International e-Conference on New Horizons And Multidisciplinary Applications In Science And Technology
In Association withInternational Journal of Scientific Research in Science and Technology
Volume 9 | Issue 6 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)

Big Data Security : An Overview

V.S.Tondre*[1], V.V.Thakare[1]

*[1]Department of Computer Science, Sant Gadge Baba Amravati University, Amravati, Brijlal Biyani Science College, Amravati, Maharashtra, India

ABSTRACT

Data is one of the most important resources in every field. The continuous growth in the importance and volume of data has created a new problem. It cannot be handled by traditional analysis techniques. This problem was, therefore, solved through the creation of a new paradigm: Big Data. However, Big Data originated new issues related not only to the volume or the variety of the data, but also to data security and privacy. In order to obtain a full perspective of the problem, we decided to carry out an investigation with the objective of highlighting the main issues regarding Big Data security, and also the solutions proposed by the scientific community to solve them. In this paper, we explain the results obtained after applying a systematic mapping study to security in the Big Data ecosystem. It is almost impossible to carry out detailed research into the entire topic of security in a Big Data system, along with the principal solutions to them proposed by the research community.

Keywords: Big Data; Security; Systematic Mapping study

I. INTRODUCTION

Datahasbecomeoneofthemost importantassetsinalmosteveryfield.Not only are they important for companies related to the computerscienceindustry,butalsofororganizations,such ascountries'healthcare, governments,education,orthe engineering sector. Data are essential with respect to carrying out their daily activities, and also helping the businesses'managementtoachievetheirgoalsandmake thebestdecisionsonthebasisoftheinformationextracted from them [1]. It is estimated that of all the data in recorded human history, 90 percent hasbeen created in the last few years. In 2003, five exabytes of data were createdbyhumans,andthisamountofinformationis,at present,createdwithintwodays.Thistendencytowards increasing the volume and detail of the data that is collectedbycompanieswillnotchangeinthenearfuture, asthe riseofsocialnetworks,multimedia,andtheInternet of Things (IOT) is producing an overwhelming flow of data [2].WearelivingintheeraofBigData.Furthermore, thisdataismostlyunstructured,signifyingthose traditional systemsarenotcapableofanalyzingit.Organizationsare willing to extract more beneficial information from this highvolumeandvarietyofdata.Anewanalysisparadigm with
which to analyze and better understand this data, therefore,emergedinordertoobtainnotonlyprivate,but alsopublic,benefits,andthiswasBigData [3].

Eachnewdisruptivetechnologybringsnewissueswithit. InthecaseofBigData,theseissuesarerelatednotonlyto thevolumeorthevarietyofdata,butalsotodataquality, data privacy, and data security. This paper will focuson thesubjectsofBigDataprivacyandsecurity.BigDatanot onlyincreasesthescaleofthechallengesrelatedtoprivacy andsecurityastheyareaddressedintraditionalsecurity management, but also create new ones that need to be approached in a new way [4]. As more data is storedand analyzed by organizations or governments, more regulations are needed to address these concerns. AchievingsecurityinBigDatahas,therefore,becomeone ofthemostimportantbarriersthatcouldslowdownthe spread of technology; without adequate security guarantees,BigDatawillnotachievetherequiredlevelof trust.BigDatabringsbigresponsibility.

According to the Big Data Working Group at the Cloud SecurityAllianceorganizationthereare,principally,four different aspects of Big Data security: infrastructure security,dataprivacy,datamanagement,andintegrityand reactive security. This division of Big Data security into four principal topics has also been used by the InternationalOrganizationforStandardizationinorderto createasecuritystandardforsecurityinBigData.

