A Comparative Survey on Security Issues in Big Data and NoSQL

Manish Bhardwaj[1]

[1]*Assistant Professor in Poornima Group of Institutions, Jaipur*

[1]*manishbhardwaj@poornima.org*

Abstract

The world of today is producing a large amount of data points to give us insights in the field of research, decision making system and value additions to buisness. To acquire these results organizations are required to be good enough to handle it efficiently and quickly. Its data everywhere and when data is there we can not ignore security or say encryption specifically. Today, we are walking on a two fold edged sword. On one hand, Big Data takes into account rich, effective examination with solid inquiries and close inconceivable outcomes. On alternate, programmers are tenaciously endeavoring to break through your organization's cloud security obstructions and harvest all that you've sewn.This paper is to overview on security and assurance issues in Big Data and NoSQL. On account of the high volume, speed and collection of huge data, security and insurance issues are particular in such spouting data establishments with various data orchestrate. In like manner, standard security models encounter issues in overseeing such generous scale data. In this paper we show some security issues in tremendous data and highlight the security and insurance challenges in gigantic data structures and NoSQL databases.

Keywords: *Big Data, NoSQL, Security, Access Control*

1. Introduction

The term Big Data is related with 3 V’s. Those three V’s are volume, velocity and vaiety of information which requires a different type of handling. Earlier database management systems are not capable enough to process the big data available currently. Every one of us is uploading thousands of photographs, videos, spreadsheets etc. Our traditional ways are not capable enough to process the data.As data is increasing day by day the issue of security of data is also increased.

As growing with new technologies and increasing the utilization of big data in many areas has also raised security and protection issues. There are numerous security and protection issues about big data. Some of these difficulties are: secured computations, data storages, granular access control and data provenance. In today’s scenario organizations are paying a lot of charges to take a data centric approach to security.

2. Big Data and NoSQL Overview

2.1 BigData

Big Data is a collection of data sets which is large in size and can be understood easily using 7 V’s. Those 7 V‘s are utilized to portray the term Big Data. These 7 V’s portray the various attributes of Big Data and are as follows:

Volume: Volume is alluded to the measure of information. The extent of information in Big Data is huge and is normally in terabytes and petabytes scale.

Velocity: Velocity alluded to the speed of information creating and preparing. In Big Data the rate of information creating and preparing is high.

Variety: Variety alludes to the distinctive sorts of information in Big Data. Big Data

incorporates semi-structured, structured and unstructured data and it can be in various forms.
Veracity: Veracity alludes to the trust of data.
Value: Value alludes to the value drives from Big Data.
Volatility: "Volatility alludes to what extent the data will be substantial and to what extent it ought to be put away".
Complexity: "A mind boggling dynamic relationship regularly exists in huge information. The difference in one information may bring about the difference in excess of one arrangement of information setting off an undulating impact".
The essential qualities of Big Data are volume, velocity and variety. When all is said in done, the qualities of huge information are communicated as three Vs.

2.2 NoSQL

NoSQL stands for "Not just SQL" and it is utilized for scalable databases. Scaling is the capacity of the framework to build throughput when the requests increment as far as information handling. To help big Data processing, two types of scaling is done: vertical and horizontal.
Horizontal Scaling: it distributes the workload across numerous servers.To maximize the output multiple systems are added altogether.
Vertical Scaling: In this approach a single system is made stronger by installing more processor, more memory and faster hardware. Good thing about NoSQL are 1)rapid reading and writing of data; 2)enabling mass storage; 3) scalability; 4)low cost".The information models that contemplated NoSQL frameworks bolster are named Key-value, Column-arranged and Document. There are numerous products that are counted as NoSQL database likeCouchDB,MongoDB, Redis, Voldermort, Riak, Cassandera, Hypertable and HBase.

Google big table has implemented Apache Hadoop as an open source for large datasets for storing and processing using clusters of commodity hardware. In distributed programming framework HDFS (hadoop distributed file system) is used for processing of huge data amount. When it comes to distributed scenario security is always a crucial issue.

3. Security Challenges

When it comes to Big Data data giants can not ignore security as a one-off project. It needs continous working towards data security. Preventive, detective and administrative key types of security controls are to be taken care of. In preventive key type security is done against mistakes or cybercriminals to access the data. In case someone gets succeed to access the data than it should become useless. This is done through encryption decryption, masking and privilage management.In detective key type security is provided by auditing database and system for a long time and creating alerts about problems found. In administrative type of key different tools ae used to enable the process and procedures for security purpose like data recovery, privilege, configuration and key managements. It is easy to secure the data at all the times because threats are becoming more prevalent and powerful day by day. To ensure that data is insulated from the threats four ways are used for most possible threats. Cloud security is built in many organizations to deploy database by implementing completely or partially. In cloud security it is ensured that data replication is being done on timely bases which will work as fail-safe in any disaster.A cloud solution should be safe, replication enabled and should give feel that they are running with in their own data center. Second way is to use built-in encryption in case of sensitive information is lost or stolen. Data is encrypted automatically and decryption is performed using by authorized user only. In third way we use central master key management which uses a master encryption key. This approach is made effective by searching a database that can work with centralized master key management and used by organization's servers. Fourth way is to use hardware encryption acceleration. Database performance can be boosted up while doing encryption and decryption. It may take advantage of any acceleration technology. Securing the big data requires authentication, authorization of database, application and users, privilage management, encryption of data, data redaction, separation of roles, API security , monitoring, auditing , alerting and reporting.

