Santanu Koley

6 Big data security issues with challenges and solutions

Abstract: Big data is a collection of huge sets of data with different categories where it could be distinguished as structured and unstructured data. As we are revolutionizing to zeta bytes from Giga/tera/peta/exabytes in this phase of computing, the threats have also increased in parallel. Besides big organizations, cost reduction is the criterion for the use of small- and medium-sized organizations too, thereby increasing the security threat. Checking of the streaming data once is not the solution as security breaches cannot be understood.

The data stack up within the clouds is not the only preference as big data technology is available for dispensation of both structured and unstructured data. Nowadays, an enormous quantity of data is provoked by mobile phones (smartphone) or equally the symphony form. Big data architecture is comprehended among the mobile cloud designed for supreme consumption. The best ever implementation is able to be conked out realistically to make use of a novel data-centric architecture of MapReduce technology, while Hadoop distributed file system also acts with immense liability in using data with divergent arrangement.

As time approaches, the level of information and data engendered from different sources enhanced and faster execution is the claim for the same. In this chapter our aim is to find out big data security that is vulnerable and also to find out the best possible solutions for them. We consider that this attempt will dislodge a stride forward along the way to an improved evolution in secure propinquity to opportunity.

Keywords: Big data, Hadoop, MapReduce, data security, big data analysis

6.1 Introduction

Massive knowledge is employed by industries at all levels that have access to big data and therefore the means to use it. Software package infrastructures like Hadoop change developers to distribute storage. The distribution process of terribly giant knowledge sets on computer clusters. It simply leverages a lot of computing nodes to perform data-parallel computing [1]. With the mixture of the ability to shop for computing power on demand [2] from public cloud suppliers, such developments greatly dramatize the adoption of huge data processing methodologies. Therefore, new security

Santanu Koley, Department of Computer Science and Engineering, Budge Budge Institute of Technology, Kolkata, India

challenges have turned out from the coupling of big data with public cloud environments classified by heterogeneous compositions of hardware with operating systems (OS), and software package infrastructures for storing and computing on knowledge.

Data security is a much-needed criterion in today's Internet-based world that is dependent on mobile phone technology. Real-time use of social media such as Facebook, Twitter, LinkedIn, blogs, WhatsApp, and further Internet-based sites produces an enormous quantity of data all the way through the world. Eric Schmidt, the erstwhile CEO of Google, once said about the production in such a rate with the aim that it is fashioned by this Internet world up to 2003 is around 5 exabytes. These data are in an increasing manner in its day-to-day applications in a multiplicative manner [3]. The reason behind this growth of unstructured (e.g., MS-Word, PDF, any Text, Media Logs) or semistructured (XML, CSV, JavaScript Object Notation, etc.) data along with structured (data with relational databases) one is mainly the grounds of facet data restricted by diverse associations due to dissimilar reasons approximating enhancing of sales, detail study, analysis, escalation of social media, continuous survey, shared projects, IoT, multimedia, and so on.

Big data expertise works with other well-known technology, that is, cloud computing, provides the user much flexibility in terms of services and money, pays as per the use of singular services endowed with, eradication of costly computer hardware, has little speculation on setup, moves threats in excess of contradictory systems, and furthermore has diminutive point of time to market make cloud systems remarkably acknowledged. Cloud computing has set out novel applications in mobile technology as the entire services and amenities are provided at its finest.

The traditional applications like RDBMS (relational database management system) come into play for structured data of tabular approach stored in .csv or .xls category of formats. Today the demand for applications like Facebook and WhatsApp involves in storing unstructured data besides the structured one. The unstructured data similar to images of diverse setup together with healthcare substantiation is analogous toward X-rays, ECG, MRI images, travel and logistics as well as moveable text layouts, forms of miscellaneous kind, video as well as audio, documents comparable to text, doc, rtf, and additional setups, manuals, contacts, automotive data, data associated with safety, accessory of electronic mail, energy/industry retails, and others. Prearranged one enlightens us concerning tables of early database management systems .CSV's and .XLS's, where row and column are applicable, as conventional database management systems represent superior.

Big data follows an essential part with cloud computing as data stores in cloud-lets. Five V's of big data provides strength to analyze the data with dissimilar approaches and finally find the results. Using the elucidation endows with Google, Doug Cutting and his panel developed an open-source project called HADOOP. Hadoop scuttles function with the MapReduce algorithm, where the data is routing in comparison with others. In short, Hadoop is used to build up applications that could carry out absolute numerical study on gigantic quantity of data.

Real-time privacy with big data analysis is required. The security problem arises when distributed frameworks are used like MapReduce function of Hadoop, which dispense huge processing tasks to dissimilar systems to save processing time and breach of security crop up. These tasks are taken as input to endpoint devices. These are the main factors of security violation as data processing, storage, and other tasks are performed here.

Storage on endpoints can stock up this streaming data into several tiers. When the next data is stored, the tiers are also changed as the criteria of the most used and least used concept are there. The manual tiering system is replaced with auto-tiering, thus transaction logs cannot cop up with and increase security crisis. Nonrelational data stores like NoSQL cannot encrypt data during distributing with endpoints when it is flowing or composed, labeling or logging also makes the same problem. At the time of storing data into storage devices, a proper encryption technique is needed. Similarly the access control, encryption, and validation are also necessary when users are associated with enterprise IT as a whole. Data mining solution is another security breach as data collected from the provider and collector provides this data to the miner. This technique involves a specified mining algorithm, which ensures the privacy and security of data.

Granular auditing is a kind of security check on logs to ensure the result on different parts, where data is stored in case of external attacks. It may be an unsuccessful hit, but auditing finds its consequences. Granular access control of big data requirements by means of NoSQL databases or the Hadoop distributed file system (HDFS) intended for a vigorous verification practice and obligatory admission power. Data provenance is done with metadata, where users can check, verify, and authenticate the data with high speed for any security issues.

Securing data storage endpoints is much necessary, but at the same time, it is critical too in case of distributed architecture. When the data volume can rise up to exabytes, autotiering system is essential. This will set streaming data involuntarily allocate to indicate and put together, organizing enormous dimensions of data uncomplicated to manage. Unverified services or contradictory protocols like crisis can occur as a result. This system generates logs that can store the data about storage in tiers that can be isolated and preserved accurately. Secure untrusted data repository is a network file system that can protect the data and logs from modifying them from external unauthorized users by constantly checking and monitoring.

Organizations that employ huge unstructured data like photographs, video, audio, and text cannot use typical structured query language (SQL), rather they utilize NoSQL. It does not have default administrative user enabled, weak authentication connecting server and client as communicating via plaintext. Weak password storage, lack of encryption support, susceptible to SQL injection, and denial of service attacks are also responsible for severe security issues. A bit solution to these may include server–client encryption algorithms approximating Rivest–

Shamir–Adleman, Advanced Encryption Standard, and SHA-256 Cryptographic Hash Algorithm, as well as Secure Sockets Layer encryption.

The distributed framework separated into diverse endpoint devices need high security. The reason behind the formulation of a trusted certificate at the point of all endpoints will ensure safekeepings. Ensuring endpoint security measures include testing resources of data or information on a regular basis and usage of reliable, trusted networking devices only through the use of a platform like mobile device management (MDM).

At every endpoint, the trusted certificates will facilitate to make sure that the data stay protected. Supplementary actions to help the association ought to exploit comprise usual resource testing and allowing merely trusted devices to associate with the network all the way through the economic consumption of an MDM program. After this hardware part, a data checking should also be there to guarantee valid data received on endpoints. Now to get rid of the hackers and protect from malware application, input devices as well as applications are defenseless. Sometimes an external attacker copies with several identifications as users and fills those endpoint storages with forged data to make the system unavailable.

6.2 Cloud computing

Cloud computing is the aspect of the present computing globe. It is a centralized approach in stipulations of data storage, improved operations, and ubiquitous, well-located, on-demand web access with smallest disbursement. It moves behind a collective pool architecture, where essential networks, special servers, enormous storage, numerous applications, and dissimilar services are incorporated [4]. The cloud model is an amalgamation of five indispensable distinctiveness, three service models, and four deployment models as looked for [5].

Cloud computing follows a service-oriented architecture (SOA), where it is divided into three main categories of services like IaaS (infrastructure as a service), PaaS (platform as a service), and SaaS (software as a service). This on-premises solution hardware or software is just like purchasing a car where driving, alteration, and changing routes as per requirement is possible as per the buyer/user. While IaaS can be an example of leasing a car from someone and drive it as we like, but in case of upgradation several others can be taken. On the other hand, PaaS can be described as taking a taxi for some time and pay the fare only, but cannot drive it either by modifying anything we dislike. Conversely, SaaS is similar to using a bus with predefined routes. Fundamentally, IaaS is employed by IT administrators; software developers are bringing into play; PaaS and SaaS make use of the simple end users.

IaaS cloud service providers are Amazon Web Services Elastic Compute Cloud (EC2), Microsoft Azure, and Google Compute Engine. The big players of PaaS are

Heroku, AWS Elastic Beanstalk, and Google App Engine. Finally, as an end user, we are aware of the names of SaaS such as Gmail, Trello, Salesforce CRM, EventPro, Office 365, and Google Docs.

The cloud does not stand for the traditional one as the name recommended is a flexible service by means of service-oriented structure. It makes use of the Internet intended for supplying certain services. The data, computer hardware along with software everything, is shared in this construction. “Share and use of applications and resources of a network environment to get work done without concern about ownership and management of the resources and applications” (M-S. E Scale, 2009) [6–8]. At this moment, the SOA put in the picture concerning the services provided through this technology is not something except contradictory transformation of several extraordinary technologies. The exceptional form of cloud service provides service and deployment models. Service model can be divided into three different varieties of perception, resembling PaaS, SaaS, and IaaS. Service models are described as the NIST model [9] for the above structural design. The discussion division for us is the IAAS cloud of the recently developed system.

The fantastically well-preferred cloud figuring is referred to as its notable highlights as pay-per-utilize model, which makes it a minimal effort, resource pooling, simple to introduce and usage capacity, arrangement of services as required by the client and ranch out capacity, QoS, wide network access, suppleness at a brief beat, self-stipulation, decided upgrade of administration, adaptable, relentlessness, easy upkeep with updegree, squat fence vitality, and so on [10].

