HDFS SECURITY APPROACHES AND VISUALIZATION TRACKING

M.ELSHAYEB[1], Dr. LEELAVATHI[1]
School of Information Technology, SEGi University, PJU5, Kota Damansara No. 9 Jalan Teknologi, 47810, Petaling Jaya, Selangor DE, Malaysia

leelavathiraj@segi.edu.my; mazen.elshayeb@gmail.com

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, the services issues that use internet and desires big data come to leading edge (data-intensive services), (HDFS) Apache's Hadoop distributed file system is in progress as outstanding software component for cloud computing joint with integrated pieces such as MapReduce. GoogleMapReduce implemented an open source which is Hadoop, having a distributed file system, present to software programmers the perception of the map and 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.

1. 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 [2]. Now a day 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 internet and desires big data come to leading edge (data-intensive services). For organizations like Amazon, Facebook and Google the internet has emerged as a large, distributed data repository, which is processing by traditional Database Management systems appears to be insufficient [1]. Big data

base, is a data that contain a very high amount of tuples (data rows), or employ a huge physical filesystem 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, now a day 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) [4]. Big Data technological advances in analysis, storage and processing contain (i) the cost of CPU power and storage in last year's decreasing very fast; (ii) the cost effectiveness and flexibility for storage and elastic computation in cloud computing and datacenters; and (iii) 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 [5].

(HDFS) Apache's Hadoop distributed file system 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 [2]. Hadoop (Highly Archived Distributed Object Oriented Programming) was produced in 2005 for assisting projects of distributed search engine by M. Cafarella and G. Cutting. It is a java framework which is an open source technology supports saving, access and getting huge resources in distributed form of Big Data at extreme degree of reducing and controlling faults, huge scalability and lower cost [3].

MapReduce is a software designed for generating and processing huge data sets [13].Google has introduced MapReduce in 2004, now a day 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 results.

Data visualization is assume 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 are 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 [10]. 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 [11]

1.1 Problem Statement

Big Data having many problems leakage, some of the problems are in processing, security, management and storage problem, in Big Data every problem has its personal undertaking of surviving [6]. By taking a deep look in security problem, to manage a huge data set in inefficient tool and safe manner there are some challenges, unexpected leakage, volunteered and more threats of data are appears in private and public database, and the insufficiency of private and public policy making the database more 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 [3]. 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 a definite challenge, methods and techniques for big data visualization need enhancements. Now a day open-sources projects and companies notice the upcoming of big data analysis via visualization[15], 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 [14], data size have extended exponentially, and with the aid of 2020 the quantity of digital bits will be comparable to the variety of stars inside the global. As the amount of bits grow every two years, for the period from 2013 to 2020 universe data will increase from 4.4 to 44 zettabytes. The huge data expansion may leads to difficulty related to human ability in dealing with the data, gain knowledge and gather information from it [26]. This paper provides information and enhancement about re-estimated tracking and recognition system of big data visualization aims to keep away from mismatch of the actual view scene and machine generated items.

1.2 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. Now a day 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 [7]. 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 [8].

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 [24].

2. 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 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 [16].

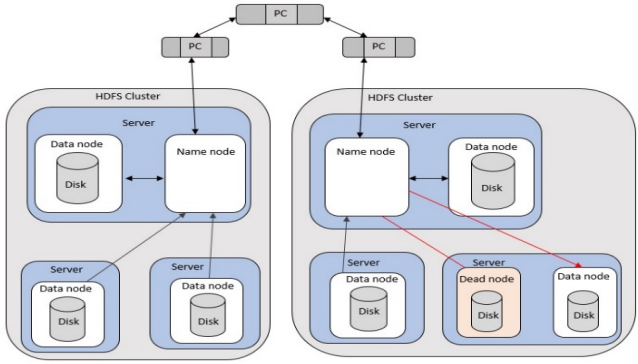

Figure 1. The HDFS architecture

2.1 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 [27].

As Figure 1. Illustrates, each cluster having one name node. 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 [27].

2.2 Name Node

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 asks 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 [9].

2.3 Data Nodes

Each block replica on a DataNode 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 [9].

A) HDFS Client

A code library that exports the HDFS file system interface, user applications access the file system using the HDFS client, [9].