4. Secutity Issues in NoSQL Databases

NoSQL means Not Only SQL. These databases are more suitable to adopt big data and categorized as key value, column oriented, document based and graph based databases.

4.1 MongoDB :

It is a record based database which handles accumulation of documents. It supports non traiditional datatype and has rapid access to Big Data.It is fast, powerful, flexible and easy to use. All data in MongoDb is kept as plain content and there is no encryption system to scramble record files.which provide access to the document to any malicious user.SSL with X.509 is used to provide secure communication to user and MongoDb cluster. It does not provide authentication in sharded mode. MD5 hash algorithm is used for encryption of passwords. Javascript is used used as internal scripting so it is potential for scripting injection attack

.

4.2 CouchDB :

It is adaptable , fault-tolerant document based open source apache project and it runs on Hadoop Distributed File Systems (HDFS). Data encryption is not supported but authentication is done through password and cookies. It is done using PBKDF2 hash algorithm and SSL protocol. It is potential for script injection and Denial of service attacks.

4.3 Cassandra:

It is an open source distribution of Big data.It is used in facebook as key value NoSQL database.MD5 hash function is used for encryption. User can easily extract the data due to no authrization technique in inter-node exchange. It is open for denial of service attacks. It uses Cassandra query language (CQL) which makes it open for SQL injection.

4.4 HBase :

It is an open source column based database inspired by Google big table. It can handle structured and semi-structured data organized and semi-organized data. It uses write ahead logging with distributed configuration. It works on SSH inter-node communication and supports user authentication using SASL with Kerberos along with Access Control list authorization.

4.5 HyperTable :

An highly efficient column oriented database which is deployed on HDFS.It store the data in a big table.It does not do data encryption and authentication. If a server crashing happens data can not be recovered. HQL (hyper query language) is used for hypertable which opens it for injection attacks. There is no DOS attack.

4.6 Voldemort :

It is used in LinkedIn. It is key value NoSQL database. It mathes keys and values and data is stored in key value pair. It supports data encryption using BerkeleyDB as storage engine.It does not supports auditing, authentication and authorization

4.7 Redis:

This is an open source key value database which does not support data encryption. All the data is stored as text and even communication between redis client and server is not encrypted. There is no access control implementation which applies a tiny layer of authentication.Injection attack is impossible in Redis.

4.8 DynamoDB:

It is a quick and adaptable NoSQL which is used in amazon. It is a key value and document data model. Their is no data encryption but https is used in communication in between client and server. Authentication and authorization are there.

4.9 Neo4J:

It is open source graphical database which does not support data encryption, authorization and auditing. SSL is used for client server communication.

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 andDOS	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

5. Conclusion

Security is very important concern while working with Big Data and NoSQL.This paper discusses survey on security and privacy issues in big data and NoSQL.Due to 3 v's volume, valocity and variety of big data,large scale data is too tough to be handled.In NoSQL databases absence of information encryption makes it vulnerable to security threats. Different NoSQL databases has taken different ways to secure themselves usis SSL, SASL etc.CouchDB utilizes SSL, Hbase utilizes SASL and Hypertable, redis and Voldemort has no verification and alternate databases has frail validation. MongoDB and CouchDB are potential for infusion and Cassandra and CouchDB are potential for refusal of administration assault. Table 1 quickly demonstrates this examination.

References

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

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

[3.] S.Kim, J.Eom, T.Chung, Big Data Security HardeningMethodology Using Attributes Relationship,InformationScience and Applications (ICISA), 2013 InternationalConference on, pp 1-2, 2013

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

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

[6.] J.H.Abawajy,A. Kelarev,M.Chowdhury, Large IterativeMultitier Ensemble Classifiers for Security of Big Data,Emerging Topics in Computing, IEEE Transactions on,Volume2, Issue 3, pp 352-363, 2014

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

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

[9.] .Singh, C.K.Reddy, A survey on platforms for big dataanalytics, Journa of Big Data, J.Han, E.Haihong, G.Le, J.Du, Survey on NoSQL Database,Pervasive Computing and Applications (ICPCA), 2011 6thInternational Conference on, pp 363-366, 2011

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

[11.] C.Rong, Z.Quan, A.Chakravorty, On Access Control Schemesfor Hadoop Data Storage, International Conference on CloudComputing and Big Data, pp 641-645, 2013

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