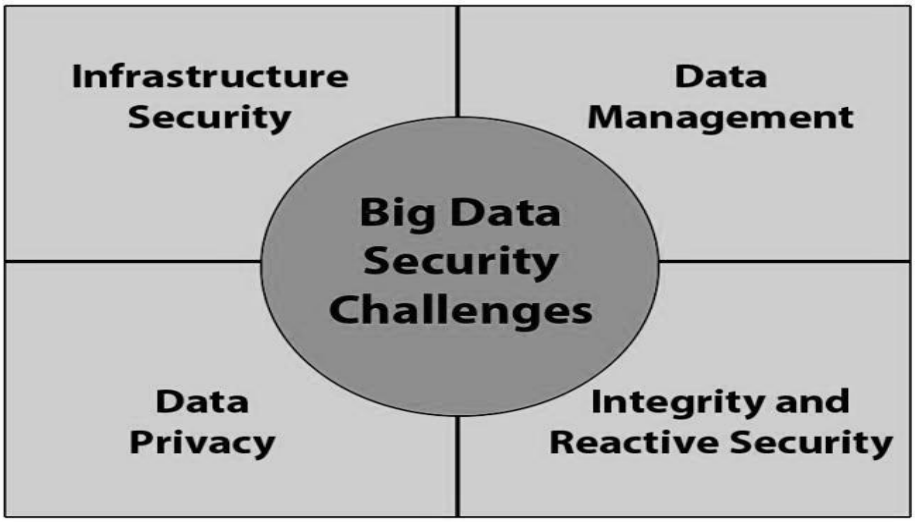

Figure1. Mainchallengesof securityinbigdata

II. INFRASTRUCTURESECURITY

When discussing infrastructure security, it is necessary to highlight the main technologies and frameworksfoundasregardssecuringthearchitectureof a Big Data system, and particularly those based on the Hadooptechnology,sinceitisthatmostfrequentlyused.In thissectionweshallalsodiscusscertainothertopics,such ascommunicationsecurityinBigData,orhowtoachieve high-availability.

A. Security forHadoop

ThegraphicshowsthatthemaintopicdealtwithbythatresearchinginfrastructuresecurityissecurityforHadoop. Hadoop can be consideredasadefectstandardforimplementingaBig Data environment in a company. The security problems related to this technology have been widely discussedbyresearchers,whohavealsoproposedvarious methodswithwhichtoimprovethesecurityoftheHadoop system. This category is probably the most transverse since, in order to protect it, the solutions use different securitymechanismssuchasauthenticityorcryptography.

Forexample,thereisaproposalforasecuritymodelfor G-Hadoop(anextensionoftheMap Reduceframeworkto run on multiple clusters) that simplifies users' authenticationandsomesecuritymechanismsinorderto protect the system from traditional attacks [5]. A few papers focus on protecting the data that is stored inthe HDFSbyproposinganewschema,asecureaccesssystem, oreventhecreationofanencryptionscheme.

B. Availability

Researchers have also dealt with the subject of availability in Big Data systems. One of the main characteristicsofBigDataenvironments,andbyextension ofaHadoopimplementation,istheavailabilityattainedby theuseofhundredsofcomputersinwhichthedataarenot only stored, but are also replicated along the cluster. Findinganarchitecturethatwillensurethefullavailability ofthesystemis,therefore,apriority.

For instance [6], in the authors propose a solution with whichtoachievehighavailabilitybyhavingmultipleactive NameNodesatthesametime.Othersolutionsarebased oncreatinganewinfrastructureofthestoragesystemso astoimproveavailabilityandfaulttolerance.

C. ArchitectureSecurity

Anotherdifferentapproachisthatofdescribinga newBigDataarchitecture,ormodifyingthetypicalone,in

Order to improve the security of the environment. The authors of propose a new architecture based on the Hadoopfilesystemwhich,whencombinedwithnetwork coding and multi-node reading, makes it possible to improve the security of the system. Another solution focuses on secure group communications in large-scale networks managed by Big Data systems, and this is achieved by creating certain protocols and changing the infrastructureofthenodes.

D. Authentication

ThevalueofthedataobtainedafterexecutingaBig Dataprocesscan,toagreatextent,bedeterminedbyits authenticity. A few papers deal with this problem by proposing solutions related to authentication. In, the authorssuggestsolvingtheproblemofauthenticationby creating an identity-based signcryption scheme for Big Data.

E. CommunicationSecurity

Thesecurityasregardscommunicationsbetween different parts of the Big Data ecosystem is a topic that often is ignored, and only a small number of papers thereforedealwiththisproblem.Onepaperapproaches thetopicbyexplainingtheregulardatalifecycleinaBig Datasystem,followingthedifferentnetworkprotocolsand applicationsthatthedatapassthrough.Theauthorsalso enumeratethemaindatatransfersecuritytechniques.

III. DATAPRIVACY

Data privacy is probably the topic about which ordinarypeoplearemostconcerned,butitshouldalsobe oneofthegreatestconcernsfortheorganizationsthatuse BigDatatechniques.ABigDatasystemusuallycontainsan enormous amount of personal information that organizations use in order to obtain a benefit from that data. However, we should ask ourselves where the limit regardingtheuseofthatinformationis.

