ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 4, No.16 , July 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

A Survey on Security Issues in Big Data and NoSQL

Ebrahim Sahafizadeh[1], Mohammad Ali Nematbakhsh[2]

[1] Computer engineering department, University of Isfahan
Isfahan,81746-73441,Iran
sahafizadeh@eng.ui.ac.ir

[2] Computer engineering department, University of Isfahan
Isfahan,81746-73441,Iran
nematbakhsh@eng.ui.ac.ir

Abstract

This paper presents a survey on security and privacy issues in big data and NoSQL. Due to the high volume, velocity and variety of big data, security and privacy issues are different in such streaming data infrastructures with diverse data format. Therefore, traditional security models have difficulties in dealing with such large scale data. In this paper we present some security issues in big data and highlight the security and privacy challenges in big data infrastructures and NoSQL databases.

Keywords: Big Data, NoSQL, Security, Access Control

1. Introduction

The term big data refers to high volume, velocity and variety information which requires new forms of processing. Due to these properties which are referred sometimes as 3 'V's, it becomes difficult to process big data using traditional database management tools [1]. A new challenge is to develop novel techniques and systems to extensively exploit the large volume of data. Many information management architectures have been developed towards this goal [2].

As developing new technologies and increasing the use of big data in several scopes, security and privacy has been considered as a challenge in big data. There are many security and privacy issues about big data [1, 2, 3, 4, 5 and 6]. In [7] top ten security and privacy challenges in big data is highlighted. Some of these challenges are: secure computations, secure data storages, granular access control and data provenance.

In this paper we focus on researches in access control in big data and security issues on NoSQL databases. In section 2 we have an overview on big data and NoSQL technologies, in section 3 we discuss security challenges in big data and describe some access control model in big data and in section 4 we discuss security challenges in NoSQL databases.

2. Big Data and NoSQL Overview

In this section we have an overview on Big Data and NoSQL.

2.1 Big Data

Big data is a term refers to the collection of large data sets which are described by what is often referred as multi 'V'. In [8] 7 characteristics are used to describe big data:
Volume, variety, volume, value, veracity, volatility and complexity, however in [9], it doesn't point to volatility and complexity. Here we describe each property.
Volume: Volume is referred to the size of data. The size of data in big data is very large and is usually in terabytes and petabytes scale.
Velocity: Velocity referred to the speed of data producing and processing. In big data the rate of data producing and processing is very high.
Variety: Variety refers to the different types of data in big data. Big data includes structured, unstructured and semi-structured data and the data can be in different forms.
Veracity: Veracity refers to the trust of data.
Value: Value refers to the worth drives from big data.
Volatility: "Volatility refers to how long the data is going to be valid and how long it should be stored" [8].
Complexity: "A complex dynamic relationship often exists in big data. The change of one data might result in the change of more than one set of data triggering a rippling effect" [8].
Some researchers defined the important characteristics of big data are volume, velocity and variety. In general, the characteristics of big data are expressed as three Vs.

2.2 NoSQL

The term NoSQL stands for "Not only SQL" and it is used for modern scalable databases. Scaling is the ability of the system to increase throughput when the demands increase in terms of data processing. To support big data processing, the platforms incorporate scaling in two forms of scalability: horizontal scaling and vertical scaling [10].
Horizontal Scaling: in horizontal scaling the workload distributes across many servers. In this type of scalability multiple systems are added together in order to increase the throughput.

Vertical Scaling: in vertical scaling more processors, more memory and faster hardware are installed within a single server.

The main advantages of NoSQL is presented in [11] as the following: "1) reading and writing data quickly; 2) supporting mass storage; 3) easy to expand; 4) low cost". In [11] the data models that studied NoSQL systems support are classified as Key-value, Column-oriented and Document. There are many products claim to be part of the NoSQL database, such as MongoDB, CouchDB, Riak, Redis, Voldermort, Cassandera, Hypertable and HBase.