“MCC at its simplest refers to an infrastructure where both the data storage and the data processing happen outside of the mobile device. Mobile cloud applications move the computing power and data storage away from mobile phones and into the cloud, bringing applications and mobile computing to not just Smartphone users, but a much broader range of mobile subscribers” [11, 12].

As the processing task isn’t finished by the cell phone, the power and memory utilization is likewise less here and in the long run, the cell phone turned out to be quick [13].

IaaS cloud innovation is the suggested technique for putting away information on cell phones, apps, and so forth as second code portion or CDroid [14] server plot. To achieve the minimal effort phrasing, we should find the cloud server in a neighborhood land, where Fujitsu server is situated must have most reduced power cost for every unit in this world [15, 16].

6.3 Big data

Big data is an expression that is used to allude different sets of data that are exorbitantly tremendous or compound for customary information handling application

programming to viably manage. Information with numerous cases (lines) offers more prominent measurable power, while data with higher multifaceted nature (more qualities or segments) may prompt a higher false disclosure rate.

Big data challenges incorporate catching of data, data stockpiling, data analysis, seeking, sharing, exchange, perception, questioning, refreshing, information security, and data source problems. Big data are related to distributed (cloud) computing as data put away onto clouds can be depicted in Figure 6.1.

Figure 6.1: Cloud computing with big data.

The present routine with regard to the expression “big data” is slanted to submit to the utilization of prescient analytics, client conduct investigation, or certain other propelled data analytic techniques that separate an incentive from data, and sometimes to a specific size of the dataset. “There is little doubt that the quantities of data now available are indeed large, but that’s not the most relevant characteristic of this new data ecosystem” [17]. Analysis of data collections can discover new relationships to “spot business patterns, anticipate ailments, and battle wrongdoing thus on” [18]. Scientists, business officials, experts of drug, publicizing, and governments alike consistently meet troubles with extensive datasets in territories including Internet look, fintech, urban informatics, and business informatics. Researchers experience restrictions in e-science work, including meteorology, genomics [19], connectomics, complex material science reproductions, science, and ecological research.

Big data is an idea to store, process, and break down gigantic sum (e.g., exabytes) of data that is almost incomprehensible with conventional RDBMS

framework. This is on the grounds that other than organized data it manages semi- or unstructured data. Big data is grouped into a few classifications like data sources, content arrangement, data stores, data organizing, and data preparing. The detailed order can be comprehended with the assistance of Figure 6.3.

Data sources are named as web and social media, a content format known as structured, semistructured, or unstructured configuration.

The data format is document oriented, column oriented, graph based, and key qualities while data staging can be depicted as cleaning, normalization, and change. At long last, data processing should be possible with the assistance of two unique strategies like batch processing and real-time processing.

Big data is able to illustrate by means of three ideologies of “V” such as volume, variety, velocity, and it can be extended to veracity, validity, volatility, and value (Figure 6.2).

Figure 6.2: The V's of big data.

- **Volume:** Volume talks on the extent of the data created from various sources and is prepared to extend as far as records, tables, or other forms. The size of the data may vary up to terabyte, petabyte, and zetabytes. Information made with definite data analysis and smartphone are the greatest maker of such sort of longitudinal data [21].

- Variety: Variety means the class of data, for example, in various structures. Furthermore, various sources will deliver big data, for example, sensors, gadgets, social networks, the web, and cell phones. For instance, data could be web logs, radiofrequency identification sensor readings, unstructured social networking data, gushed video, and sound. Other than unstructured data, both structured and semistructured assortment of data stores into big data.
- Velocity: This implies how regularly the data is produced, for example, data analysis. For instance, each nanosecond, millisecond, second, minute, hour, day, week, month, and year. Handling recurrence may likewise contrast from the client necessities. A few data should be prepared real time, batch, or stream too and some may not.
- Veracity: Veracity is the data in doubt, that is, uncertain, untrusted, and unclean. The uncertainty is due to data inconsistency and incompleteness, ambiguities, latency, deception, and model approximations. It is the management of the reliability and predictability of inherently imprecise data types.
- Validity: Validity of input data based on accurate processing is done on the data and provides a particular output as the product. The validity of data is very near to the veracity of the data. Through big data, one should be spared attentively regarding validity. For example, in the banking sector, data accumulated from various banks must be valid to show the growth of the nation. The finance ministry may plan their strategy for the next financial year regarding the valid data provided.
- Volatility: Volatility is very common when data is needed to transform into other types on a regular basis. If it is not accounted for analytical results, perhaps invalid at the instant they are produced. Such types of circumstances are very common, where businesses like the stock market or a telecom company (call data records related to one day). Volatility is directly linked by means of the challenges of invalidity and veracity.
- Value: In big data, value means exact data that are in reality and has some meaningful aspects that take out from some usable system. Apparently, this must be the output of big data processing. Value is an essential feature in the big data. Extracting the exact value is only possible when proper roads are there; here road refers to IT infrastructure arrangement to collect big data. These can be possible in dealing with a business that is meant for a return on investment.

There is another factor called inconstancy, which can be an issue for the individuals who break down and analyze the data. This alludes to the irregularity, which can be appeared by the data on occasion, along these lines hampering the way toward having the capacity to deal with the data adequately.

Big data are portrayed by various perspectives: (a) data sources, (b) data are various, (c) data can't be sorted into standard social databases, (d) content

arrangement, (e) data stores, (f) data organizing, and (g) data are created, caught, and processed quickly. The real classification of big data can be illustrated in the hierarchical structure as shown in Figure 6.3.

Figure 6.3: Big data classification [20].

The distributed file system structure is utilized to store in big data utilizing two unique strategies like HDFS and MapReduce programming system. The two are utilized to store and keep up the entire data structure (Figure 6.4).

Figure 6.4: Life cycle of big data.

6.3.1 Hadoop distribution file system

HDFS is responsible to collect the data in lumps. Here data is parted into squares of 64 MB each. It is made perfect for executing its file framework on a hardware

platform where the cost is lower as well. It can endure the exceedingly imperfection framework as it keeps up block replication. The Internet searcher for the web application as named Apache-Nutch venture was the explanation behind the production of this engineering. Apache-Hadoop is the great introduction as a subproject of the resulting web crawler. An occasion of HDFS is made of a major number of server PCs; some of the time the amount can leave a couple of thousand as well. These servers reserve the data of the said record framework. They are not set up for a solitary collaboration by the clients, yet for cluster preparing. The data documents are a lot greater past evaluation, as they achieved the size close to TBs. Here the low latency of data access is ignored; however, high throughput is upgraded and that is much fundamental for a framework like HDFS. Data coherency and this throughput are the establishment of the idea of “compose once and read various occasions” idea of the files.

The Hadoop architecture contains data sources, Hadoop system, and big data understanding. The data sources contain site clickstream data, content administration framework, outside web substance, and user created content. The Hadoop system incorporates HDFS that is controlled with big data landing zone and MapReduce algorithms. These algorithms are constrained by keywords investigate, content characterizations, or subjects and client division. The big data knowledge holds keyword-applicable substance-rich client focused on presentation pages.

HDFS is set up to utilize Java technology that takes care of business. The basic plan of this document framework is given a name of NameNode – DataNode followed on master–slave development. Serving read–write operation/application from the file frameworks, customer is performed by the pair. The guidance from NameNode is completed by performing block creation, cancelation, and replication. They are expected to execute on item frameworks that might be OS as GNU/Linux as NameNode is somewhat software same as DataNode. NameNode runs file system namespace tasks like opening, shutting, and renaming files and directories. It is a master server and administrates the file framework namespace and controls access to documents by customers.

NameNode finishes up the mapping of different blocks to DataNodes. To guarantee the DataNodes working precisely, that is ordinarily single per bunch, the NameNode engages time-to-time refreshing of heartbeat and a block report from DataNodes in an ordinary interim. DataNode might be called as a vendor for the hub with its arrangement, as a gathering of DataNode is upright for amassing of a few blocks. These blocks contain part documents dispersed in changed blocks.

HDFS depicts a file system namespace and, put aside, client data to be put away in these documents are in blocks.

NameNode does not flood with user data; it is an arbitrator for all of HDFS meta-data. HDFS is planned such a way that it has a solitary NameNode in a bunch to the most noteworthy degree that disentangles the design of the framework. Hadoop

system has the layer, in particular HDFS layer and MapReduce layer. The second one is the known execution engine in a multinode cluster. A job tracker colleague tasks trackers in both master and slave parts; then again, name node in the HDFS layer partner's data nodes in those parts as appeared in the outline in Figure 6.5.

Figure 6.5: Master–Slave architecture of HDFS.

6.3.2 MapReduce

MapReduce is a data handling worldview, where a major measure of data is transformed into little. This programming model is combined with the task for doling out and causing cumbersome data collections with a parallel, distributed algorithm on a bunch [23, 24]. Hadoop is a framework that can store and control a lot of data all around effectively, in light of basic master–slave engineering. It is the center of Hadoop. Thoughtfully comparable methodologies have been very outstanding since 1995 with the Message Passing Interface [25] standard having decreased [26] and dissipate activities [22].

MapReduce is separated into a few applications, patterns, examples of overcoming adversity, utilities, capacities, highlights, and executions. Utilizations like questions and examination, works as map and reduce, includes in the vein of a programming model, huge-scale dispersed information handling, straightforward yet limited, parallel programming, extensible, motivated in useful programming however not comparable and once in a while think in recursive arrangements.

Executions looking like with Google, apache-Hadoop, and a wide range of innovations are utilized effectively in MapR. Different maps that lessen systems like Signalcollect and storm are utilized here. Security and privacy challenges in big data biological system are depicted in Figure 6.6 in the above section.

Figure 6.6: Security and privacy challenges in big data ecosystem [22].

6.3.2.1 Big data security structure

Big data security structure can be characterized into various classifications. Some of them can be named as infrastructure security, data protection, data management, and integrity and responsive security [27]. The outline depicts the design in detail. Foundation of security might be portrayed in cases like dispersed systems and nonsocial information stores, although information protection shows information mining and examination, cryptographically authorized information security and granular access control. Data executives depict data tiering, exchange logs, granular examining, and data provenance. Respectability and responsive security delineate constant protection and endpoint gadgets. The arrangement can be portrayed in Figure 6.7.

