Journal of Engineering and Applied Sciences 14 (22): 8266-8272, 2019
ISSN: 1816-949X

A Study on Security Approaches for Big Data Hadoop Distributed File System

Leelavathi and M. Elshayeb
SEGi University, Kota Damansara, Selangor, Malaysia

Abstract: Big data refers to large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies in order to analyse complex data and to identify patterns it is very important to securely store, manage and share large amounts of complex data. In recent years an increasing of database size, according to the various forms (text, images and videos) in huge volumes and with high velocity, problems of services that use internet and require big data come to forefront (data-intensive services). Hadoop Distributed File System (HDFS) is evolving as a superior software component for cloud computing combined along with integrated parts such as MapReduce. Hadoop which is an open-source implementation of Google MapReduce including a distributed file system, provides to the application programmer the abstraction of the map and the reduce. The research shows the security approaches for big data Hadoop distributed file system and the best security solution, also this research will help business by big data visualization which will help in better data analysis. In today's data-centric world, big-data processing and analytics have become critical to most enterprise and government applications.

Key words: Technologies, information, management, large amounts, dentify, complex data

INTRODUCTION

The security of sharing, managing and storing huge amount of complicated data is extremely important in turn to recognize patterns and analyse complicated data (Inukollu *et al*., 2014). Nowaday, the database capacity rising, according to the different kind of forms such as (images, videos and text), due to the large volume and huge velocity, the services issues that use web and needs big data come to forefront (data-intensive services). For companies like Amazon, Facebook and Google the web has emerged as a large, distributed data repository which is processing by traditional database management systems appears to be insufficient (Pokorny, 2013).

Big data base is a data that contain a very high amount of tuples (data rows) or employ a huge physical files system storage space. The majority definition of big data is a database that contains more than one terabyte or occupies number of billion rows, although, over time, naturally this definition changes, nowaday many of organizations have a huge database, it is the organization's information obtained and treated through new techniques to get the best value in perfect way.

Big data point to a huge amount of information management and analysis technologies that go over the proficiency of traditional data processing technologies. Big data have three different ways than traditional technologies: the number of data (volume), the speed of data transference and generations (velocity) and the types of structured and unstructured data (variety) (Laney, 2001). Big data technological advances in analysis, storage and processing contain the cost of CPU power and storage in last year's decreasing very fast, the cost effectiveness and flexibility for storage and elastic computation in cloud computing and data centers and (the new frameworks development such as no SQL and Hadoop which give advantage for users in these distributed computing systems saving huge quantities of data through adaptable parallel processing. Several changes have produced by these advances between big data analytics and traditional analytics.

Apache's Hadoop Distributed File System (HDFS) is in progress as outstanding software component for cloud computing joint with integrated pieces such as MapReduce. Google MapReduce implemented an open source which is Hadoop having a distributed file system, present to software programmers the perception of the map and the reduce (Inukollu *et al*., 2014). Hadoop (Highly Archived Distributed Object Oriented Programming) was produced by Goug Cutting and Mike Cafarella in 2005 for helping a distributed search engine project. It is an open source Java framework technology supports saving, access and getting huge resources from big data in distributed form at extreme degree of fault tolerance, huge scalability and lower cost (Saraladevi *et al*., 2015).

MapReduce is a software designed for generating and processing huge data sets (Dean and Ghemawat,

Corresponding Author: Leelavathi, SEGi University, Kota Damansara, Selangor, Malaysia

2004). Google has introduced MapReduce in 2004, nowaday it is the programming design most selected for generating huge data sets. Programmer explained both of map function and reduce function as maps a data set into different data set and a reduce function that gathers intermediate outcomes into a final result.

Data visualization is a critical part in utilizing big data to get a full view of customers. In a lot of big data scenarios relationships are valuable aspect. Social network maybe the most conspicuous example and are extremely hard to understand in text or other formats, although, visualization can help make emerging network trends and patterns obvious (Wang *et al*., 2015). When visualization takes the correct place in the big data technology, then, we will be able to move forward with concept by using more technological tools to collect more information from charts and graphs as a result the data that being seen and how it is getting processed will be changed to better way (Ajibade and Adediran, 2016).

