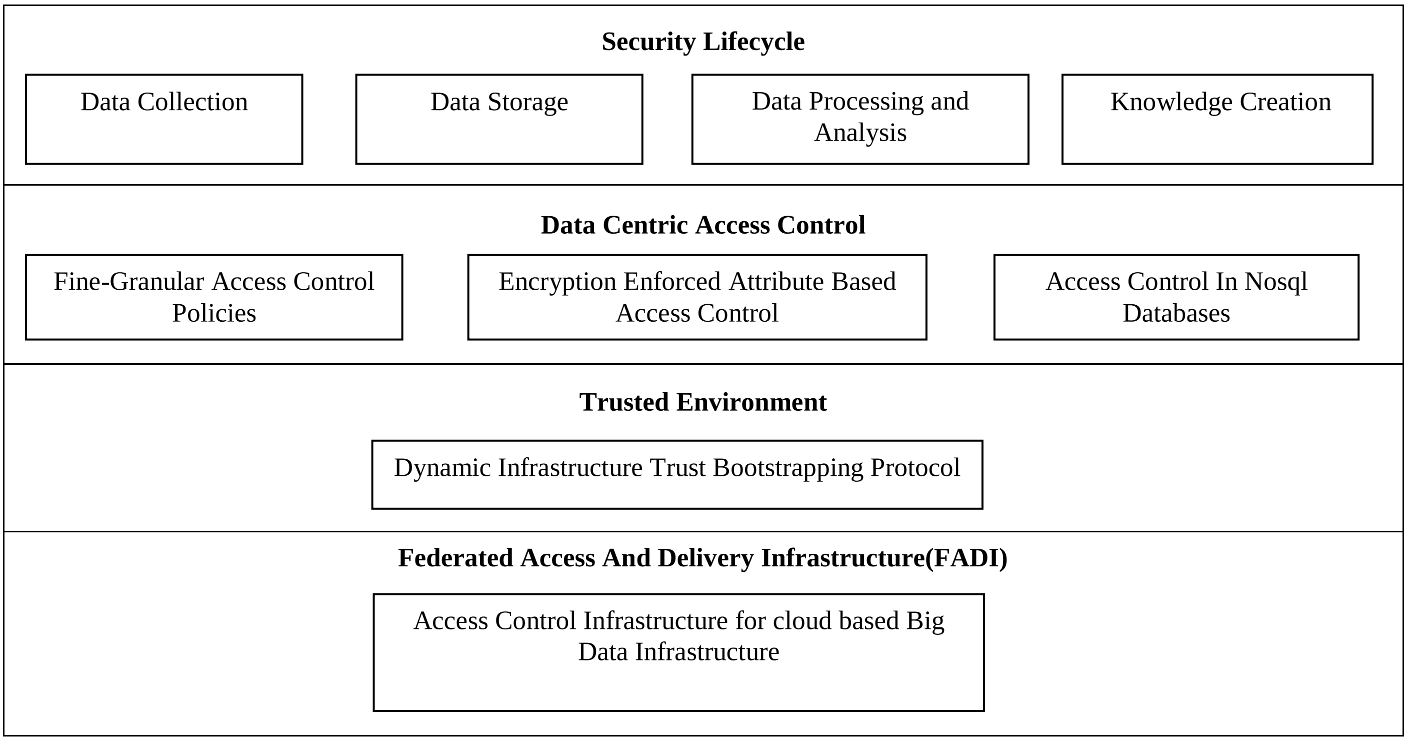
Security Lifecycle
Data Collection
Data Storage
Data Processing and Analysis
Knowledge Creation
Data Centric Access Control
Fine-Granular Access Control Policies
Encryption Enforced Attribute Based Access Control
Access Control In Nosql Databases
Trusted Environment
Dynamic Infrastructure Trust Bootstrapping Protocol
Federated Access And Delivery Infrastructure(FADI)
Access Control Infrastructure for cloud based Big Data Infrastructure

Figure 3: Framework for Big Data Infrastructure Security Components

The two basic access control and policy models can be defined based on data type and format: resource and/or document based access control, including intra document; and cell or record based access control for data stored in databases. Here we identify XACML policy language as suitable for document/intra-document access control. For databases we need to combine their native access control mechanisms and general document based access control.