Apache Hadoop is an open source implementation of Google big table [12] for storing and processing large datasets using clusters of commodity hardware. Hadoop uses HDFS which is a distributed file system to store data across clusters. In section 6 we have an overview of Hadoop and discuss an access control architecture presented for Hadoop.

3. Security Challenges and Access Control Model

There are many security issues about big data. In [7] top ten security and privacy challenges in big data is presented. Secure computation in distributed framework is a challenge which discusses security in map-reduce functions. Secure data storage and transaction logs discuss new mechanism to prevent unauthorized access to data stores and maintain availability. Granular access control is another challenge in big data. The problem here is preventing access to data by users who should not have access. In this case, traditional access control models have difficulties in dealing with big data. Some mechanisms are proposed for handling access control in big data in [2, 3, 4, 13 and 14].

Among the security issues in big data, data protection and access control are recognized as the most important security issues in [4]. Shermin In [14] presents an access control model for NoSQL databases by the extension of traditional role based access control model. In [15] security issues in two of the most popular NoSQL databases, Cassandra and MongoDB are discussed and outlined their security features and problems. The main problems for both Cassandra and MongoDb mentioned in [15] are the lack of encryption support for data files, weak authentication between clients and servers, simple authentication, vulnerability to SQL injection and DOS attack. It is also mentioned that both of them do not support RBAC and fine-grained authorization. In [5] the authors have a look at NIST risk management standards and define the threat source, threat events and vulnerabilities. The vulnerabilities defined in [5] in term of big data are Insecure computation, End-point input validation/filtering, Granular access control, Insecure data storage and communication and Privacy preserving data mining and analytics.

In some cases in big data it is needed to have access control model based on semantical content of the data. To enforce access control in such content centric big data sharing, Content-Based Access Control (CBAC) model is presented in [2] using data content. In this case the semantic content of data plays the major role in access control decision making. "CBAC makes access control decisions based on the content similarity between user credentials and data content dynamically" [2].

Attribute relationship methodology is another method to enforce security in big data proposed in [3] and [4]. Protecting the valuable information is the main goal of this methodology. Therefore [4] focuses on attribute relevance in big data as a key element to extract the information. In [4], it is assumed that the attribute with higher relevance is more important than other attributes. [3] uses a graph to model attributes and their relationship. Attributes are expressed as node and relationship is shown by the edge between each node and the method is proposed by selecting protected attributes from this graph. The method proposed in [4] is as follow:

"First, all the attributes of the data is extracted and then generalize the properties. Next, compare the correlation between attributes and evaluate the relationship. Finally protect selected attributes that need security measures based on correlation evaluation" [4] and the method proposed in [3] is as follow:

"All attributes are represented as circularly arranged nodes. Add the edge between the nodes that have relationships. Select the protect nodes based on the number of edge. Determine the security method for protect nodes" [3].

A suitable data access control method for big data in cloud is attribute-based encryption.[1] A new schema for enabling efficient access control based on attribute encryption is proposed in [1] as a technique to ensure security of big data in the cloud. Attribute encryption is a method to allow data owners to encrypt data under access policy such that only user who has permission to access data can decrypt it. The problem with attribute-based encryption discussed in [1] is policy updating. When the data owner wants to change the policy, it is needed to transfer the data back from cloud to local and re-encrypt the data under new policy and it caused high communication overhead. The authors in [1] focus on solving this problem and propose a secure policy updating method.

Hadoop is an open source framework for storing and processing big data. It uses Hadoop Distributed File System (HDFS) to store data in multiple nodes. Hadoop does not authenticate users and there is no data encryption and privacy in Hadoop. HDFS has no strong security model and Users can directly access data stored in data nodes without any fine grain authorization [13, 16]. Authors in [16] present a survey on security of Hadoop and analyze the security problems and risks of it. Some security mechanism challenges mentions in [16] are large scale of the system, partitioning and distributing files through the cluster and executing task from different user on a single node. In [13] the authors express some of security risk in Hadoop and propose a novel access control scheme for storing data. This scheme includes Creating and Distributing Access Token, Gain Access Token and Access Blocks. The same scheme is also used with Secure Sharing Storage in cloud. It can help the data owners control and audit access to

their data but the owners need to update the access token when the metadata of file blocks changes.