Problem statement: Big data having many problems leakage, some of the problems are in processing, security, management and storage problem in big data each issue has its own task of surviving (Ji *et al*., 2012). By taking a deep look in security problem, to manage a huge data set in safe manner and inefficient tool there are some challenges, unexpected leakage, volunteered and more threats of data appears in private and public database and the insufficiency of private and public policy making the database easier for hackers. When data moves from homogeneous data to the heterogeneous the security to face untrusted people is highly complicated, many applications and technologies for big data set is not often designed and developed with more policy and security certificates (Saraladevi *et al*., 2015). The remainder of this research is organized as follows. Section II present the importance of re-estimated tracking and recognition system during presentation of big data visualization, the difficulty ok big data analysis shows an undeniable challenge, methods and techniques for big data visualization need enhancements. Nowaday open-sources projects and companies notice the future of big data analysis via. visualization (Husain *et al*., 2015), a quick decision in employment is needed for big data characteristics as the information of data can lose of importance and become less up to date fast (Turner *et al*., 2014), data size have increased exponentially and by 2020 the amount of digital bits will be comparable to the number of stars in the world. As the amount of bits grow every 2 years for the period from 2013-2020 universe data will increase from 4.4-44 zettabytes. The huge data expansion may lead to difficulty related to human ability in dealing with the data, gain knowledge and gather information from it (Olshannikova *et al*., 2015). This study provides information and enhancement about re-estimated tracking and recognition system of big data visualization aims to avoid mismatch of the real view scene and computer generated objects.

Significance of the study: This study will help organizations in understanding the security approaches for big data Hadoop distributed file system, also this research will help business by big data visualization which will help in better data analysis. Nowaday big data analytic and processing have become very important to many companies and government applications. Thus, a well and successfully big data security is needed for defending the storage and operating on huge scale. Currently, MapReduce is regularly used for operating such big data (Zhao *et al*., 2014). MapReduce implemented one of the best known apaches which is Hadoop and has been extended/used by scientists as the base of their own research work (Shan *et al*., 2010).

The second section of this research will help to understanding visualization through giving information and enhancement of tracking and recognition system for big data visualization, to re-estimate data during presentation of data and to help in identifying the best visual presenting way for re-estimated data which help decision makers derive more value from big data. The main goal of data visualization is to associate with information purely and proficiently through plots information graphics and statistical graphics. Data operating ways embrace different disciplines including computer science, economics, applied mathematics and statistics. Those are the roots for data analysis techniques such as data mining, neural networks, machine learning, signal processing and visualization methods (Shneiderman, 2009).

Literature review: At the time of writing, the term 'big data' is closely everywhere inside reports and articles created by information technology experts and researchers. The broad scale of data-reliant tools and the omnipresent nature of digital technologies have also, made the term far reach everywhere within other disciplines including biology, management, medicine information science, sociology and economics. In spite of this, there are some challenges for handling a huge data set in safe and secure way and ineffective applications, big data needs a highly and fast deployment of infrastructure which will help in operating and saving big data in the computing environment.

The Hadoop Distributed File System (HDFS) is a distributed file system developed to work on commodity