Big data is the most recent innovation utilized by associations that get vulnerability as we are uninformed of the vast majority of the things. The vast majority of the devices bringing into play are open source, and accordingly discover assaults of the hubs where information stores. Data stores here conveyed in nature; subsequently, ill-advised client verification happens. There is a significant prospect for pernicious data input and inadequate data approval.

The examination of the advancement of big data uncovers the high adequacy of big data as far as data handling. Be that as it may, the data preparing and information stockpiling of big data raises the issue of the data security ruptures and infringement of clients' protection. At the equivalent time, the fundamental exercises

Figure 6.7: MapReduce architecture.

learned are the questionable idea of big data in light of the fact that, from one perspective, the advancement of big data raises data security dangers; in any case, then again, big data can possibly improve the data security in the event that they are legitimately utilized. The capability of big data is gigantic and the consideration regarding the data security is basic for the powerful improvement of big data and the avoidance of various dangers.

6.3.2.2 Real-time privacy

Constant supervision of real-time security is much demanding when dissimilar protection procedure engenders a gigantic amount of observant. The perceptive may go ahead numerous tribulations. They are frequently unobserved as the users cannot deal with the cut-off quantity. This may have a huge setback and also enhance further amid big data as the volume and velocity of data pour out. On the other hand, big data expertise endows with a prospect to facilitate this competence to carry out and consent to prompt processing and analytics of a remarkable type of data and this, in turn, can be worn out to formulate an existing real-time variance discovery associated with scalable safekeeping analytics as well. In real-time privacy, analytics and the use cases are diverse in dissimilar business applications where that particular industry will get benefitted (Figure 6.8).

For example, e-commerce and consumer marketing industry obtains huge support and profits in terms of monetary benefits. The same thing has to happen with healthcare industry when doctors need accurate data in terms of report generation for medical tests where doctors to prescribe. This situation is much similar when tax paying for a country as well. The tax calculations, returns, payback, advance

Figure 6.8: Different big data security and privacy challenges [22].

tax, and claims sometimes get frauds done on tax payments. Here the problem arises when accessing the data between different parties, resources of those data, and accessing the data in office or none office hours.

Today we move to computation on real-time data where big data faces most challenges. Here, real-time updating or keeping an eye on the websites and web pages is completed. The gigantic quantity of data (sometimes tera/petabytes) is composed on or after a variety of resources, sorted out, scrutinized by means of numerous data mining, data classification, and prediction algorithms, and consequently, reports are kept up of all these analyses. These prepared reports are exceptionally helpful when decision-making standards are satisfied. The carrying out of an organization depends to a great extent on those accounts. Language processing is a real-time data processing language used to process data streams coming from multiple sources (Figure 6.9).

IBM's Stream Processing Language has three singular varieties of operators: utility, relational, and arithmetic, which take data through input source operator and give output through output source operators. These multiple operators present in between the source filter, aggregate, join multiple data streams according to the need of the user. As per the necessities, the formulations of the operators can be executed manually by the users. Processing of streaming data is put in a more competent technique in big data, whereas it also props up ad hoc queries.

Here the end users can write their own query in SQL as in custom database application and straightforwardly submit them to relevant web applications too. Thus, it can get further flexibility and power. But there is a larger security concern of having ad hoc queries. The entire practice can be expressed in Figure 6.10.

Figure 6.9: Time security monitoring (real-time privacy).

Figure 6.10: Real-time streaming and continuous computation.

Solution

At that place should be a control on the admittance to the data; moreover, it ought to be monitored. To put off illicit right of entry to the data, threat intelligence should be employed. Use big data analytics to set up confidence acquaintances to come together to make sure that merely authorized links to take place on a cluster. Scrutinizing tools like security information and event management (SIEM) way out can be brought into play to stumble on uncharacteristic associations. They may include:

6.3.2.2.1 Secure authentication gateways

Powerless confirmation system is a standout among the most widely recognized components that contribute toward information ruptures. Uncovering the vulnerabilities present in client confirmation work, a programmer can possibly access delicate information. Imperfect execution of client confirmation process must be counteracted at the plan organize. Guarantee that there are no broken confirmation tokens, which can be misused by any unapproved clients.

6.3.2.2.2 Utilize principle of least privilege

We ought to be in a perfect world to keep up a layered access control and actualize standard of least benefit. It advocates constraining client access to the insignificant dimension that will permit ordinary working. As it were, we should give a client just those benefits that are fundamental for that client to deal with his/her duties. It would keep unscrupulous IT experts from enjoying unlawful information mining exercises.

6.3.2.2.3 Utilize retrospective attack simulation

Not all associations can work in-house framework to help big data activities because of monetary requirements. Big data venture depends on an outsider cloud-based (public or private) arrangement; at that point review assault recreation can be utilized to discover vulnerabilities with the outsider application facilitated on the cloud. On the off-chance that the assault succeeds, at that point you ought to examine the issue further to locate changeless goals. Review reproduction would assist with identifying plausible shortcomings in the framework before a veritable programmer endeavors to abuse the helplessness.

6.3.2.2.4 Utilize latest antivirus protection

Numerous antivirus merchants have concocted security arrangements that are explicitly focused toward big data activities. So dependably ensures big data condition with the most recent Antivirus suite. Ensure that the updates and fixes are introduced when they are made accessible by the producer.

6.3.2.2.5 Utilize principle of least privilege

Big data is a developing business sector and the advancements are continually developing, making it hard for the current security answers to stay aware of the expanding request. Intermittent reviews will assist identifying new vulnerabilities as they make their essence felt. Subsequently, it can realign the security consistence with the present security guidelines.

6.3.2.2.6 Secure coding practices

While evaluating the code, one should ask himself the accompanying essential inquiries like, Am I ready to comprehend the code effectively? Is the code composed after the coding norms/rules? Is a similar code copied more than twice? Will I unit test/troubleshoot the code effectively to discover the underlying driver? Is this capacity or class too enormous? On the off-chance that truly, is the capacity or class having such a large number of duties? On the off-chance that one may feel that the appropriate response isn't tasteful to any of the above inquiries; at that point you can propose/prescribe code changes.

6.3.2.2.7 Input source authorization

The utilization to constrain which wellsprings of information are substantial for employment accommodation, including workstations, gadget perusers, hubs, and interior perusers. For instance, it should need to keep certain clients from entering employments from a specific workstation.

To approve the accommodation of work from explicit information sources, request that the security administrator can enact the class and characterize a profile for each information source. Furnish a security manager with info source and gadget names.

6.3.2.2.8 Employ data analytics

Big data analytics is the regularly mind-boggling procedure of looking at extensive and shifted data collections or enormous information to reveal data, including concealed examples, obscure connections, advertise patterns, and client inclinations that can enable associations to settle on educated business choices.

As a conclusion to real-time privacy, solutions endow with the use of security devices, secure coding habits, input source authorization, and employ data analytics. Figure 6.8 briefs the story in short.

6.3.2.3 Distributed framework

The distributed computing framework is the modern approach today, where not only hardware, the software is distributed to. Here a software component shares

multiple computers in a network in order to get the faster processing capability. Distributed programming frameworks (DPF) tie together the primary, secondary storage, processing of data in an enormous quantity by size. Here in this framework (MapReduce), they split the parallel computation; storage functions process mammoth volumes of big data. It could identify wicked mappers, and look over the data entrusted mappers is the prevalent setback of big data and can potentially get in the way big data seclusions endeavor.

For example, MapReduce in Figure 6.7 is a programming model and a related usage for handling and creating substantial data collections with a parallel, distributed algorithm on a cluster. Normally, it segregates the input datasets into self-regulating pieces. These components of a set are processed by map jobs in a comprehensive equivalent approach. MapReduce, in its initial stage, is naught but a Mapper for each lump understands the data by a process called reading; there is a little calculation is executed with the data already read and amount produced a list of key/value couple. In the following stage, a reducer consolidates the qualities that are in the correct circumstance to each unique key and yields the outcome.

The attack anticipation assesses in two special ways: mapper security and the security of the data in the existence of an untrusted mapper. The untrusted mappers might revisit unidentified consequences, which will, consecutively, bring about an inaccurate cumulative outcome. When using large-scale datasets, it is incredible to recognize the outcome of momentous smash up in scientific and financial calculations. Different marketing and advertisement agencies collect and analyze buyer–retailer data to reach consumers they have already marked. These errands include a high measure of parallel estimations over expansive informational collections and are especially appropriate for MapReduce systems, for example, Hadoop. The information mappers may contain spillages that might be deliberate or unexpected in nature. For instance, a mapper may release an extremely particular incentive by breaking down individual data, undermining clients' security.

Figure 6.11: Distributed frameworks' secure computations.

To carry out a gigantic quantity of data, DPF makes use of parallel computations. DPF makes use of parallelism in computations with storage space to practice an enormous amount of data (Figure 6.11).

There are two different methods presented to make certain the trustworthiness of mappers: trust organization and mandatory access control (MAC). Throughout the first part, that is, trust establishment, “workers” must be genuine along with prearranged belongings by “masters,” and only when they’re experiencing can they be doled out mapper responsibilities. Subsequent to this requirement, periodic updates must be constructed to ensure mappers, again and again, and congregate the recognized procedure.

Alternately, MAC, the predefined security approach, will help to follow out. On the other hand, while MAC makes certain that the input of mappers is safe and sound, it does not put off data loss from mapper output. To keep away from this, it is important to influence data de-identification techniques that will set off the erroneous info from being circulated among nodes.

Solution

The seclusion and security problem take into account of a number of questions like auditing, access control, authentication, authorization, and privacy once bring into play the mapper and reducer process. The way out pays trusted third-party monitoring and security analytics (Apache Shiro, Apache Ranger, and Sentry) just as protection arrangement implementation with security to put off information spillage. In that regard is inalienable lack of clarity in administering compound application reconciliations on the creation-scale distributed framework. The system is to work by methods for an undertaking class programming model that has the office to grasp remaining task at hand strategies, tuning, and general observing and organization. At that point, when we make an application for a solitary office or different capacities, we influence an IPAF (to entomb protocol acceptability framework). The solution provided with the following:

6.3.2.3.1 Trust establishment

This blended gift additionally commands the connection between big data and trust. On the one hand, a lot of trust-related data can be used to build up creative data-driven methodologies for notoriety-based trust of the board. On the other hand, this is naturally attached to the trust we can put in the sources and nature of the basic data. There may be situation when trusted nodes in distributed structure malfunction with compute working nodes; the solution for this is to make regular updates on each node periodically.