A. *XACML policies for fine granular access control*

It is essential that data centric access control model provides authorization features based on structured data content and not just subjects/users, data identifiers, actions and lifetimes. The policies must be incorporated based on access control mechanisms rather than just logic expressions of attributes. Authorized users must get output based on their queries else should get error. In this respect, managing SDI/BDI (Scientific Data Infrastructure (SDI) or Big Data Infrastructure (BDI)) big data using attribute-based policy languages like XACML is applicable. Dynamically Provisioned Access Control Infrastructure (DACI) [10] uses advanced features of the XACML based policies that allow describing access control rules for complex multi-domain resources, including trust delegation, domain, multi-domain identity and session context [11, 12].

B. *Access control in NoSQL databases*

The popular NoSQL databases provide different levels of security and access control for structured data storage MongoDB [13], Cassandra, Accumulo. On user management and on protected data granularity like table-level or row-level security, most of them have coarse-grain authorization features. Accumulo [14] provides the most advanced features to allow cell-level security with which accesses from keys to values are only granted when the submitted attributes satisfy predefined Boolean expressions provided as a security label of the cell key index. However, the current policy language in Accumulo is at early development stage and lacks of features for distributed, multi-domains environments.

C. *Encryption enforced access control*

The main tool for guaranteeing confidentiality of data is data encryption. Encryption takes a piece of data, usually called the plaintext, together with a cryptographic key and produces a scrambled version of the data called the cipher text. It is possible to decrypt the data to recover the plaintext using the key, but without the key, all information about the original data is secreted by the cipher text, other than its length. This security property, commonly known as semantic security guarantees that, without the key, an adversary cannot learn any property of the underlying data even if he has a lot of imminent as to what the data may be. This is risky in that applications where data may have some predefined structure. In financial transactions where only partial information about the type of data is confronted, it is difficult to measure this data in real-world scenarios [4]. Data in-rest may remain unprotected when stored on remote facilities, so the solution to this problem can be found with using the encryption enhanced access control policies which in addition to the traditional access control, uses attributes based encryption [15].

4.3 *Trusted environment*

A generic bootstrapping protocol that can be adopted by frameworks is required that is known as Trusted Infrastructure Bootstrapping Protocol.

- ***Trusted Infrastructure Bootstrapping Protocol***

Such a protocol would have two key requirements. First remote machine must be able to authenticate and can also verify its trustworthiness by using this protocol and second, it must provide a mechanism for transferring and executing the initializing the framework. Dynamic Infrastructure Trust Bootstrapping Protocol (DITBP) includes supporting mechanisms and infrastructure that takes benefit of the TCG Reference Architecture (TCGRA) and Trusted Platform Module (TPM). One of the key components in the Trusted Computing Group Reference Architecture (TCGRA) is the Trusted Platform Module (TPM). Cryptographic functionality in hardware is provided by the physical device known as a TPM. The generation of encryption keys and the ability to store measurements of the current state are two of the key features provided by the TPM. Four key components are here for the bootstrapping process. The process in this protocol enables to begin the bootstrapping process only when a client machine authenticates a remote machine, determine that the machine is in a trusted state.

A. *Domain Authentication Server (DAS)*

It provides a trusted root for the third party's domain. It contains relevant information such as the public key for that machine's non-migratable key pair

B. *Bootstrap Initiator (BI)*

BI is the application that is transferred to the remote machine in order to confirm the machine's status before any infrastructure or software is deployed.

C. *Bootstrap Requester (BREQ)*

BREQ runs on the machine responsible for provisioning remote infrastructure and it is a client application. It communicates with its counterpart on the remote machine and handles the first/initial stage of the bootstrapping process.