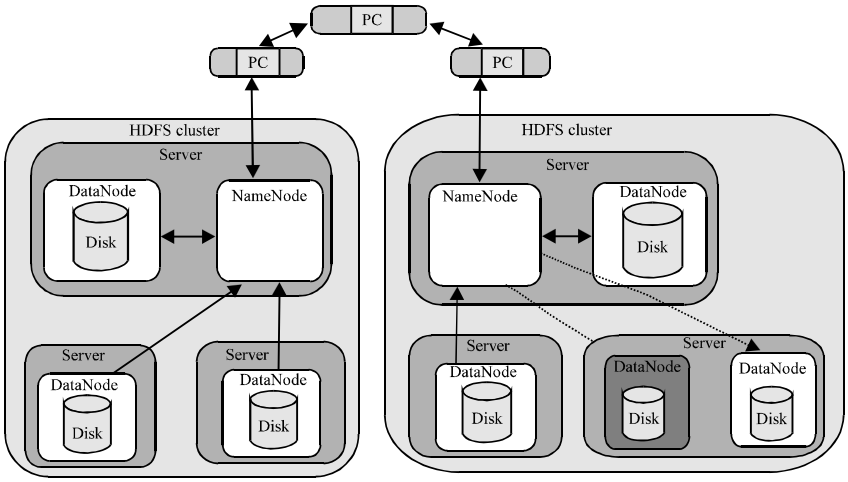

Fig. 1: The HDFS architecture

hardware. HDFS similar to many other available distributed file systems. However, there are expressive changes from other distributed file systems. HDFS contain NameNode where metadata saved on a dedicated server and DataNodes are other servers saves the application data on it. TCP-based protocols are used to totally communicate and connect between all servers with each other (Tantisiriroj *et al*., 2008).

HDFS architecture: An HDFS cluster contains a single node called NameNode, it works on regulates users access to data and controlling the file system namespace. On the other hand, DataNodes works on saving data as blocks inside files (Hanson, 2011).

As Fig. 1, illustrates, each cluster having one NameNode. This development facilitates a clear model for controlling each namespace and arbitrating data distribution. NameNodes and DataNodes are tools components developed to work in a decoupled manner on commodity machines across heterogeneous operating systems (Hanson, 2011).

NameNode: The HDFS namespace is a hierarchy of directories and files. Directories and files are showed on the NameNode by inodes which store attributes as modification and access times, disk space quotas, namespace and permissions. The file content is divided into big blocks and each block is individually replicated at twice DataNodes, the NameNode working on fixing and arranging the mapping of file blocks to DataNodes and the namespace tree. In the process of writing data, the user askes the NameNode to submit a suite of three DataNodes to host the block replicas. The user then stores data to the DataNodes in a pipeline fashion. In the current design for each cluster it has a single NameNode. The cluster can contain tens of thousands of HDFS clients per cluster and thousands of DataNodes. HDFS keeps the entire namespace in RAM (Shvachko *et al*., 2010).

DataNodes: Each block replica on a DataNodes is represented by two files in the local host's native file system. The first file holding the data and the second file is block's metadata including checksums for the block data and the block's generation stamp. A DataNode identifies block replicas in its possession to the NameNode by sending a block report. The generation stamp, the block id and the length for each block replica the server hosts are inside of the block report. Every hour the NameNode get information from a subsequent block report about the view of where block replicas are located on the cluster (Shvachko *et al*., 2010).

HDFS client: The code library that exports the HDFS file system interface, user applications access the file system using the HDFS client (Shvachko *et al*., 2010).

- Checkpoint node
- Backup node

HDFS security issues: In Hadoop architecture the base layer is the HDFS which is highly sensitive to security issue as it contains different classifications of data. Also, the risk of data access when data inserted in one Hadoop environment it is being easily for unwanted disclosure and theft to take place. The replicated data is also not secure which needs more security for protecting from vulnerabilities and breaches. Because of the low security

level inside a Hadoop technology most of the organizations and governments field's never using Hadoop environment for saving important data. They getting security help in outside of Hadoop environment like intrusion detection system and firewalls. The HDFS is represented by some authors in the Hadoop environment is providing security to protect from vulnerabilities and theft only by using encryption techniques to encrypting the nodes and blocks and other encrypted block levels and file system but till now there is no perfect algorithm known to maintain the security in Hadoop environment. In order to increase the security some approaches are mentioned as Saraladevi *et al*. (2015).