B) CheckpointNode

C) BackupNode

2.4 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. Additionally the threat of information access, when data inserted in one Hadoop environment it is being easily for unwanted disclosure and theft to take place. The replicated data is likewise not protected which wishes extra protection for defensive from vulnerabilities and breaches. Because of the low security level inside a Hadoop technology most of the organizations and governments fields' 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 best set of rules regarded to hold the safety in Hadoop environment. With the intention to raise the security some approaches are noted below [3].

2.5 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 Name node. 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 [17]. The client makes use of the Kerberos authenticated connection if he needs to acquire token means. Kerberos can be used to authenticate a name node 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 name node. 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 name node 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 [3].

2.6 Walled Garden

The approach that is mostly used now a days 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 [23].

2.7 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 [11].

3. Data Visulasion Methods

3.1 Line Chart

A line chart shows the connection between each variable at the chart. Line charts are regularly used to make comparability between plenty of objects at one time. for example, there are 10 statistics points to plot or display, the greatest way to make those factors comprehensible is to simply show them in an order using a table [22].The truth that one has some statistics points to devise or show does now not mean that line graph is the satisfactory to pick out, however you ought to bear in mind the number of statistics points which you need to show which will inform the satisfactory visual method to pick out. Data points are in general being related by means of a straight line, and line chart is really an extension of Scatter plot. A few particular symbols and icons are being used to represent data points in a line chart [11].

3.2 Pie Chart

It is as familiar as a circle graph. A pie chart present data statistics and information in a way that isn't always hard to examine known as “pie-slice” form and the numerous sizes of slice indicates how tons of a detail is in existence. While the slice is huge, then it indicates of the information was gathered. Also it is used to compare values of information and the instant some values are represented on pie chart, then you may be capable of view which of the objects is the least famous or that is extra famous [12]. By imparting extra statistics, document purchasers do now not need to bet the meaning and cost of each slice. If you pick out to use a pie chart, the slices should be a percent of the entire [22].

3.3 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 [24]. 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.

3.4 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 [25].

3.5 Tracking Algorithm

A feature tracking method that solves the correspondence trouble based totally on main attributes of features, along with function, length, and mass. The important thing of the algorithm is the usage of a prediction scheme and the usage of a multi-pass look for persevering with paths. The method is extremely interactive; the scientist can guide the tracking method through changing standards and parameters, ensuing in distinct tracking answers. This tracking algorithm is based totally on an easy assumption: capabilities evolve constantly, i.e. their conduct is predictable. This means that after a path of an object is observed, it is able to make a prediction to the next frame and look for capabilities in that body that correspond to the prediction. A prediction may be made for the subsequent frame at the end of the direction, but also for the preceding body at the start of the direction. This indicates it could search forward and backward in time [21]. 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 [20].

GNN algorithm description:

1) Receiving data for current scan.

2) 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.

3) For each cluster:

3.1) 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.

3.2) 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.

4) Track Initiation.

Measurements, which are not associated with existing tracks, generate new tracks [20].

4. Conclusion

In this paper it present the big data details, approaches and visualization used in world wide. The problem are also noted to give example about the big data problems in mean time. The security problems is pointed more in order to raise the protection in big data. The protection can be improved by identifying the best security approach or by combining the right approaches together in Hadoop Distributed File System which is the main layer in Hadoop. There are quite a few demanding situations for big data processing and analysis. As all of the data is presently visualized via machines, it leads to problems in the extraction of information, this paper also obtained relevant Big Data Visualization re-estimated tracking and recognition methods, the paper give attention to visualization techniques for you to get most advantages provide by visualization. Another critical factor is to truly specify what information kind desires to represent which visualization method, that interpret most understandings, keeping in mind the significance of visualization.