6.3.2.3.2 Mandatory access control

MAC is a lot of security approaches obliged by framework arrangement, setup, and confirmation where access to sensitive data can take place. Macintosh arrangement, the board, and settings are built up in one secure system and restricted to framework

chairmen. Macintosh characterizes and guarantees a unified implementation of private security strategy parameters.

6.3.2.3.3 Privacy preserving transformations

Because of data gathering from various sources, odds of protection break have expanded. It is hard to apply existing protection models (security safeguarding procedures) in big data investigation due to 3Vs: volume (substantial measure of information), variety (organized, semistructured, or unstructured information), and velocity (quick age and preparing of information) – qualities of big data.

Thus, distributed framework solutions afford trust establishments like commencement, occasional trust update, obligatory access control, and protection safeguarding changes. This course of action is explained at this point in Figure 6.10.

6.3.2.4 Endpoint devices

The above framework of big data accumulates data from an assortment of resources. They are generally called as endpoint devices. The technique of collecting data split into two classes of perils such as data collection with validation and filtering of data. The first part (data collection with validation) is to collect data from several endpoints connected in a distributed network, where millions of hardware and software are associated with it in an enterprise network.

Here another problem arises when an input validation is performed on the piled-up data. Substantiation of the input data is important as infected data are too compiled with some malware application with computer viruses that may harm the data sources. Now the data should be filtered and modified as per the given format of the data requirements. The second part (filtering of data) provides the exact outcome improbable to validate and filter data. The data mapping can be done with the help of knowledge processing and business goals or assess the scope of customer data. Previous signature-oriented data filtering may perhaps be unsuccessful to solve the input validation and data filtering to slow down completely.

Solution

Various solutions include the following:

6.3.2.4.1 Corruption free software

Tamper-proof software is required to apply the accumulated data from diverse endpoints. Finding out the proper software is awfully essential with the existing setup. The depravity of free software is much asked for swift operation, virus detection and removal, and use of original software obligatory.

6.3.2.4.2 Trust certificate and trusted devices

The trust certificates are necessary with the use of trusted devices. These are really difficult when collecting data from various sources, but during the filtration process, one can ensure the said certificate and devices. Filtration process can include firewall protection.

6.3.2.4.3 Analytics to detect outliers

The purpose of data analytics may help to find outliers. When special interpretations on data that move away from normal path and produce extreme values are called outliers, they will point out changeability during a measure, tentative inaccuracy, or a novelty. In alternative terms, an outlier is an associate inspection that deviates from a taken whole outline on a section.

Outliers can be divided into two categories, namely univariate and multivariate. Univariate outliers are originating on one occasion staring at an allocation of principles during a particular attribute area. Multivariate outliers are set up during an n -dimensional area (of n -characteristics). Staring at allotments in n -dimensional areas is often terribly tough for the human intelligence, that's why we need to prepare a replica to try and do it for us.

Outliers can even be available in completely diverse flavors, looking on the surroundings: point outliers, contextual outliers, or collective outliers. Point outliers are distinct data tip that lay off from the remainder of the distribution. Contextual outliers are noise in data, like punctuation symbols previously become conscious text analysis or background noise signal once doing speech identification. Collective outliers are subsets of novelties in data like a symbol that will point out the invention of recent trend.

6.3.2.4.4 Cryptographic protocol

Cryptographic protocols such as triple DES (put back the main data encryption standard utilizes three individual keys with all 56 bits. The entire key length implies 168 bits, yet masters would fight that 112 bits if key quality is progressively like it), RSA (an open key encryption estimation considered as amiss computation in light of its usage of two or three keys), Blowfish (symmetric consider parts messages along with squares of 64 bits and encodes them freely), Twofish (keys used in this figuring may be up to 256 bits in length and as a symmetric framework, only a solitary key is required), and AES (advanced encryption standard, incredibly successful in 128-piece structure; similarly, uses keys of 192 and 256 bits for significant encryption purposes) are there to compile the encrypted data collected from other origins.

They may be exceedingly useful when using semistructured or unstructured data. This procedure of endpoint input validations is exposed in Figure 6.12.

Figure 6.12: Endpoint input validation/filtering.

6.3.2.5 Data into several tiers

Tiered storage of data allocates to two or more dissimilar categories of storage media. They depend upon the business value that data collected so far. This business value enables the data tiering at an exceptionally low cost, but collecting the maximum data possible within. Sometimes the significance of data alters over time. There are some events that establish the altering worth of data, such as how often we access the data and for what time, what capacity you need to fend for its functions and exceptional circumstances that call for its quick retrievals, such as an account audit or financial account. Tiered storage includes the physical aspects, where hard disk, tape, or disk storage, as well as logical aspects like cloud storage, is associated.

Tiering approach of data storage in said devices depends upon the type of user data with its applications. As time progresses, this data tiering complexity increases (in a few cases, it decreases too) and it is classified into unusual types; thus considered necessary to store data in different types of storage medium. Generally, this is done manually, but the system may upgrade to automatic data storage tiering. The foremost companies help autotiering system, namely IBM, NetApp, and EMC. They can submit application, that is, policy-based rules to shift data involving dissimilar tiers automatically.

The advantages of the tiered storage system are cost efficient, operationally effective, and elastic. The tiered data storage system compels massive diminutions in storage costs compared with untiered storage. Today, the organizations realize instead of a single sort of depot for such data are a misuse of money for the majority of jobs (Figure 6.13).

Tiering storage is operationally very much efficient as per the different need of time and type. Such storage consists of the data that give out essential business functions that end up in high-performance storage media, such as solid-state drives

Figure 6.13: Multi-temperature splits big data.

(SSD) through archival data among little significance that ends up in tape storage or low-cost cloud storage services.

The normal manual train system is flexible enough, but automated tiering system has the largest flexibility ever, to shift data among diverse storage media as its value changes.

Big data describes distributed processing frameworks like Hadoop that are worn for data processing and tiering storage for big data applications. Hadoop's substantial storage potential appears on or after its clustering architecture, wherein data are distributed and lay-up in a network of compound computing possessions. If subscribed with a specific cloud setup, the data tiering part may be stored in the cloud data center; otherwise, Hadoop setup may be installed on-premise for enterprise-specific area.

At the very beginning, the datasets are created and afterward, they are used on a regular basis by several users from the diverse user from dissimilar endpoints. Seeing that large datasets come in Hadoop clusters, a component of the data is stock up on individual computing machines or nodes in a cluster.

Yet, dissimilar variety of data along with the regularity of data access is very frequent but with the progression of time, the value of data diminishes. Data tiering is supported by Hadoop, and by splitting the cluster into different tiers based on the frequency of use, it can considerably shrink the costs to pile up this data. This technique of reduction of cost in data storage is continuous in nature. It is publicized as (Figure 6.14).

Hot data: The real-time data analysis needs high-frequency entrance of data. The regular reporting or ad hoc queries are also associated with this type of data. The utmost computing power is utilized with this kind of data.

Warm data: This type of data accessed seldom. But it is functional at times so that it is necessary to be stored on hard disks or sometimes in SSD storage too. The computing power needed is obviously much less than hot data.

Cold data: Cold data is essentially a library type of data called archival data. This character of data is used in enterprises at times when the periodic report is generated or an enterprise wants to retain for compliance or once-off queries. This data is located in storage by means of negligible processing power obtainable of it.

Frozen data: The frozen data means it will almost never be used. This kind of data does not store in basic computer storage areas like HDD. Such kind of data is

Figure 6.14: Data tiering hierarchy.

almost unusable, so the power consumption is near to zero. They can be stored in a node that uses minimal computing power, and less processing task is performed. Here data is stored in an achievable manner (Figure 6.15).

Figure 6.15: Mapping data to a storage tier based on temperature.

Solution

The cost reduction in data storage tiering is done for easy handling of data. Data is separated according to the frequency of use of data with a specific time frame stored in different nodes. Reduction of cost for data storage depends on the storage; when data stores on nodes with minimum computing power or in case of achieving data, the data with minimum computing power saves a huge amount of money. The data can move among said tiers via Hadoop tools, for instance, Mover. If the system turns out to be relatively dynamic, it sets aside for superior competence in the big data storage technique.

Now big data are classified into singular storage tiers per the occurrence of its usage. This is an excellent opening tip for the enterprise that wants to store information in a less pricey manner. After sometime, when this process is rationalized, user organization need not think about its storage and reallocation of the same data with its cost minimization as they move between tiers.

The development of every job is directly proportionate with the addition of data as big as it constructs the peak of big data that gathers by themselves. A small part within big data is useful in

this instance as the cold and especially archive data exist. Thus there is a perspective to categorize and tier enterprise level big data as soon as it gets into the clusters, leading to even greater efficiencies.

Storage tiering has immense prospective in a business world, where industries are under pressure to accomplish useful imminent commencing the bulky bind of data they gather on a habitual basis. Now the data will simply maintain to breed in volume and velocity in terms of big data. Tiered storage brings cost optimizations to the table that can guarantee organizations achieve the correct equilibrium between performance, capability, and cost on their big data clusters.

Storage tiering has extraordinary potential in this present reality where organizations are feeling the squeeze to increase valuable bits of knowledge from the extensive swathes of data they gather all the time; data that will just keep on developing in volume and speed. Layered capacity conveys cost enhancements to the table that can guarantee associations that accomplish the correct harmony between execution, limit, and cost on their big data groups.

6.3.2.6 Cryptographically enforced access control and secure communication

The big data analytics system can be made automated in terms of collecting data, but as a result, it enhances data loss during this period. Different encryption techniques and proper training of users can minimize the risk associated with it. Adequate protection method must be introduced as big data stored into clouds should be tested accurately. This can be done for extra protection of data on the user side while cloud service provider confers customary precautions review within the time frame.

Cloud service providers can be imposed penalties if the security standard does not meet up to the expected standards. The right to use of power strategy must be set up for authorized user access to both internal and external user sites. User authentication for the data coming from different sites must be controlled up to some level. To protect data from unauthorized access, a second-level security mechanism is very much useful, namely encryption (Figure 6.16).