Kerberos mechanism: A network authentication protocol that grant the node to transfer files through a non-secured channel by a ticket which is a tool used to prove the special identification between the nodes is called Kerberos. It is a procedure that is used to improve the HDFS security. The remote procedure call is used to attain a connection between the client and NameNode. The block transfer is used to attain a connection from the client to the data node. In this case the Kerberos authenticates a RPC connection (Al-Janabi and Rasheed, 2011). The client makes use of the Kerberos authenticated connection, if he needs to acquire token means. Kerberos can be used to authenticate a NameNode by ticket granting ticket or service ticket. After long running of jobs both ticket granting ticket and service ticket can be renewed while Kerberos is renewed. New ticket granting ticket and service ticket are as well supplied and provided to all task. After receiving a request from task and network traffic is evaded the key distribution centre issues the kerberos service ticket using ticket granting ticket by using tokens. The ticket remains constant and only the time period is extended in the NameNode. One of the greatest advantages is that the token cannot be renewed by any attacker, if stolen. To provide security for file access in HDFS, other methods can be used. To identify which data node hold the files of the block, the data node has to contact the NameNode as it only authorizes access to file permission and a block token is issued where the data node authenticates the token. This token allows the data note to identify the authorization status of the client to the data to be accessed. A name token is issued by the data node to authorise it to enforce permissions for correct control access on its data blocks. Both tokens are sent back to the client with the data block locations and that they're authorized person to access it. These methods increase security as they help prevent unauthorised access from clients (Saraladevi *et al*., 2015).

Walled garden: The approach that is mostly used nowdays is called 'Walled garden' security model. It is quiet similar to the 'moat' model from mainframe security. This is a place where the cluster entirely on its own network, firewalls or API gateways tightly control logical access, access controls for user or application authentication usage. When put in use, virtually the model provides no security in the Hadoop cluster. Data and infrastructure security depend on the 'protective shell' of a network or application that surround it. The simplicity of the model is its greatest advantage. It helps all types of firms to implement the model with the tools and skills that already exist without the performance or Hadoop cluster functional degradation. The disadvantage of this model is that once the firewall or application failure occurs the system itself is exposed to the public. Moreover, it doesn't prevent authorized users from misusing the system or even modifying the data stored in the cluster. It is mostly cost effective and simple to businesses that do not worry about security (Securosis, 2016).

Data visualization: In talking about the data visualization, then, the meaning is the progress of data presented in pictorial layout or in a graphical design. The main advantage of data visualization is to help organization and managements who are responsible for taking decision to easily view data analytics presented in visual way for better understanding the complex ideas and identify the new patterns. After the visualization become more cooperating, then we need to updates and upgrade the visual concept by more technological application to collect more information from the charts and graphs, therefore, it leads to better changes the data being viewed and how it processed (Ajibade and Adediran, 2016).

MATERIALS AND METHODS

Data visualization

Line chart: A line chart displays the relationship between each variable on the chart. Line charts are frequently used to make comparison between lots of items at the same time. Take, for example, there are 12 data points to plot or show, the best way to make those points understandable is to just display them in an order using a table (SAS., 2014). The fact that one has some data points to plot or display does not mean that line graph is the best to pick but you should consider the number of data points that you want to display which will tell the best visual method to pick. Data points are mostly being connected by a straight line and line chart is actually an extension of Scatter plot. Some specific symbols and icons are being used to represent data points in a line chart (Ajibade and Adediran, 2016).

Pie chart: It is as well-known as a circle graph. A pie chart shows information statistics and data in a way that is not difficult to read called "pie-slice" form and the various sizes of slice shows how much of an element is in existence. When the slice is big, then it shows of the data was gathered. It is also used to compare values of data and the moment some values are represented on pie chart, then you will be able to view which of the items is the least popular or which is more popular (Cardenas *et al*., 2013). By providing additional information, report consumers do not have to guess the meaning and value of each slice. If you choose to use a pie chart, the slices should be a percentage of the whole (SAS., 2014).

Tree map: A tree map is a visualizing technique that has the attribute of showing data in hierarchy in a nested or layered rectangle form (Shneiderman, 2009). It is a very effective technique that is used to visualize structures of hierarchies. User are able to compare nodes and sub nodes at different depth and also, they are able to identify expected results and patterns. A lot of data set have the hierarchy characteristics and the objects are thereby divided into different divisions, sub divisions, etc.