D. *Bootstrap Responder (BRES)*

BRES is responsible for authenticating the machine to a remote client and verifying that the client is authorized to bootstrap the machine. Once each end point has been authenticated, the BRES will receive, decrypt and decompress the payload sent by the client [5].

4.4 FADI for cooperation and services integration

- *Access Control Infrastructure for cloud based Big Data Infrastructure*

Federated Access and Delivery Infrastructure (FADI) is defined as Layer 5 in the generic SDI (Scientific Data Infrastructure) Architecture model for e-Science (e-SDI). It includes federation infrastructure components, together with this policy and collaborative user groups support functionality. When implemented in clouds, multiple providers and both cloud and non-cloud based infrastructure components may involved in the FADI and SDI [7]. ICAF provides a common basis for building adaptive and on-demand provisioned multi-provider cloud based services. Figure 4 illustrates the general architecture and the main components of the FADI that includes infrastructure components to support inter-cloud federations services such as Cloud Service Brokers, Trust Brokers, and Federated Identity Provider. Each service/cloud domain contains an Identity Provider IDP, Authentication, Authorization, Accounting (AAA) service and typically communicates with other domains via service gateway.

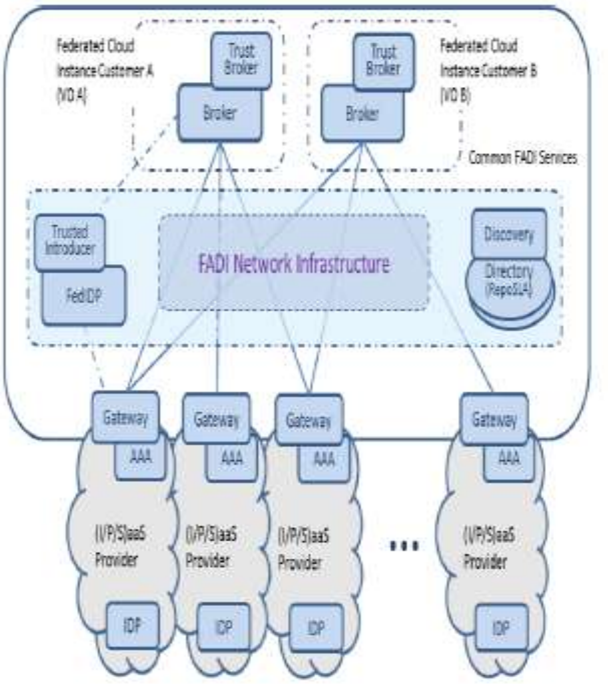

Figure 4: Federated Access and Delivery Infrastructure [9]

V. Conclusion

Big data term refers to the process of managing huge amount of data. However, the Big Data has many challenges when it comes to deploying security controls in real-world environment. Challenges like security of data storage, data mining and analytics, transaction log and secure communication do exist. In this paper we study various security challenges around Big Data security and entire stack in broad components. The paper also proposes a framework for Big Data infrastructure Security components which the reader can further expand and customize to their organizational environment. Our future work encompasses implementation of such a framework using Hadoop technology.

References

[1] Elmustafa Sayed Ali Ahmed1 and Rashid A.Saeed, "A Survey of Big Data Cloud Computing Security", International Journal of Computer Science and Software Engineering (IJCSSE), Volume 3, Issue 1, December 2014

[2] https://www.vormetric.com/data-security-solutions/use-cases/big-data-security

[3] Peter Membrey, Yuri Demchenko, Canh Ngo, Architecture Framework and Components for the Big Data Ecosystem Draft Version 0.2, 12 September 2013

[4] Ariel Hamlin, Nabil Schear, Emily Shen, Mayank Varia, Sophia Yakoubov, Arkady Yerukhimovich, Cryptography for Big Data Security, December 17, 2015

[5] Peter Membrey , Keith C. C. Chan1 , Canh Ngo , Yuri Demchenko , Cees de Laat Hong Kong Polytechnic University , University of Amsterdam, Trusted Virtual Infrastructure Bootstrapping for On Demand Services. The 7th International Conference on Availability, Reliability and Security , August 2012