References

1. Pokorny, J. (2013). NoSQL databases: a step to database scalability in web environment". International Journal of Web Information Systems, Vol. 9 Iss 1 pp. 69 – 82.
2. Venkata, N.I.; Sailaja Arsi; and Srinivasa R.R. (2014). Security Issues Associated with Big Data in Cloud Computing. International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, 10.5121.
3. Saraladevia, B.; Pazhanirajaa, N.; Paula, P.V.; Bashab, M.S.S.; Dhavachelvanc, P. (2015). Big Data and Hadoop-a Study in Security Perspective. Crossref, DOI link: https://doi.org/10.1016/j.procs.2015.04.091.
4. Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. Stamford, CT: META Group. Retrieved from https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.
5. Alvaro, A.C.; Pratyusa, K.M.; Rajan, S. (2013) Big Data Analytics for Security Intelligence. CT: Cloud Security Alliance.
6. Changqing, J.; Yu, L.; Qiu, W.; Awada, U.; and Li, K. (2012). Big Data Processing in Cloud Computing Environments. Proceedings of the International Symposium on Parallel Architectures, Algorithms and Networks. 10.1109.
7. Zhao, J.; Lizhe, W.B.; Tao, J.; Chen, J.; Sun, W.; Ranjan, R.; Kołodziej, J.; Streit, A.; and Georgakopoulos, D. (2014). A security framework in G-Hadoop for big data computing across distributed Cloud data centres. Journal of Computer and System Sciences, 10.1016.
8. Shan, Y.; Wang, B.; Yan, J.; Wang, Y.; Xu, N.; and Yang, H. (2010). FPMR: MapReduce Framework on FPGA. A Case Study of RankBoost Acceleration, 10.1145/1723112.1723129.
9. Shvachko, K.; Kuang, H.; Radia, S.; and Chansler R. (2010). An introduction to the Hadoop Distributed File System. 10.1109/MSST.2010.5496972.
10. Wang, L.; Wang, G.; Alexander, C.A. (2015). Big Data and Visualization: Methods, Challenges and Technology Progress Digital Technologies. Vol. 1, No. 1, 33-38.
11. Ajibade, S.S.; Adediran, A. (2016). An Overview of Big Data Visualization Techniques in Data Mining. Vol. 4, Issue 3, pp: 105-113.
12. Intel IT Center, (2013). Big Data Visualization: Turning Big Data into Big Insights. White Paper, pp.1-14.
13. Dean, J.; and Ghemawat, (2014). The big picture for big data: visualization, S. MapReduce: Simplified data processing on large clusters. In Proceedings of the Sixth Symposium on Operating System Design and Implementation, 10.1126.
14. Turner, V.; Reinsel, D.; Gantz, J.F.; and Minton, S. (2014). The Digital Universe of Opportunities:

Rich Data and the Increasing Value of the Internet of Things. IDC Analyze the Future.

15. Husain, S.S.; Kalinin, A.; Truong, A.; and Dinov, I.D. (2015). SOCR data dashboard: an integrated Big Data archive mashing medicare, labor, census and econometric information. 10.1186/s40537-015-0018-z.
16. Tantisiriroj, W.; Patil, S.; and Gibson, G. (2008) Data-intensive file systems for Internet services: A rose by any other name. Technical Report CMUPDL- 08-114, Parallel Data Laboratory, Carnegie Mellon University, Pittsburgh, PA.
17. Al-Janabi, Rasheed, and M.A. (2011). Public-Key Cryptography Enabled Kerberos Authentication. Journal of Developments in E-systems engineering.
18. Deyhim, P. (2013). Best Practices for Amazon EMR.
19. Michael, C.C.; Wyk, K.V.; and Radosevich, W. (2013). Black Box Security Testing Tools. Retrieved December 28, 2005, from https://www.us-cert.gov/bsi/articles/tools/black-box-testing/black-box-security-testing-tools.
20. Konstantinova, P.; Udvarev, A.; Semerdjiev, T. ND. A Study of a Target Tracking Algorithm Using Global Nearest Neighbor Approach.
21. Reinders, F.; Frits, H.; and Hans, J.W.S. ND. Visualization of Time-Dependent Data using Feature Tracking. 10.1007/PL00013399.
22. SAS, (2014). Data Visualization Techniques: From Basics to Big Data with SAS® Visual Analytics.
23. Securosis, (2016). Securing Hadoop: Security Recommendations for Hadoop Environments.
24. Shneiderman, Ben, Plaisant, and Catherine, (2009). Tree maps for space-constrained visualization of hierarchies.
25. Turaga, P.; Chellappa, R.; Subrahmanian, V.S.; and Udrea, O. (2008). Machine recognition of human activities: A survey. Journal of Circuits and Systems for Video Technology, vol. 18, no. 11, pp. 1473–1488.
26. OlshannikovaEmail, E.; Ometov, A.; Koucheryavy, Y.; Olsson, T. (2015). Visualizing Big Data with augmented and virtual reality: challenges and research agenda. 10.1186/s40537-015-0031-2.
27. Hanson, J. (2011). An introduction to the Hadoop Distributed File System. Retrieved from https://www.ibm.com/developerworks/library/wa-introhdfs.