RESULTS AND DISCUSSION

Activity detection: Activity detection is the process of searching for the existing instances of an activity in time-varying data sets. While there is a slight difference between activity detection and activity recognition, fundamentally, it is assumed that they both serve the same purpose: extracting the instances of a certain activity or a set of activities. The subtle difference between activity detection and activity recognition lies in the definition of activity. If there are multiple activities are defined in the data, then, the action of gathering and naming the instances of any of these activities in the data is defined as activity recognition. If there is only a single activity to detect (or if the purpose is finding "any" activity in the data regardless of its label), then, the action of gathering the instances of the activity is simply an activity detection process. Note that the terms "activity detection", "activity recognition", "action detection" and "action recognition" have also been used interchangeably (Turaga *et al*., 2008).

Tracking algorithm: A feature tracking method that solves the correspondence problem based on primary attributes of features such as position, size and mass. The key of the algorithm is the use of a prediction scheme and the use of a multi-pass search for continuing paths. The process is highly interactive, the scientist can guide the tracking process by changing criteria and parameters, resulting in different tracking solutions. This tracking algorithm is based on a simple assumption: features evolve consistently, i.e., their behavior is predictable. This implies that once a path of an object is found, it can make a prediction to the next frame and search for features in that frame that correspond to the prediction. A prediction can be made for the next frame at the end of the path but also for the preceding frame at the beginning of the path. This means it can search forward and backward in time (Reinders *et al*., 2001). The track updating process typically begins with a procedure that is used to choose the best observation to track association. This procedure is known as data correlation and is conventionally comprised of two steps called gating and association (Konstantinova *et al*., 2003).

GNN algorithm description:

- Receiving data for current scan
- Clusterisation-measurements to tracks allocation

At the beginning all tracks are clusters. In two nested cycles for all tracks and for all measurements using gating criterion it is defined, if some measurement falls in the gate of the given track. When two tracks have common measurement in their gates their clusters are merged in supercluster.

For each cluster

Measurements to tracks association: At this stage, the elements of the cost matrix for the assignment of the measurements to tracks in the current cluster is defined by equation. Solve assignment problem using Munkres algorithm.

Track filtering: Taking from the Munkres solution the associated measurement for each track state update is performed using extended Kalman filter in the frame of Interacting Multiple Model (IMM) approach.

Track initiation: Measurements which are not associated with existing tracks, generate new tracks (Konstantinova *et al*., 2003).

CONCLUSION

This study shows the big data information and characteristics used in world wide. The issues are also, mentioned to give idea about the big data issues in real time. The security issue is pointed more in order to increase the security in big data. The security can be improved by identifying the best security approach or by

combining the right approaches together in Hadoop distributed file system which is the base layer in Hadoop. There are a lot of challenges for big data processing and analysis. As all the data is currently visualized by computers, it leads to difficulties in the extraction of data, this study, also, obtained relevant big data visualization re-estimated tracking and recognition methods, the study concentrate on visualization techniques in order to get maximum benefits offer by visualization. Another important aspect is to clearly specify what data type needs to represent which visualization method that interpret maximum understandings, keeping in mind the significance of visualization.

REFERENCES

Ajibade, S.S. and A. Adediran, 2016. An overview of big data visualization techniques in data mining. Intl. J. Comput. Sci. Inf. Technol. Res., 4: 105-113.

Al-Janabi, S.T.F. and M.A.S. Rasheed, 2011. Public-key cryptography enabled kerberos authentication. Proceedings of the IEEE Conference on Developments in E-Systems Engineering (DeSE), December 6-8, 2011, IEEE, Dubai, UAE., ISBN:978-1-4577-2186-1, pp: 209-214.

Cardenas, A.A., P.K. Manadhata and S. Rajan, 2013. Big data analytics for security intelligence. Master's Thesis, Cloud Security Alliance, Stamford, Connecticut.

Dean, J. and S. Ghemawat, 2004. MapReduce: Simplified data processing on large clusters. Proceedings of the 6th International Symposium on Opearting Systems Design Implementation Vol. 6, December 06-08, 2004, San Francisco, California, USA., pp: 10-10.