[6] Vipul Goyal Omkant Pandey Amit Sahai, Brent Waters, Attribute-Based Encryption for Fine-Grained Access Control of Encrypted Data, Proceeding CCS '06 Proceedings of the 13th ACM conference on Computer and communications security

[7] Yuri Demchenko, Canh Ngo, Cees de Laat, Juan Rodriguez, Luis M. Contreras, Intercloud Architecture Framework for

Heterogeneous Cloud based Infrastructure Services Provisioning On-Demand, 27th International Conference on Advanced Information Networking and Applications Workshops, 2013.

[8] Marc X. Makkes, Canh Ngo , Yuri Demchenko , Rudolf Strijkers, Robert Meijer, Cees de Laat, Defining Intercloud Federation Framework for Multi-provider Cloud Services Integration, The Fourth International Conference on Cloud Computing, GRIDs, and Virtualization, 2013

[9] Cees de Laat , Yuri Demchenko , Canh Ngo , Peter Membrey , Daniil Gordijenko, Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure, August 2013

[10] Canh Ngo, Yuri Demchenko, Cees de Laat, Tomasz Wiktor Wlodarczyk, Wolfgang Ziegler Security Infrastructure for On-demand Provisioned Cloud Infrastructure Services, Third IEEE International Conference on Coud Computing Technology and Science, 2011

[11] Canh Ngo, Yuri Demchenko, Cees de Laat, Toward a Dynamic Trust Establishment Approach for Multi-provider Intercloud Environment, IEEE 4th International Conference on Cloud Computing Technology and Science, 2012\

[12] Yuri Demchenko, Leon Gommans, Cees de Laat, Using SAML and XACML for Complex Resource Provisioning in Grid based Applications, Eighth IEEE International Workshop on Policies for Distributed Systems and Networks, Eighth IEEE International Workshop on Policies for Distributed Systems and Networks, June 2007.

[13] Boyu Hou, Kai Qian, Lei Li, Yong Shi, Lixin Tao, Jigang Liu, MongoDB NoSQL Injection Analysis and Detection, IEEE 3rd International Conference on Cyber Security and Cloud Computing, 2016

[14] Apache Accumulo [online] http://accumulo.apache.org/

[15] Chase, M., Multi-Authority Attribute Based Encryption. Proceeding TCC'07 Proceedings of the 4th conference on Theory of cryptography.

[16] Yazan Alshboul, Yong Wang, Raj Kumar Nepali, Big Data LifeCycle: Threats and Security Model, Twenty-first Americas Conference on Information Systems, Puerto Rico, 2015

[17] Raghav Toshniwal, Kanishka Ghosh Dastidar, Asoke Nath, Big Data Security Issues and Challenges, International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 2, Volume 2, February 2015.

[18] http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/

[19] Joey Jablonski, Security for Big Data, A Dell Big Data White Paper

[20] Julio Moreno, Manuel A. Serrano, Eduardo Fernández-Medina, Main Issues in Big Data Security, www.mdpi.com/journal/futureinternet, 2016

[21] Venkata Narasimha Inukollu, Sailaja Arsi and Srinivasa Rao Ravuri, Security Issues Associated With Big Data In Cloud Computing, International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014

[22] Kalyani Shirudkar, Dilip Motwani, Big-Data Security, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 3, March 2015

[23] Securing the Big Data Life Cycle, Mit Technology Review Custom, Oracle, 2015

[24] Rossen Naydenov, Dimitra Liveri, Lionel Dupre, Eftychia Chalvatzi, Christina Skouloudi Big Data Security, Good Practices and Recommendations on the Security of Big Data Systems, European Union Agency for Network and Information Security (ENISA), 2015 December 2015

[25] Top 10 Big Data Security and Privacy Challenges, Cloud Security Alliance, 2012 https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Top_Ten_v1.pdf