The reason behind doing the same is the use of raw data as well as analyzed data. Confidentiality and integrity should be imposed on data as a measure of data protection.

Another aspect of the data security mechanism is the use of antivirus and firewall protection. Today the trend is the creation of special attacks like ransomware – a type of malware that is very common. They are breaching the defense of computers throughout the world on a regular basis. Security mechanism includes some other small techniques like disconnection of user devices with servers restraining important information when they are unused. The security mechanism must pay attention to the fortification of the function, rather than just safekeeping the device. Proactive and reactive security skill should be supplied to big data.

Data visibility manages to different entities differently as there are two methods for organizing, like systems, individuals, and organizations. The primary technique systematizes the visibility of data in the form of preventive admittance to the main scheme, for example, the OS itself or hypervisor. The secondary technique encapsulates the data itself in a self-protective shell by means of cryptography.

Figure 6.16: Cryptography as encryption and decryption.

Both the approaches put in their remuneration and detriments. In the past, the foremost procedure was simpler to carry out and, as integrated among cryptographically protected statement. These may be customary on behalf of the prevalence of computing and communication infrastructure.

The conventional security methods legitimate for an assortment of protective intimidation. They include replay attacks, password-guessing attacks, spoofing log-ins, intercession that has chosen plaintext attacks (Kerberos-specific attack), and session key's exposure. These attempts are really common in big data security techniques. They are pertinent to any of the customary verification schemes. The peculiar and more protected system, namely Kerberos fails sometimes. Today there is a need for implementing a scheme that can forestall the attempts.

Solution

Instead of using cryptography in security control, steganography can be used where data is covered up by some other media. In the case of cryptography, data are actually changed in its format but steganography creates an illusion for the intruders who want unauthorized access to user data.

To address the security problem like access control in big data is setting a revision in the secure remote password (SRP) protocol to have room for the access control of the clients in authentication level. This is the aim of assigning the access labels to the big data users to limit their access rights in the big data environment. Simple remote password protocol is a secure password-based authentication and key management protocol (Figure 6.17).

This protocol authenticates the clients to the server using a password-like mystery. This mystery must be known to the client only. No other secret information is needed to remember by the client. The server stores the verifiers for every user to

Figure 6.17: Cryptographically enforced access control and secure communication.

authenticate the client, but if this verifier is compromised by an attacker, it cannot be used to by the attacker to impersonate as a client. The foremost benefit of the SRP protocol above further verification methods is that there is no need to store any password equivalent data and the systems are immune to the password attacks. When the client is verified by the server, a cryptographically strong secret is exchanged by the SRP protocol between the communication parties to pass securely.

Ultimately, the cryptographically enforced access control and secure communication solution present as a conclusion are identity and attribute-based encryptions, encrypting techniques supporting search and filter, fully homomorphic encryption, and group signatures with trusted third parties.

6.3.2.7 Transaction logs

The protected data storage and transaction logs are really much linked up in the midst of data warehousing, management, processing, and hence make the security tribulations. Sometimes external security threats may distort the memory due to unauthorized access within big data. This deformed data is then transferred from one source to another while the data transmission occurs between different nodes and users. The tiered approach has already described how data is stored in different tiers. This case is likewise for the transaction logs, and both data and transaction logs are stacked out in multitiered storage media. The data moves frequently among tiers manually.

This procedure assures the developer direct authority over vitally what data is moved and when. On the other hand, as the data set volume produces exponentially, scalability and accessibility require autotiering for big data storage management. The autotiering solution does not have any track of where the data is stocked up. This leaves a new problem of protected data storage. For instance, on that point is an establishment that wants to ingest data from different departments of their own. There are dissimilar types of data (every bit per data tiering approach) available within the

departments. The data that are virtually not used and mostly used both exist in parallel. An autotiering storage system will help a great deal and can put away the money via transferring the less used data to a lower grade and thus along. It may find that the data stored in a lower tier is sensitive information. Usually the lower level data have less security as companies do not desire to spend much for the unused ones, only in this event, they must convert the policy to send data toward lower level or increase the security of data.

Exchange logs can develop wild when not appropriately kept up. Each time data is changed in the database, records get added to the exchange log. On the off-chance that an exchange is kept running against an extremely huge table, the exchange log must record and store those data changes simultaneously until the exchange is finished. Since the log composes its data to disk, this can gobble up a ton of disk space all around rapidly (Figure 6.18).

Figure 6.18: Transaction log (basic working procedure).

Solution

Data tiering in big data analysis plays a major role, especially when an autotiering scheme is adopted. This autotiering must be fully log based so that the movements of data between tiers are noted and side by side the sensitive data should be marked properly and increase the security of the tier itself. Now, this may be the problem for huge data to provide high security, for this case data should be isolated and kept in a safe place. The encryptions, especially policy-based encryption, and a signature on data safe secure data storage and transaction logs. The proof of data possession and periodic audits of data are also recommended (Figure 6.19).

6.3.2.7.1 Recuperation models

To begin with, pick the correct database recuperation model. There are three that can be utilized with an SQL server database, yet the most normally utilized models are SIMPLE and FULL. The third alternative, BULK_LOGGED, is normally utilized briefly for expansive activities where execution is basic. The penance to get this execution help is potential information misfortune if something turns out badly.

Figure 6.19: Secure data storage and transaction logs.

6.3.2.7.2 Development limits

One method for evaluating a proper most extreme exchange log measure is to set it at any rate the extent of the biggest list in the table. Since revamping the list requires in any event a similar measure of room in the exchange log as the file itself, the exchange log will dependably have enough space to finish the task.

6.3.2.7.3 Backup

At last, take steady and robotized reinforcements. There are numerous methods for setting this up (and is past the extent of this post). On the off-chance that the recuperation model is set to SIMPLE, the main alternative is to perform reinforcements of the information; backing up the exchange log is beyond the realm of imagination. Every reinforcement will be a solitary depiction of the information in the database at the season of the reinforcement.

In the event that the recuperation model is set to FULL, at that point of exchange log reinforcements enable the database to be completely reestablished to a specific point in time and aren't obliged to just a single explicit minute in time. Log reinforcements likewise play out the significant errand of stamping existing log records as dormant. Idle log record space would then be able to be reused by the exchange log when it returns to that area of the log document since it keeps in touch with the sign in a consecutive request. It's likewise essential to take note of that playing out a "full reinforcement" from the reinforcement alternatives does not back up the exchange log, notwithstanding when the database is in the FULL recuperation model.

6.3.2.8 NoSQL

The contemporary big data analysis is the problem of synchronization between database systems, which host the data and make available SQL querying, by means of data analytic [28] packages that carry out numerous forms of non-SQL processing, like data mining and statistical analysis.

Nonrelational data stores recognized as NoSQL databases are even under pressure concerning security infrastructure.

The basic difference between NoSQL and SQL maybe thought of the difference in nonrelational and relational structure of data. Data stored here in has NoSQL more like document layout, whereas SQL has more table setup. This provides NoSQL to be more flexible and trouble free to deal with the new data models than in SQL. NoSQL is a open-source software; it signifies the minimum cost associated with it and can act with low configuration hardware too. As a consequence, small software companies make use of it. Quick processing on big data is obtainable using NoSQL tools. Elastic scalability is also useable with this database. Here no database models are employed with NoSQL and thus a huge time is preserved. All the above points are just opposite to SQL (shown in Figure 6.20). Examples of NoSQL database tools such as MongoDB, CouchDB, CloudDB, and Bigtable are set up to use a decent variety of difficulties presented by the examination world and subsequently security was never part of the model at any time of its origination stage (Figure 6.21). Most engineers utilizing NoSQL databases by and large incorporate security in the middleware. No financing is offered by NoSQL databases for implementing it expressly in the database. Such security rehearses represent extra difficulties. As far as hotel and preparing the colossal majority of data, associations managing enormous unstructured informational collections may pick up the preferred standpoint by moving from a customary social database to a NoSQL database. Subsequently, the security of NoSQL databases depends on outside upholding systems.

Figure 6.20: Basic difference between NoSQL and SQL.

For instance, well-molded answers for NoSQL infusion are as yet not built up. Each NoSQL databases was in parliamentary law to drop down the security episodes. The

Figure 6.21: Security best practices for nonrelational data stores.

general public must continue through security arrangements for the middleware adding things to its motor and toughen NoSQL database itself to coordinate social databases without settling on its utilitarian qualities.

NoSQL database encryption solutions

Nonrelational or NoSQL databases are adaptable database arrangements that are composed for a lot of changed data types. Since they include more than customary database tables – utilizing objects and lists rather – they require an alternate way to deal with enormous data security.

Clients would now be able to ensure data in any NoSQL database, including driving database merchants, for example, MongoDB, Cassandra, Couchbase, and HBase.

Leverage file framework-level encryption answers for guaranteeing the documents, organizers, and offers that contain the files and articles listed in the NoSQL outline.

Coupled with strategy-based access controls, clients hold a fine purpose of control regardless of the gigantic information volumes.

Application-level big data encryption or tokenization arrangements append security legitimately to the data before it ever is spared into the NoSQL diagram.

Operations stay straightforward to the end-client while the database holds its capacity to direct inquiries, and convey data without reductions in execution.

6.3.2.9 Proper encryption technique

To ensure that the most touchy private data is completely secure and just open to the approved elements, data must be encoded depending on access control approaches. To guarantee confirmation, course of action, and decency among the dispersed substances, a correspondence system that is cryptographically verified must be completed. Delicate data is for the most part put away decoded in the cloud.

Lack of designing security measures is also creating security problems such as encryption, policy enablement, compliance, and risk management. If these

things are needed, they should be built on their own. Data masking policies and aggregating datasets may be used as security measures. Here re-identifying individuals are the proper tools that may perhaps locate datasets back simultaneously. Confining solitude might show the way to augmented safety measures menace, particularly if the data caught up be full of responsive and commercial information. Assortment of data is another important security measure, where data provider provides structured or unstructured data, but both are used by high-, middle-, or low-level users. This is the newest among all technologies using in today's computing world. It is very clear when any technologies are not well understood, certainly they become vulnerable.