4. Secutity Issues in NoSQL Databases

NoSQL stands for "Not Only SQL" and NoSQL databases are not meant to replace the traditional databases, but they are suitable to adopt big data when the traditional databases do not appropriate [17]. NoSQL databases are classified as Key-value database, Column-oriented database, Document based and Graph database.

4.1 MongoDB

MobgoDB is a document based database. It manages collection of documents. MongoDB support complex datatype and has high speed access to huge data.[11] flexibility, power, speed and ease of use are four properties mentioned in [18] for MongoDB. All data in MongoDB is stored as plain text and there is no encryption mechanism to encrypt data files [19].
All data in MongoDB is stored as plain text and there is no encryption mechanism to encrypt data files. [19] This means that any malicious user with access to the file system can extract the information from the files. It uses SSL with X.509 certificates for secure communication between user and MongoDB cluster and intra-cluster authentication [17] but it does not support authentication and authorization when running in Sharded mode [15]. The passwords are encrypted by MD5 hash algorithm and MD5 algorithm is not a very secure algorithm. Since mongo uses Javascript as an internal scripting language, authors in [15] show that MongoDb is potential for scripting injection attack.

4.2 CouchDB

CouchDb is a flexible, fault-tolerant document based NoSQL database [11]. It is an open source apache project and it runs on Hadoop Distributed File Systems (HDFS) [19].
CouchDB does not support data encryption [19], but it supports authentication based on both password and cookie [17]. Passwords are encrypted using PBKDF2 hash algorithm and are sent over the network using SSL protocol [17]. CouchDB is potential for script injection and denial of service attack [19].

4.3 Cassandra

Cassandra is an open source distributed storage for managing big data. It is a key value NoSQL database which is used in Facebook. The properties mentioned in [11] for Cassandra are the flexibility of the schema, supporting range query and high scalability.
all passwords in Cassandra are encrypted by the use of MD5 hash function and passwords are very weak. If any malicious user can bypass client authorization, user can extract the data because there is no authorization mechanism in inter-node message exchange.[17] Cassandra is potential for denial of service attack because it performs one thread per one client [19] and it does not support inline auditing.[15] Cassandra uses a query language called Cassandra Query Language (CQL), which is something like SQL. The authors of [15] show that injection attack is possible on Cassandra like SQL injection using CQL. Cassandra also has problem in managing inactive connection [19].

4.4 HBase

HBase is an open source column oriented database modeled after Google big table and implemented in java. Hbase can manage structured and semi-structured data and it uses distributed configuration and write ahead logging.
Hbase relies on SSH for inter-node communication. It supports user authentication by the use of SASL (Simple Authentication and Security Layer) with Kerberos. It also supports authorization by ACL (Access Control List) [17].

4.5 HyperTable

Hypertable is an open source high performance column oriented database that can be deployed on HDFS. It is modeled after Google's big table. It use a table to store data as a big table [20].
Hypertable does not support data encryption and authentication [19]. It does not tolerate the failure of range server and if a range server crashes it is not able to recover lost data [20]. Eventhough Hypertbale uses Hypertable Query Language (HQL) which is similar to SQL, but it has no vulnerabilities for the injection [19]. Additionally there is no denial of service is reported for Hypertable [19].

4.6 Voldemort

Voldemort [23] is a key value NoSQL database used in LinkedIn. This type of databases match keys with values and the data is stored as a pair of key and value. Voldemort supports data encryption if it uses BerkeleyDB as the storage engine. There is no authentication and authorization mechanism in Voldemort. It neither supports auditing [21].

4.7 Redis

Redis is an open source key value database. Data encryption is not supported by Redis and all data stored as plain text and the communication between Redis client and server is not encrypted [19]. Redis does not implement access control, so it provides a tiny layer of authentication. Injection is impossible in Redis, since Redis protocol does not support string escaping concept [22].