Organizationsshouldnothavetotalfreedomtousethat information without our knowledge, although they also needtogainsomebenefitfromtheuseofthatdata.Several techniques and mechanisms with which to protect the privacyofthedata,andalsoallowcompaniestostillmake a profit from it have therefore been developed, and attempttosolvethisprobleminvariousdifferentways.

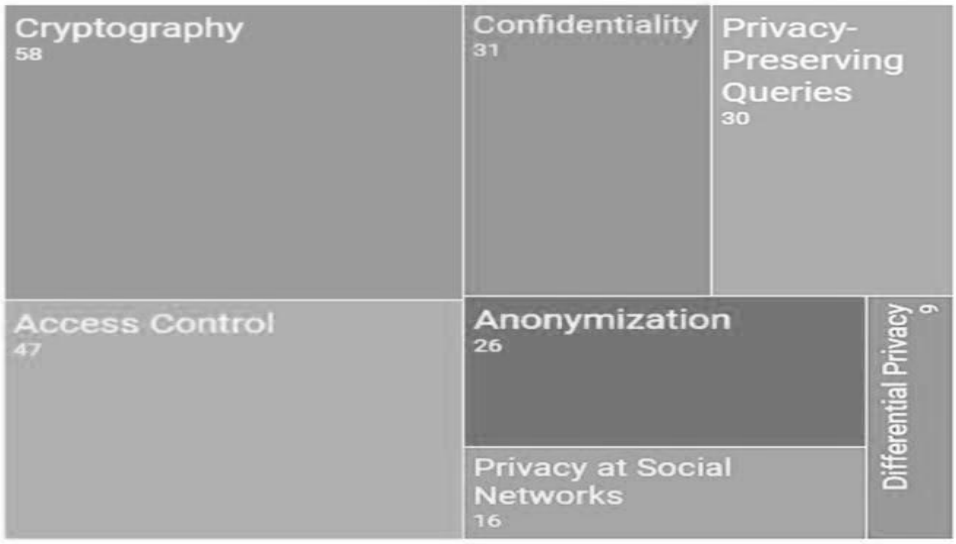

Figure 2 Main topics on data privacy.

A. Cryptography

The most frequently employed solution as regards securingdataprivacyinaBigDatasystemiscryptography. Cryptography has been used to protect data for a considerableamountoftime.Thistendencycontinuesin the case of Big Data, but it has a few inherent characteristics that make the direct application of traditionalcryptographytechniquesimpossible.

Oneexampleoftheuseofcryptographycanbefound in, in which the authors propose a bitmap encryption scheme that guarantees users' privacy. Other authors' researchisfocusedonhowtoprocessdatathatisalready encrypted.Onepaper,forexample,explainsatechnique with which to analyze and program transformations withPig Latininthecaseofencrypteddata.

B. AccessControl

Access control is one of the basic traditional techniques used to achieve the security of a system.- Its main objective is to restrict non-desirable users' access to the system. In the case of BigData, the access control problem is related to the fact that there are only basic forms of access control. In order to solve this problem, some authors propose a frameworkthatsupportstheintegrationofaccesscontrol features. Other researchers focus their attention on the Map Reduceprocessitself,andsuggestaframeworkwith which to enforce the security policies at the key-value level.

C. Confidentiality

Although privacy is traditionally treated as a part of confidentiality,wedecidedtochangetheorderowingto the tremendous impact that privacy has on the general public'sperceptionofBigDatatechnology.

Theauthorsthatapproachthisproblemoftenpropose newtechniquessuchascomputingonmaskeddata(CMD), which improves data confidentiality and integrity by allowingdirectcomputationstobemadeonmaskeddata, or new

schemes, such as Trusted Scheme for Hadoop Cluster (TSHC) which creates a new architecture framework for Hadoop in order to improve the confidentialityandsecurityofthedata.

D. Privacy-Preserving Queries

The main purpose of a Big Data system is to analyzethe data in order to obtain valuable information. However, while we manipulate that data, we should not forget its privacy.Afewpaperspayattentiontotheproblemofhow to make queries whilst simultaneously not violating the privacyofthedata.