The principle issue to scramble data is the win or bust recovery strategy of encoded data, which limits clients from effectively performing finely grained activities, for example, sharing records or ventures. Quality-based encryption (ABE) reduces this issue by using an open key cryptosystem where ascribes identified with the data scrambled serve to unscramble the keys. Then again, the decoded less touchy data helpful for investigation must be conveyed in a safe and settled upon way utilizing a cryptographically secure correspondence system.

Hadoop encryption solutions

The SafeNet data protection portfolio can secure data at different focuses in the Hadoop engineering – from Hive and HBase to singular nodes in the data spill with Gemalto, clients have a decision. Incorporate straightforward application-level security by means of APIs to ensure data without changing their database structure.

Select a column-level answer for Hive that licenses typical questioning. Obtain a file system-level arrangement with policy-based access controls. Every Hadoop big data encryption and tokenization arrangement is totally straightforward to the enduser and configuration saving encryption usefulness implies that customers will keep on benefitting from the examination instruments that suck up additional incentive from developing information stores.

Come up a center ground among obscurity and validation through a group signature, a cryptographic configuration where individuals can sign their data yet can just be named as individuals from a gathering. Just believed outsiders can perceive the substance (Figure 6.22).

6.3.2.10 Data mining

The use of big data can make out intrusion of security, obtrusive advertising, diminished common opportunities, what's more, increment state and corporate control. The special user data that appears really simple can be made usable when they make in use to predict something after analyzing it.

Figure 6.22: Transparent data encryption [29].

Anonymizing data for investigation isn't adequate to continue client privacy. For instance, Netflix (an American worldwide entertainment organization that blossoms with spilling media and video on interest on the web) confronted an issue when clients of their anonymized dataset collection were perceived by associating their Netflix motion picture scores with IMDB scores. Consequently, it is critical to decide rules and proposals for averting incidental security divulgences. User data marshaled by organizations and government experts are steadily mined and dissected by inside investigators and furthermore by outside contractual workers. A malevolent insider or unapproved accomplice can manhandle these datasets and get private information from clients. Also, knowledge organizations require the amassing of immense amounts of information. Hearty and adaptable, protection-saving calculations will expand the likelihood of gathering pertinent data to complement client security.

Solution

Data mining is a useful means to dispense with the vast amounts of big data, as important information can be pulled out, and then utilized to project future developments. The analysis of this data can help to solve problems and to shape strategies for predicting future trends (Figure 6.23).

A class of data mining is named as basket analysis. This analysis process reveals with reference to consumer buying options over a combination of different commodities at

Figure 6.23: Data mining – system architecture.

a time for specific occupations. The pattern matching ability of data mining provides retailers spending capability and increase out the products bought collectively. Mining gives the analyzed data to the marketing section of the business, based on those different items they advertise to the customer.

Another brand of data mining is sales forecasting; the operative procedure for this is very simple. It memorizes which product a consumer bought at what timing. This forecasting technique now determines when the purchaser failed to buy the same product again and once more. For instance, coffee retailing is a lot higher in winter than in summer, so retailers are acquainted with increased stock for the winter months. It can be useful for the retailer to keep funds in monthly basis when sales (maintaining inventories) budgeting plays a significant part in modern commercial enterprise.

Finally, data mining helps to predict the customer needs at times and build proper databases in a particular area with the quantity and quality of different products (Figure 6.24).

This enables retailers, dealers, distributors, as well as manufacturers to know close to consumer needs and also launching new products at times. The encryptions are necessary for data at rest, so that no external attack can harm data. Access control and authorization mechanisms are likewise required to go through. The severance of duty, ideology, and comprehensible policy for logging access to datasets are the big factors in data mining. For the awareness of re-identification issues, differential privacy should also be preserved.

Figure 6.24: Scalable and compostable privacy-preserving data mining and analytics.

6.3.2.11 Granular auditing

In real-time security settings, the system keeps an eye on each and every attack with proper notification to the administrator of the system. Theoretically, the system is 100% protected, but in reality, it may not happen every time. There are many things to settle on like missed hit auditing of information is very much essential. Periodic system audit information is required to analyze the system to enhance security. What information is to be checked, what to ignore, what and when went wrong at times with other processes involved in it are also obligatory. The acquiescence, regulations, and forensic exploration are significant. System auditing is mandatory in case of distributed data processing applications in big data applications as data volume is so big that a standalone system cannot process them properly. Distributed systems enable the computer hardware like gateway, router, server; client computer system of numerous types, thus auditing system is put into practice over these systems as well. The auditing system must involve with the software systems too, which includes application software enabled and also the system software (OS).

The data-level safekeeping assistance could be comprehended by means of the pattern given. To analyze database activities, a standard statement is produced by the sample organization. The database administrators (DBA) analyze and review activity reports of their databases. The data accumulation and filtration are different jobs; DBAs have the farm duties to sort out data that is not important to them.

Figure 6.25 describes two managers, namely DBA manager and audit director. The DBA manager for Oracle databases can only see the data that belongs to the same database. This procedure is the same for the DB2 and MySQL databases as well. The DBA manager has permission to access all the data stored in the different databases. There is an audit manager associated with the auditing section for in-house audits. The SOX and HIPAA are two different auditors work under the same

Figure 6.25: Granular audit process.

audit manager. The SOX auditor is accountable for the monetary data, i.e. sales, salaries, and orders, no matter where that data are stocked up. The HIPAA auditor is used in healthcare systems. It audits, data associated with patient information, no issue where that data is accumulated (Figure 6.26).

Figure 6.26: Granular audits.

Solution

Granular access audits in the majority of the cases can facilitate to come across the attack. Audit trials disclose the reason behind the nondetection of data at the beginning phase. The most significant mechanism of auditing may be talked about as the completeness of the required audit, well-timed approach to audit information, the reliability of the information, and endorsed entrance to the audit information.

Successful audit trial guarantees to incorporate the appropriate procedure and technologies in the big data infrastructure, including function logging, SIEM, forensic tools, and enabling SYSLOG on routers.

Split big data and audit data to differentiate among responsibilities. The audit data looks upon information concerning what has turned out in the big data infrastructure; however, it must be reserved taking, apart from the “regular” big data. A dissimilar network subdivision or cloud should be set up to host the audit system infrastructure.

6.3.2.12 Granular access control

The big data scheme acts as a significant task in data over the network of networks and storages. They carry out an excellent job regarding execution and versatility. But alas! It finds nearly no protection in mind. Existing RDBMS applications are secure enough with several protection attributes regarding access control, clients, tables, and rows and even at the cell level. However, a change of essential challenges put off big data solution to endow with comprehensive access control. The most important and foremost involvedness with course-grained get to components is that data that could somehow or another be shared is frequently cleared into an increasingly prohibitive gathering to ensure security.

Here granular access control is essential for diagnostic frameworks to adjust to this continuously progressive complex security environment. Keeping a log of jobs and experts of clients is one of the dangers alongside keeping up access marks crosswise over diagnostic changes. Big data examination and distributed (cloud) computing are progressively cantered around taking care of assorted data-set collections, both as far as assortment of diagrams and necessities. Legitimate and strategy confinements on data originate from different sources. Security arrangements, sharing understandings, and corporate strategy additionally forces prerequisites on data that dealt with. Dealing with this overabundance of the confinements has so far brought about expanded expenses for creating applications and a walled pationursery approach in which few individuals can take an interest in the analysis.

Solution

Granular access control as with any background is significant in a cloud environment to make certain that only the right people can have the right to use the right information and that the wrong people can't. The granular access control is vital to achieving this, and the cloud security alliance has acknowledged three specific tribulations in the realm of cloud data admission authority such as tracking confidentiality necessities that meant for individual essentials, supervising roles and authorities of users, and properly implementing secrecy requirements with MAC (Figure 6.27).

The resolution includes addressing these tribulations: associations may want to put into practice get to controls in the foundation layer, rearrange multifaceted nature in the application space, and receive confirmation and required access control. Nonetheless, scaling these technologies to the levels necessary to accommodate big data can present their own set of unique challenges. Use the typical single sign-on (SSO) method to condense the organizational work caught up in supporting a huge user base. SSOs transfer the trouble of user verification from administrators to publicly obtainable scheme.

As a conclusion, the data must pick the right dimension of granularity at the row level, column level, and cell levels. At the base, comply with cross section to get to impediments; progressively complex data changes are being considered in dynamic research. The authentication, authorization, and compulsory access control are mandatory.

Figure 6.27: Granular access control.

6.3.2.13 Data provenance

The information and process integrity are needed in big data. Thus, the term data provenance comes into play, which integrates them by reporting the entities, frameworks, and procedures working on and adding to data of intrigue. The lifetime of data and its starting place must confer as unchangeable chronological evidence. The large provenance graphs and diagrams are analyzed to perceive metadata dependence for protection, and discretion application is computationally exhaustive. The beginning of an application or the place (process) of creation must need to know to the assortments of key security applications. This source of data is important for companies based on financial trading. They need to know the origin, process, precision of data for further research on market trends and future forecasting. Data provenances endow with the safety measures and they are completed within the time frame and involve speedy algorithms to switch the provenance metadata restraining this information.

Big data provenance architecture may be partitioned into few parts as big data admittance, distributed big data platform, provenance, and application employing provenance. Every partition will get in touch with particular live data, then database engines calculate the provenance data from there. The above diagram describes a reference design; system developers force choice as well as make mind up on component in each sub-framework that is dependent on the focused provenance practice circumstances along with competence.

The subsystem called big data access describes the most proficient method to get the various sorts of big data for provenance and distributed systems. System developers are required to find out the finest method, or tool to access the detailed datasets. Further work is needed for experimentation requisite on the datasets. The synchronization involving data and computer system is extremely desirable. Based on this synchronization, the provenance following the execution is an important aspect.

Distributed big data sub-framework manages development and execution support for big data applications and capacity support for big data provenance over appropriated big data platform (Figure 6.28). Before executing with the definite big data engine, the application designers will choose one big data work process framework to construct their submission. Some systems similar to Spark are capable to take action as both workflow production tool and distributed data parallel (DDP) execution engine. By means of DDP programming models, a big data submission can be constructed by covering inheritance tools or straight programming. The provenance storage too requires selecting the appropriate (distributed) databases or file handling systems (Table 6.1).