Hanson, J.J., 2011. An introduction to the Hadoop distributed file system. IBM Computer Hardware Company, Armonk, New York, USA. https://www.ibm.com/developerworks/library/wa-introhdfs/

Husain, S., A. Kalinin, A. Truong and I.D. Dinov, 2015. SOCR data dashboard: An integrated big data archive mashing medicare, labor, census and econometric information. J. Big Data, 2: 1-18.

Inukollu, V.N., S. Arsi and S.R. Ravuri, 2014. Security issues associated with big data in cloud computing. Intl. J. Network Secur. Appl., 6: 45-56.

Ji, C., Y. Li, W. Qiu, U. Awada and K. Li, 2012. Big data processing in cloud computing environments. Proceedings of the 12th International Symposium on Pervasive Systems, Algorithms and Networks (ISPAN), December 13-15, 2012, IEEE, San Marcos, Texas, USA., ISBN:978-1-4673-5064-8, pp: 17-23.

Konstantinova, P., A. Udvarev and T. Semerdjiev, 2003. A study of a target tracking algorithm using global nearest neighbor approach. Proceedings of the International Conference on Computer Systems and Technologies (CompSysTech'03), June 19-20, 2003, Rousse, Bulgaria, pp: 290-295.

Laney, D., 2001. 3D data management: Controlling data volume, velocity and variety. META Group, Stamford, Connecticut.

Olshannikova, E., A. Ometov, Y. Koucheryavy and T. Olsson, 2015. Visualizing big data with augmented and virtual reality: Challenges and research agenda. J. Big Data, 2: 1-27.

Pokorny, J., 2013. NoSQL databases: A step to database scalability in web environment. Intl. J. Web Inf. Syst., 9: 69-82.

Reinders, F., F.H. Post and H.J. Spoelder, 2001. Visualization of time-dependent data with feature tracking and event detection. Visual Comput., 17: 55-71.

SAS., 2014. Data visualization techniques: From basics to big data with SAS® visual analytics. SAS Institute Software Company, Cary, North Carolina, USA.

Saraladevi, B., N. Pazhaniraja, P.V. Paul, M.S. Basha and P. Dhavachelvan, 2015. Big data and hadoop-a study in security perspective. Procedia Comput. Sci., 50: 596-601.

Securosis, 2016. Securing hadoop: Security recommendations for hadoop environments. Securosis, Arizona, USA.

Shan, Y., B. Wang, J. Yan, Y. Wang and N. Xu *et al*., 2010. FPMR: MapReduce framework on FPGA. Proceedings of the 18th Annual ACM/SIGDA International Symposium on Field Programmable Gate Arrays (FPGA '10), February 21-23, 2010, ACM, Monterey, California, USA., ISBN:978-1-60558-911-4, pp: 93-102.

Shneiderman, B., 2009. Treemaps for space-constrained visualization of hierarchies. Master's Thesis, University of Maryland, College Park, Maryland, USA.

Shvachko, K., H. Kuang, S. Radia and R. Chansler, 2010. The hadoop distributed file system. Proceedings of the 26th IEEE Symposium on Mass Storage Systems and Technologies, May 3-7, 2010, Incline Village, NV., pp: 1-10.

Tantisiriroj, W., S. Patil and G. Gibson, 2008. Data-intensive file systems for internet services: A rose by any other name. MCS Thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Turaga, P., R. Chellappa, V.S. Subrahmanian and O. Udrea, 2008. Machine recognition of human activities: A survey. IEEE Trans. Circuits Syst. Video Technol., 18: 1473-1488.

Turner, V., J.F. Gantz, D. Reinsel and S. Minton, 2014. The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. International Data Corporation, Framingham, Massachusetts, USA.,.

Wang, L., G. Wang and C.A. Alexander, 2015. Big data and visualization: Methods, challenges and technology progress. Digit. Technol., 1: 33-38.

Zhao, J., L. Wang, J. Tao, J. Chen and W. Sun *et al*., 2014. A security framework in g-hadoop for big data computing across distributed cloud data centres. J. Comput. Syst. Sci., 80: 994-1007.