4.8 DynamoDB

DynamoDB is a fast and flexible NoSQL database used in amazon. It supports both key value and document data model [24]. Data encryption is not supported in Dynamo but the communication between client and server uses https protocol.

Authentication and authorization is supported by dynamo and arequests need to be signed using HMAC-SHA256 [21].

4.9 Neo4J

Neo4j [25] is an open source graph database. Neo4j does not support data encryption and authorization and auditing. The communication between client and server is based on SSL protocol. [21].

5. Conclusion

Increasing the use of NoSQL in organization, security has become a growing concern. In this paper we presented a survey on security and privacy issues in big data and NoSQL. We had an overview on big data and NoSQL databases and discussed security challenges in this area. Due to the high volume, velocity and variety of big data, traditional security models have difficulties in dealing with such large scale data. Some researchers presented new access control model for big data which was introduced in this paper.

In the last section we described security issues in NoSQL databases. As it was mentioned the most of NoSQL databases has the lack of data encryption. To have a more secure database it is needed to encrypt sensitive database fields. Some of databases have vulnerability for injection. It is needed to use sufficient input validation to overcome this vulnerability. Some of them have no authentication mechanism and some of them have weak authentication mechanism. So to overcome this weakness it is needed to have strong authentication mechanism. CouchDB uses SSL protocol, Hbase uses SASL and Hypertable, redis and Voldemort has no authentication and the other databases has weak authentication. MongoDB and CouchDB are potential for injection and Cassandra and CouchDB are potential for denial of service attack. Table 1 briefly shows this comparison.

Table 1: The Comparison between NoSQL Databases

DB/Criteria	Data Model	Authentication	Authorization	Data Encription	Auditing	Communication protochol	Potential for attack	Data Model
MongoDb	Document	Not Support	Not Support	Not Support	-	SSL	Script injection	Document
CouchDB	Document	Support	-	Not Support	-	SSL	Script injection and DOS	Document
Cassandra	Key/Value	Support	Not Support	Not Support	Not Support	SSL	Script injection (in CQL) and DOS	Key/Value
Hbase	Column Oriented	Support	Support	Not Support	-	SSH	Not reoprt for DOS and injection	Column Oriented
HyperTable	Column Oriented	Not Support	-	Not Support	-	-	-	Column Oriented
Voldemolt	Key/Value	Not Support	Not Support	Support	Not Support		-	Key/Value
Redis	Key/Value	Tiny Layer	Not Support	Not Support	Not Support	Not Encrypted	-	Key/Value
DynamoDB	Key/Value Document	Support	-	Not Support	-	https	-	Key/Value Document
Neo4J	Graph	-	Not Support	Not Support	Not Support	SSL	-	Graph

References

[1] K.Yang, Secure and Verifiable Policy Update Outsourcing for Big Data Access Control in the Cloud, Parallel and Distributed Systems, IEEE Transactions on , Issue 99, 2014

[2] W.Zeng, Y.Yang, B.Lou, Access control for big data using data content, Big Data, IEEE International Conference on, pp. 45-47, 2013

[3] S.Kim, J.Eom, T.Chung, Big Data Security Hardening Methodology Using Attributes Relationship, Information Science and Applications (ICISA), 2013 International Conference on, pp 1-2, 2013

[4] S.Kim, J.Eom, T.Chung, Attribute Relationship Evaluation Methodology for Big Data Security, IT Convergence and Security (ICITCS), 2013 International Conference on, pp 1-4, 2013

[5] M.Paryasto, A.Alamsyah, B.Rahardjo, Kuspriyanto, Big-data security management issues, Information and Communication Technology (ICoICT), 2nd International Conference on, pp 59-63, 2014

[6] J.H.Abawajy,A. Kelarev, M.Chowdhury, Large Iterative Multitier Ensemble Classifiers for Security of Big Data, Emerging Topics in Computing, IEEE Transactions on, Volume 2, Issue 3, pp 352-363, 2014