The provenance sub-framework settles on which provenance information will be reported and how to record it. The provenance can be separated into three extents: data, lineage, and the environment. Revamping the definite condition of the investigation over these three measurements is basic to repeat any information-driven logical analysis. Data provenance catches the condition of the input halfway

Figure 6.28: Data provenance architecture [30].

and yield data at the season of the investigation. The fundamental pressure calculation is favored by this provenance relying on the data informational index. Lineage provenance arranges the computational bustle of the trial, which is caught by putting away the guidelines that actuated on these information datasets. System provenance unites data concerning the careful condition of the framework design, which assesses both equipment determinations and framework programming details (OS, libraries, outsider apparatuses, and so on).

Large prospects of big data provenance carry on how we could make use of it. Big data provenance can be worn for provenance question, test multiplication, provenance mining, try checking, information quality administration, analyze adaptation to noncritical failure, and numerous others as an uncommon sort of big

Table 6.1: Big data provenance work comparison [30].

	Provenance recording	Applicable big data engines	Provenance usage	Environment provenance recording
Kepler	Parallel recording in a MySQL database	Unmodified Hadoop	Parallel query through MySQL Cluster	N/A
RAMP	Parallel recording in files	Extended Hadoop	Backward provenance tracing	N/A
HadoopProv	Parallel recording in files	Modified Hadoop	Parallel query through index files	N/A
Pig Lipstick	Parallel recording in files	Unmodified Pig/Hadoop	Graph operation-based query	N/A

data. Each capability can be an individual application or an internal module in a superior scheme.

Big data workflows need to discover modeling and capturing provenance information by a small number of key learning. Here we make available a little outline of the efforts that attempt the challenges. Several new studies on big data provenance are not comprehensive in view of the fact that we spotlight on those that are in the circumstance of DDP also, and that have test analysis.

Kepler distributed provenance framework is done on big data provenance. It recommends a data model that can catch provenance contained by MapReduce occupations just as the provenance of non-MapReduce work process errands. It builds up the Kepler DDP engineering to record and inquire provenance in a disseminated manner on a MySQL cluster. It additionally offers an API to request the formed provenance. The WordCount application and a bioinformatics application called BLAST carry into play with the versatility of gathering and questioning provenance.

RAMP (reduce and map provenance) is an augmentation to Hadoop that underpins provenance catch and following for MapReduce work processes (). RAMP catches fine-grained provenance by wrapping Hadoop APIs. This constant wrapper-based methodology is perfectly clear to Hadoop and clients. RAMP implements various realities working expense all through provenance catch and empowers skillful in reverse following.

HadoopProv changes over Hadoop to incorporate provenance catch and investigation in MapReduce employments. The goal is to diminish provenance-limit expense. It follows to treat provenance in the map and reduces fragment freely. It likewise defers the structure of the provenance chart to the question stage by mixing transitional keys of the map and reduces provenance files.

Pig lipstick suggests a provenance system that combines database-style (fine-grained conditions) and work process style provenance (coarse-grained conditions)

Figure 6.29: Data provenance.

over Pig Latin. It advances a far-reaching and squashed diagram-based portrayal of fine-grained provenance for work processes that yield a more extravagant chart model than the OPM-Open Provenance Model standard utilized in work processes. It characterizes three graph revolution operations to ease the study of fiction workflow analysis queries.

Solution

Cloud computing involves big data applications within it; thus, provenance of metadata matures in volume. They tend to be huge and thus intricacy grows very high. There are three most important intimidation to protect provenance metadata in big data relevance such as faulty infrastructure mechanism, infrastructure external hit, and infrastructure in the interior hit.

Secure data provenance is necessary to solve these threats. The system needs to improve trust; the external attack should be condemned. The usability of secure provenance can be accomplished by protecting origin finding technique, and data admittance is tuned up to some degree. The system is based on clouds, and the infrastructure of the cloud setup solves the problem where collection and preventing outside attacks, pairing a fast, lightweight authentication technique with current provenance tools are present. The insider attacks must be prohibited, for that a dynamic, scalable access control product is essential. To maintain the employability, connection ability of provenance graphs fine-grained access control technology offers data access attribution in big data application.

Design goals must be identified in big data provenance. The layer-based architecture is followed, where the access control mechanism is needed to address security. It can hold equally the structured and unstructured types of data, but this can handle simple queries only. For handling complex queries we will add additional components.

Finally, as a conclusion, it can be assured that the authentication techniques, message digests, access control through systems, and cryptography are the supporting points.

6.3.2.14 Other problems

Cloud computing technology enables dynamic collection of colossal data from different ends or nodes. These types of data generated by different organizations include unstructured as well as semistructured data. These data are commonly referred to as big data. These data are needed for security from both ends (internal

as well as external), privacy for specific data. Cloud computing involves equally with affirmative and unenthusiastic belongings.

Fake data generation is another problem in the big data analysis, where security breaches made by cybercriminals fabricate data and store it in the same place as it was there. The big data technology used so far is implemented through open-source code that can be modified by an individual to have proper knowledge about it. The attack is done using any one of the user sites that cannot be recognized, checked, or imposed penalties in big data; at the same time, the server side is also unaware about what measures to be maintained.

The most astonishing thing about this methodology is that the phony data doesn't have to coordinate the qualities of genuine data precisely. That data-to-data coordinating is exceptionally difficult to incorporate with manufactured data, so it is uplifting news that it is commonly not required. Rather, you simply need to coordinate key performance indicators (KPIs) among phony and genuine data when utilized as contribution for the procedure of intrigue. Note that these KPIs are in respect to the particular issue and models you are working with. Android malware [31] sometimes cause problems in big data, and proper use of antivirus helps to remove them.

The volume of big data market increases exponentially where the growth of enterprise-level data is the fastest. It is relied upon to twofold at regular intervals, from 2,500 exabytes in 2012 to 40,000 exabytes in 2020. Figure 6.30 suggests that about big data market today (2018) is about \$24.3 billion and will be \$49.1 billion in 2020 [32]. These will not only tend to increase complexity but also the problem of accessing and storing security measures that should be improved up to some extent.

Till now the discussion is all about hardware, and software associated with it, but another part of big data security is left off. The data, software, and hardware everything maintained by human beings are sometimes taking the system unanimous. The devices logged in but not in use yield access to further workforce as well as the threat of admittance data passing through unrestricted Wi-Fi. Automatic user log table with actual work seconds should be taken for an internal purpose for individual systems, and proper training on usage of data is also be useful.

Another way out of the big data security problem is to use software tools that can analyze and monitor in real time so that the system can get the happenings instantly. Besides the actual hazard, the alarming system can find some artificial data works as threats called false alarm. It must have the capability to differentiate between false alarms and real threats. The tools produce a huge amount of network data; again, it is problematic to maintain this data. Now each organization may not afford the cost of tools as well as it needs to update the hardware system to execute 24×7 checking capability. Big data analytics itself can resolve for the said problem by enhancing the services of network protection. Log tables are maintained to take data about changes made on the network; they can also be used for anomalous network connections.

Numerous large organizations, which are exercised on big data, their leading apprehension are the security of the cloud-based systems. Malevolent assaults on

Figure 6.30: Year-wise big data market growth prediction.

IT systems are progressively thornier and an unmarked malware is being fashioned every now and then. Regrettably, endeavors that employed through big data come across such concern more or less on a daily basis. However, there is a clarification for each predicament and pronouncement one is certainly significant for the security apprehension.

The confrontation setup next to security, access control, compression, encryption, and compliance need to concentrate on in an organized manner as computing atmosphere turns out to be contemptible, and application atmosphere converted to the network along with system and analytics setting grow to be collectively more with the cloud. For the data section, further mechanism may possibly be added to grip composite queries.

Here, in the above section, we have marched from end to end and all the possible big data security challenges and have put down a number of suggestions intended for building big data processing and computation extra reliable and in turn making its infrastructure more secure. Some of the security issues discussed so far are very common in nature and are particular to big data crop up as of the numerous infrastructure tiers – (both computing and storage) employ in support of big data processing, the new-fangled computation infrastructures like NoSQL databases drawn on prompt throughput that are indispensable meant for huge dimensions of data are not comprehensively tenable as of most important security intimidation. The nonscalability of real-time scrutinizing practice, the diverse arrangement of devices that produce data, uncertainty by means of miscellaneous lawful limitations that come may lead to ad hoc approach for security and privacy.

For explicit big data tribulations, there must have a big ecosystem that exists. The subject matter here provides to illuminate unambiguous characteristic of the susceptible regions in the whole big data dispensation communications that require being investigated on behalf of definite intimidation.

The dispute with big data is that the unstructured character of the information creates it intricate to classify, model, and map the data as soon as it is confined to moreover accumulate. The dilemma is ended most horrible by the actuality that the data generally appear as of external sources, over and over again building it convoluted to substantiate its correctness.

Conflict resembling protection of data storage, data mining and analytics, transaction log and secure communication do subsist. The study on an assortment of safety measures faces up to roughly big data security and intact stack in large apparatus. Big data infrastructure security mechanism can be more stretched and modified to their directorial upbringing.

One of the modern approaches to secure big data is the use of blockchain. It could transform the technique we move toward big data. The quality of data would be improved and safekeeping on data has immediately two benefits to individuals and businesses as the blockchain is all around to grip information that can be digitized. The blockchains' prevalent improvement is its decentralization and thus no one owns the data entry or the integrity as it is established constantly by each computer on the network. Certainly, as much as necessary, blockchain and big data are a match made in heaven. The genuine issue at the present time is who will be the first to endow with the good number of appropriate and paramount skilled artificial intelligence/machine learning model working on top of distributed, transparent, and immutable blockchain generated data layers. The business to do this will roll in investments and engender enormous proceeds.

6.4 Conclusion

Big data analysis is flattering essential means for automatic determination of astuteness that is concerned in the recurrently stirring outline and secreted convention. This can facilitate companies to obtain an improved resolution, to envisage and recognize revolutionize and to categorize new fangled prospects. Dissimilar procedure in support of big data analysis as well as numerical analysis, batch processing, machine learning, data mining, intelligent investigation, cloud computing, quantum computing, and data stream preparing become possibly the most important factors. There is a gigantic open door for the big data industry in addition to plenty of possibility for research and enhancement.