One way in which to achieve this protection is by encryptingthedata,asdiscussedpreviously,butthisadds a new problem: how do we analyze the encrypted data? Someauthorsproposethatthisproblemcanbesolvedby means of a secure keyword search mechanism over that encrypteddata.

E. Anonymization

One of the most extended ways in which toprotect the privacy of data is by anonymizing it. This consists of applyingsomekindoftechniqueormechanismtothedata inordertoremovethesensitiveinformationfromitorto hideit.BigDatausuallyimpliesalargeamountofdata,and this problem, therefore, increases in Big Data environments.

Theauthorsofproposeahybridmethodthatcombines the two most frequently used anonymisation schemes: top-down specialization (TDS) and bottom-up generalization(BUG).

F. DifferentialPrivacy

The objective of differential privacy is to provide a methodwithwhichtomaximizethevalueofanalysisofa set of data while minimizing the chances of identifying users'identities.Afewpapersfocusonachievingprivacy inBigDatabyapplyingdifferentialprivacytechniques.For example, in the authors attempt to distort the data by addingnoise.

IV. DATA MANAGEMENT

This section focuses on what to do once the data is containedintheBigDataenvironment.Itnotonlyshows how to secure the data that is stored in the BigData system, but also how to share that data. We shall also discussthedifferentpoliciesandlegislationthatauthors suggestinordertouseBigDatatechniquessafely.

V. INTEGRITYAND REACTIVE SECURITY

OneofthebasesonwhichBigDataissupportedisthe capacity to receive streams of data from manydifferent originsandwithdistinctformats:eitherstructuraldataor non-structural data. This increases the importance of checkingthatthedata'sintegrityisgoodsothatitcanbe used properly. This topic also covers the use case of applying Big Data in order to monitor security so as to detectwhetherasystemisbeingattacked.

A. Integrity

Integrity has traditionally been defined as the maintenance of the consistency, accuracy, and trustworthinessofdata.Itprotectsdatafromunauthorized alterationsduringitslifecycle.Integrityisconsideredtobe oneofthethreebasicdimensionsofsecurity(alongwith confidentiality and availability). Ensuring integrity is

criticalinaBigDataenvironment,andauthorsagreeasto thedifficultyofachievingtheproperintegrityofdatawhen attemptingtomanagethisproblem.
For example, they propose an external integrity verification of the data [7] or a framework to ensure it during a Map Reduceprocess [8].

B. Attack Detection

Asoccurswithallsystems,BigDatamaybeattackedby malicioususers.Someauthors inherit characteristics of Big Data and suggest certain indicators that may be a sign that the Big Data environmentisunderattack.
For instance, in [9] the authors a computational system develops. That captures the provenance data related to a Map Reduce process. There are also researchers who proposeanintrusiondetectionsystemespeciallyintended for the specific characteristics of a Big Data environment [10].

C. Recovery

The main purpose of this topic is to create particular policies or controls in order to ensure that the system recoversassoonaspossiblewhenadisasteroccurs.Many organizations currently store their data in Big Data systems, signifying that if a disaster occurs the entire company could be in danger. We have found only a few papersthatcoverthisproblem.Forexample,in [10]there are some recommendations regarding what can be done to recover from a desperate situation.

VI. CONCLUSION

The infrastructure security, the main problem dealt with by researchers would appear to be security for Hadoop systems. This is not surprising since, as stated previously, Hadoop can be considered as a de facto standard in industry. The remaining problemsaddressed in this topic are usually solved by modifying the usual schemeofaBigDatasystemthroughtheadditionofnew securitylayers.
The most frequently dealt with by researchers would appear to be privacy. There are a lot of different perspectivesasregardsensuringprivacy.Authorsusually propose different means of encryption, based on traditionaltechniquesbutwithafewchangesinorderto adaptthesetechniquestotheinheritcharacteristicsofa Big Data environment. Owing to the large number of papersfoundonthistopicincomparisontotheothers,we believethatitisadvisabletosplitthiscategoryupinto,on the one hand, data privacy itself, and on the other, cryptographyandaccesscontroltechniques.
The entire lifecycle of the data used in a Big Data system, from its collection to its sharing, and also includes how to properly govern the securityofthatdata.Withregardtocollectionandtoits sharing, authors propose the creation of new schemas, frameworks, and protocols with which to secure data. Other authors also suggest toughening up thelegislation concerning the privacy of the data used by companies. Furthermore,wehavefoundalackofpapersdealingwith theneedtocreateaframeworkthatcoverssecuritydata governanceinaBigDatasysteminitsentirelifecycle.
In this section, the main topic discussed by researchers would appear to be the integrity of data. In order to secure that integrity, they propose various kinds of verification to ensure that the data has not been modified.
Possibility of detecting the attacks that a Big Data system may undergo. This is probably a consequence of the high availability that a Big Data system usually achieves, but this topic should not be overlooked.