[7] Cloude Security Allience, Top Ten Big Data Security and Privacy Challenges, www.cloudsecurityalliance.org, 2012

[8] K. Zvarevashe, M. Mutandavari, T. Gotora, A Survey of the Security Use Cases in Big Data, International Journal of

Innovative Research in Computer and Communication Engineering, Volume 2, issue 5, pp 4259-4266, 2014

[9] M.D.Assuncau, R.N.Calheiros, S.Bianchi, A.S.Netto, R.Buyya, Big Data computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing, 2014

[10] D.Singh, C.K.Reddy, A survey on platforms for big data analytics, Journa of Big Data, 2014

[11] J.Han, E.Haihong, G.Le, J.Du , Survey on NoSQL Database, Pervasive Computing and Applications (ICPCA), 2011 6th International Conference on, pp 363-366, 2011

[12] F.Chang, J.Dean, S.Ghemawat, W.C. Hsieh, D.A. Wallach, Bigtable: A Distributed Storage System for Structured Data, Google, 2006

[13] C.Rong, Z.Quan, A.Chakravorty, On Access Control Schemes for Hadoop Data Storage, International Conference on Cloud Computing and Big Data, pp 641-645, 2013

[14] M. Shermin, An Access Control Model for NoSQL Databases, The University of Western Ontario, M.Sc thesis, 2013

[15] L.Okman, N.Gal-Oz, Y.Gonen, E.Gudes, J.Abramov, Security Issues in NoSQL Databases, Trust, Security and Privacy in Computing and Communications (TrustCom), IEEE 10th International Conference on, pp 541-547, 2011

[16] M.RezaeiJam, L.Mohammad Khanli, M.K.Akbari, M.Sargolzaei Javan, A Survey on Security of Hadoop, Computer and Knowledge Engineering (ICCKE), 2014 4th International Conference on , pp 716-721, 2014

[17] A.Zahid, R.Masood, M.A.Shibli, Security of Sharded NoSQL Databases: A Comparative Analysis, Conference on Information Assurance and Cyber Security (CIACS), pp 1-8, 2014

[18] A.Boicea, F.Radulescu, L.I.Agapin, MongoDB vs Oracle - database comparison, Emerging Intelligent Data and Web Technologies (EIDWT), 2012 Third International Conference on, pp 330 – 335, 2012

[19] 19P.Noiumkar, T.Chomsiri, A Comparison the Level of Security on Top 5 Open Source NoSQL Databases, The 9th International Conference on Information Technology and Applications(ICITA2014) , 2014

[20] A.Khetrapal, V.Ganesh, HBase and Hypertable for large scale distributed storage systems, A Performance evaluation for Open Source BigTable Implementations, Dept. of Computer Science, Purdue University, http://cloud.pubs.dbs.uni-leipzig.de/node/46, 2008

[21] K.Grolinger, W.A.Higashino, A.Tiwari,M.AM Capretz, Data management in cloud environments: NoSQL and NewSQL data stores, Journal of Cloud Computing: Advances, Systems and Applications, 2013

[22] http://redis.io/topics/security

[23] http://www.project-voldemort.com

[24] http://aws.amazon.com/dynamodb

[25] http://neo4j.com

Ebrahim Sahafizadeh, B.S. Computer Engineering (Software), Kharazmi University of Tehran,2001, M.S. Computer Engineering (Software), ran University of Science & Technology, Tehran, 2004. Ph.D student at Isfahan University. Faculty member, Lecturer, Department of Information Technology, Payame Noor University , Boushehr.

MohammadAli Nematbakhsh, B.S. Electrical Engineering , Louisiana Tech University, USA, 1981, M.S. Electrical and Computer Engineering. University of Arizona, USA, 1983, Ph.D. electrical and Computer Engineering, University of Arizona, USA, 1987. Micro Advanced Computer, Phoenix, AZ, 1982-1984, Toshiba Co, USA and Japan, 1988-1993, Computer engineering Department, university of Isfahan, 1993-now