Big data security solutions we have discussed so far may solve the current problem up to some extent. But it is to be said these are not the only solutions. They can

vary on a case-to-case basis, depending on the hardware devices or software applications used so far. The secure coding in a different language may fluctuate; encryption protocols are singular as per requirement. The layer of data may afford an unusual solution at times as authentication or authorization of data may be different. The proper analysis of static and streaming datasets in big data technology can assure to develop applications on medical and other scientific relevance and thus there can be a huge scope of business opportunities. The need for developing data-driven algorithms for creating applications on several other fields can open a new field of study.

Wish the problems and solutions given will be useful for the researchers and the big data fraternity at its best. The future research plan is to extend the other solutions possible as a new problem arises every corner of the world connected or to be connected.

Nomenclature

IoT	Internet of things
SaaS	Software as a service
Paas	Platform as a service
IaaS	Infrastructure as a service
NIST	National Institute of Standards and Technology
HDFS	Hadoop distributed file system
RDBMS	Relational database management system
MCC	Mobile cloud computing
QoS	Quality of service
SQL	Structured query language
IBM	International business machines
SIEM	security information and event management
DPF	Distributed programming frameworks
MAC	Mandatory access control
SSD	Solid-state drive
HDD	Hard disk drive
ABE	Attribute-based encryption
DBA	Database administrators
SSO	Single sign-on
DDP	Distributed data parallel
RAMP	Reduce and map provenance

References

- [1] A Cloud Security Alliance Collaborative research – Big Data working group. “Expanded Top Ten Big Data Security and Privacy challenges”, 2013.
- [2] Rountree, Derrick., & Castrillo, Ileana. (2014), “The Basics of Cloud Computing”, <https://doi.org/10.1016/B978-0-12-405932-0.00001-3>, ScienceDirect Elsevier Inc., pp 1-17

- [3] <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003>, Last Access: March (2019).
- [4] Koley, Santanu., Rashi, Agarwal., & Renu, Jain. "Cloud Computing: A New Era of Saving Precious Lives in the developing World". *Skit Research Journal*, 2015, 5(1), 2015(ISSN 2278-2508), 1–7.
- [5] Irion, K. Government Cloud Computing and National Data Sovereignty. *Policy and Internet*, 2012, 4(3–4), 40–71. DOI: 10.1002/poi3.10.
- [6] Zhou, Wenhao., Xiao, Yongbing., & Shen, Yulan. (2017) "Application of Cloud Computing in Telecom Operators" Creative Commons Attribution-NonCommercial 4.0 International License Computer System Networking and Telecommunications, WHIOCE Publishing Pte. Ltd. (<http://creativecommons.org/licenses/by-nc/4.0/>).
- [7] <https://www.economist.com/special-report/2010/02/25/data-data-everywhere.>, Last Access: May (2019).
- [8] Fung, PoTso., SimonJouet, Dimitrios P., & Pezaros, (2016) "Network and server resource management strategies for data centre infrastructures: A survey" Vol. 106, pp. 209–225.
- [9] Bohn, Robert. "Cloud Computing – A NIST Perspective and Beyond" Advanced Network Technologies Division, National Institute of Standard and Technology, U.S. Department of Commerce, 2016. (<https://www.nitrd.gov/nitrdgroups/images/8/82/NISTCloudComputing.pdf>).
- [10] Koley, Santanu., & Ghosh, Shivnath. (2015) "A Life Saving Cloud Computing Architecture for Mobility of Sufferings", 9th INDIACom; INDIACom-2015; IEEE Conference ID: 35071, 2015 2nd International Conference on "Computing for Sustainable Global Development", 11th–13th March, 2015, ISSN 0973-7529; ISBN 978-93-80544-14-4, pp. 5.239–5.244.
- [11] Akherfi, Khadija., Gerndt, Micheal., & Harroud, Hamid. "Mobile cloud computing for computation offloading: Issues and challenges." *Applied Computing and Informatics*, 2016, 14(1), 1–16.
- [12] Wang, Yating., & Ing-Ray Chen, Ding-Chau. "A Survey of Mobile Cloud Computing Applications: Perspectives and Challenges". *Wireless Personal Communications: An International Journal* archive, 2015, 80(4), 1607–1623.
- [13] Benslimane, Younes., Plaisent, Michel., Bernard, Prosper., & Bahli, Bouchaib. "Key Challenges and Opportunities in Cloud Computing and Implications on Service Requirements: Evidence from a Systematic Literature Review", *CLOUDCOM '14 Proceedings of the 2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, IEEE Computer Society, Washington, DC, USA, 2014. ©2014, ISBN: 978-1-4799-4093-6 doi>10.1109/CloudCom.2014.115, pp. 114–121.
- [14] Koley, Santanu., & Ghosh, Shivnath. (2015) "Cloud Computing with CDroid OS based on fujitsu Server for Mobile Technology", *Bilingual International Conference on Information Technology: Yesterday, Today, and Tomorrow*, 19–21 February 2015, pp. 163–169.
- [15] Koley, Santanu., & Singh, Navjot. "Cdroid: Used In Fujitsu Server For Mobile Cloud". *Ge-International Journal of Engineering Research*, 2014, 2(7), ISSN: (2321-1717), pp. 1–6.
- [16] Koley, Santanu., & Ghosh, Shivnath. "Cloud Computing with CDroid OS based on Fujitsu Server for Mobile Technology". *Skit Research Journal*, 2014, 4(2), 2014(ISSN 2278-2508).
- [17] <https://moreign-technologies.co.za/2019/03/02/markets-trends-of-big-data-usage-in-health-care-during-the-next-5-years/>, *Markets Trends Of Big Data Usage In Health Care During The Next 5 Years*, Last Access: May (2019).
- [18] Reichman, O.J., Jones, M.B., & Schildhauer, M.P. "Challenges and Opportunities of Open Data in Ecology". *Science*, 2011, 331(6018), 703–5. Bibcode:2011Sci ... 331.703R. doi:10.1126/science.1197962. PMID 21311007.
- [19] Gaskell, Jill., Kersten, Phil., Larrondo, Luis F., Canessa, Paulo., Martinez, Diego., Hibbett, David., Schmoll, Monika., Kubicek, Christian P., Martinez, Angel T., Yadavi, Jagjit., Master,

- Emma., Magnusonk, Jon Karl., Yaver, Debbie., Berka, Randy., Lail, Kathleen., Chen, Cindy., LaButti, Kurt., Nolan, Matt., Lipzen, Anna., Aerts, Andrea., Riley, Robert., Barry, Kerrie., Henrissat, Bernard., Blanchette, Robert., Grigoriev, Igor V., & Cullen, Dan. "Draft genome sequence of a monokaryotic model brown-rot fungus *Postia* (*Rhodonia*) *placenta* SB12". *Genomics Data-ELSEVIER*, 2017, 14, 21–23.
- [20] Demchenko, Y., Ngo, C., Laat, C. de., Membrey, P., & Gordijenko, D. (2014). "Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure" In W. Jonker & M. Petković (Eds.), *Secure Data Management*, Springer International Publishing. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-06811-4_13, pp. 76–94.
- [21] Xueheng, Hu., Meng, Lei., & Aaron, D. Striegel.(2014) "Evaluating the Raw Potential for Device-to-Device Caching via Co-location" <https://doi.org/10.1016/j.procs.2014.07.042> ELSEVIER Procedia Computer Science, Vol 34, pp. 376–383.
- [22] Wang, Jianwu., Crawl, Daniel., Purawat, Shweta., Nguyen, Mai., & Altintas, Ilkay. (2015) "Big Data Provenance: Challenges, State of the Art and Opportunities" INSPEC Accession Number: 15679551, DOI: 10.1109/BigData.2015.7364047
- [23] Terzi, Duygu Sinanc., Demirezen, Umut., & Sagioglu, Seref. "Evaluations of big data processing." *Services Transactions on Big Data* (ISSN 2326-442X), 2016, 3, 1, 44–53.
- [24] Geist, A. et al. (1996) "MPI-2: Extending the message-passing interface" In: L. Bougé, P. Fraigniaud, A. Mignotte, & Y. Robert (eds) *Euro-Par'96 Parallel Processing*. Euro-Par 1996. *Lecture Notes in Computer Science*, Vol. 1123, DOI https://doi.org/10.1007/3-540-61626-8_16, Online ISBN 978-3-540-70633-5. Springer, Berlin, Heidelberg.
- [25] Kailai, Xu. (2017) <http://stanford.edu/~kailaix/files/MPI.pdf>, Last Access: May (2019).
- [26] Alfredo Cuzzocrea, C. "Privacy and Security of Big Data: Current Challenges and Future Research Perspectives". *ACM Digital Library*, 2014, ISBN: 978-1-4503-1583-8 doi>10.1145/2663715.2669614, pp. 45–47.
- [27] Boyd, Dana., & Crawford, Kate. (2011) "Six Provocations for Big Data " *Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*. doi:10.2139/ssrn.1926431.
- [28] Cuzzocrea, A., Song, I., & Davis, K.C. (2011) "Analytics over Large-Scale Multidimensional Data: The Big Data Revolution!" In: *Proceedings of the ACM International Workshop on Data Warehousing and OLAP*, pp. 101–104.
- [29] Adhikari, M., Koley, S., & Arab, J. *Sci Eng* <https://doi.org/10.1007/s13369-017-2739-0> (2017), "Cloud Computing: A Multi-workflow Scheduling Algorithm with Dynamic Reusability". *Arabian Journal for Science and Engineering* (Springer), Journal, 2017, 13369, Article: 2739.
- [30] Koley, S., Nandy, S., Dutta, P., Dhar, S., & Sur, T. (2016), "Big Data architecture with mobile cloud in CDroid operating system for storing huge data," 2016 International Conference on Computing, Analytics and Security Trends (CAST), Pune, India, doi: 10.1109/CAST.2016.7914932, pp. 12–17.
- [31] Bacci, Alessandro., Bartoli, Alberto., Martinelli, Fabio., Medvet, Eric., Mercaldo, Francesco., & Visaggio, Corrado Aaron. "Impact of Code Obfuscation on Android Malware Detection based on Static and Dynamic Analysis" 4rd International Conference on Information Systems Security and Privacy, Funchal, Portugal, 2018.
- [32] Transparent data encryption or always encrypted?, Source: <https://azure.microsoft.com/en-us/blog/transparent-data-encryption-or-always-encrypted/>, Last Access: March (2019).