This paper provides an explanation of the research carried out in order to discover the main problems and challenges related to security in Big Data, and how researchers are dealing with these problems. This objective was achieved by following the systematic mapping study methodology, which allowed us to find the papers related to our main goal.

Having done so, discovered that the principal problems are related to the inherent characteristics of a Big Data system. and also, to the fact that security issues were not contemplated when Big Data was initially conceived. Many authors, therefore, focus their research on creating means to protect data, particularly with respect to privacy, but privacy it is not the only security problem that can be found in a Big Data system; the traditional architecture itself and how to protect a Hadoop system is also a huge concern for the researchers.

However, also detected a lack of investigations in the field of data management, especially with respect to government. They were considered opinion that this is not acceptable, since having a government security framework will allow the rapid spread of Big Data technology.

In conclusion,the Big Data technology seems to be reaching a mature stage, and that is the reason why there have been a number of studies created the last year. However, that does not mean that it is no longer necessary to study this paradigm, in fact, the studies created from now should focus on more specific problems. Furthermore, Big Data can be useful as a base for the development of the future technologies that will change the world as we see it, like the Internet of Things (IOT), or de-demand services, and that is the reason why Big Data is, after all, the future.

VII. REFERENCES

[1] . Mayer-Schönberger,V.;Cukier,K.BigData:ARevolution that Will Transform How We Live, Work, and Think;HoughtonMifflinHarcourt:Boston,MA,USA,2013.

[2] . Hashem,I.A.T.;Yaqoob,I.;Anuar,N.B.;Mokhtar,S.;Gani, A.;UllahKhan,S.Theriseof“bigdata”oncloudcomputing: Reviewandopenresearchissues.Inf.Syst.2015,47,98–115.

[3] . Eynon, R. The rise of Big Data: What does it meanfor education,technology,andmediaresearch?Learn.Media Technol.2013,38,237–240.

[4] . Wang, H.; Jiang, X.; Kambourakis, G. Special issueon Security,Privacyand Trustin network-basedBig Data.Inf. Sci.Int.J.2015,318,48–50.

[5] . Zhao,J.;Wang,L.;Tao,J.;Chen,J.;Sun,W.;Ranjan,R.; Kołodziej, J.; Streit, A.; Georgakopoulos, D. A securityframework in G-Hadoop for big data computing across distributed Cloud data centres. J. Comput. Syst. Sci.2014,80,994–1007.

[6] . Wang,Z.;Wang,D.NCluster:UsingMultipleActiveName Nodes to Achieve High Availability for HDFS. In Proceedings of the 2013 IEEE 10th International Conference on High Performance Computing and Communications&2013IEEEInternationalConferenceon Embedded and Ubiquitous Computing (HPCC_EUC), Zhangjiajie,China,13–15November2013;pp.2291–2297.

[7] . Liu,C.;Yang,C.;Zhang,X.;Chen,J.Externalintegrity verificationforoutsourcedbigdataincloudandIOT. FutureGener.Comput.Syst.2015,49,58–67.

[8] . Wang,Y.;Wei,J.;Srivatsa,M.;Duan,Y.;Du,W. IntegrityMR: Integrity assurance framework for big data analyticsandmanagementapplications.InProceedingsof the 2013 IEEE International Conference on Big Data, SiliconValley,CA,USA,6–9October2013;pp.33–40.

[9] . Liao, C.; Squicciarini, A. Towards provenance-based anomalydetectioninMapReduce.InProceedingsofthe 201515thIEEE/ACMInternationalSymposiumonCluster, CloudandGridComputing(CCGrid), Shenzhen,China,4–7May2015;pp.647–656.

[10] . Tan,Z.;Nagar,U.T.;He,X.;Nanda,P.;Liu,R.P.;Wang,S.; Hu, J. Enhancing big data security with collaborative intrusiondetection.IEEECloudComput.2014,1,27–33.