Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 50 (2015) 596 – 601

Procedia Computer Science

2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Big Data and Hadoop-A Study in Security Perspective

B. Saraladevi[a], N. Pazhaniraja[a], P. Victer Paul[a], M.S. Saleem Basha[b], P. Dhavachelvan[c]

[a] *Department of Information Technology, Sri Manakula Vinayagar Engineering College, Puducherry, India.*
[b] *Department of Computer Science, Mazoon Univesity College, Muscat Oman.*
[c] *Department of Computer Science, Pondicherry University, Puducherry, India.*
{saramit91, victerpaul, pazhanibit, m.s.saleembasha, dhavachelvan}@gmail.com

Abstract

Big data is the collection and analysis of large set of data which holds many intelligence and raw information based on user data, Sensor data, Medical and Enterprise data. The Hadoop platform is used to Store, Manage, and Distribute Big data across several server nodes. This paper shows the Big data issues and focused more on security issue arises in Hadoop Architecture base layer called Hadoop Distributed File System (HDFS). The HDFS security is enhanced by using three approaches like Kerberos, Algorithm and Name node.

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

*Keywords:*Big data;Hadoop;HDFS;Security

1. Introduction

Big data [1] is a current technology and also going to rule a world in future. It is the Buzz word hiding both technical and marketing data inside it. The data that is small which collected in big size forms a terms called Big data and in real time its rate of growth is increased from Gigabytes in 2005 to Exabyte in 2015(forecast) which is reported by IDC research in Universe. Unfortunately big data holds large terabytes of data which cannot be maintained or stored in traditional database and it is travelled towards more latest technology which holds large datasets in it. In 1944 Fremont Rider [2] mentioned that American University library were doubling in size every sixteen years. He represented in 2040 this library will hold more than 200,000,000 volumes of books which will occupy 6000 miles of shelves in library.

Fig.1 Big Data Market

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)
doi:10.1016/j.procs.2015.04.091

The Fig 1 shows the future of big data market which is announced by International Data Corporation in March 2012. There are more 1 billion people using a mobile for transferring information per month where these data are monitored by telecommunication big data centre and allows the data centre to store more than 621 petabytes of data per year. The big data analytics [3] allows quickly identifying the risks and opportunities and also increasing capabilities of predictive analysis and Big Data Characteristics [22].

Table.1 Difference between Traditional Data and Big data

Components	Big data	Traditional Data
Queries	Largely Abandoned SQL	Traditional SQL
Architecture	Distributed	Centralized
Data Types	Structured,Semi-Structured and Unstructured	Structured
Data Model	No schema	Fixed Schema
Data Relationship	Unknown or complex Relationships	Known Relationship
Data volume	Petabytes or Exabytes	Terabytes
Data Traffic	More	Less
Data Integrity	Less	High

1.2 Big Data Issues

There are many issues arising in big data. They are Management issues, Processing Issues [24], Security issues, and Storage issues [25]. Each issue has its own task of surviving in big data and mainly focusing on security issues.

a. Management Issues

The biggest data management [23] is the collection of large volumes of Structured, Semi structured and unstructured data from the organization, Government sector and Private and Public Administration. The motto of big data management is ensuring a high data quality, data ownership, responsibilities, standardization, documentation and accessibility of data set. According to Gartner [4]"Big data" Challenge Involves More than Just Managing Volumes of Data mentioned in his Article.

b. Storage Issues

The Storage is achieved using virtualization in big data where it holds large set of Sensor information, media, videos, E-business transaction records, Cell Phone Signal Coordinates. Many Big data Storage Companies Like EMC [12], IBM, Netapp, Amazon Handles a data in a Large volume by using some tools like NoSQL, Apache Drill, Horton Works [13], SAMOA, IKANOW, Hadoop, Map reduce, Grid Gain.

c. Processing Issue

The big data processing analyzes the big data size in Petabyte, Exabyte or even in Zettabyte either in Batch Processing or Stream Processing.

d. Security Issues

There are fewer challenges for managing a large data set in secure manner and inefficient tools, public and private database contain more threats and vulnerabilities, volunteered and unexpected leakage of data, and deficiency of Public and Private Policy makes a hackers to collect their resources whenever required. In Distributed programming frameworks, the security issues start working when massive amount of private data stored in a database which is not encrypted or in regular format. Securing the data in presence of untrusted people is more difficult and when moving from homogeneous data to the Heterogeneous data certain tools and technologies for massive data set is not often developed with more security and policy certificates. Sometimes data hackers and system hackers involves in collecting a publicly available big data set, copy it and store it in a devices like USB drives, hard disk or in Laptops. They involves in attacking the data storage by sending some attacks like Denial of Service [14],Snoofing attack and Brute Force attack [15]. If the unknown user knows about the key value pairs of data it makes them to collect atleast some insufficient information. When the Storage of data increases from single tier to Multi storage tier the security tier must also be increased. In order to reduce these issues some cryptographic Framework techniques and robust algorithm must be developed in order to enhance the security of data for future. Similarly some tools are developed like Hadoop; NoSQL technology can be used for big data storage. In our

proposed work some ideas are given to overcome security issues in Hadoop environment.

2.Hadoop

Hadoop (Highly Archived Distributed Object Oriented Programming) was created by Goug Cutting and Mike Cafarella in 2005 for supporting a distributed search Engine Project. It is an Open source Java Framework technology helps to store, access and gain large resources from big data in a distributed fashion at less cost,high degree of fault tolerance and high scalability. Hadoop [5] handles large number of data from different system like Images,videos, Audios, Folders, Files, Software, Sensor Records, Communication data, Structured Query, unstructured data, Email& conversations, and anything which we can't think in any format. All these resources can be stored in a Hadoop cluster without any schema representation instead of collecting from different systems. There are many components involved in Hadoop like Avro, Chukwa, Flume, HBase, Hive, Lucene, Oozie, Pig, Sqoop and Zookeeper. The Hadoop Package also provides Documentation, source code, location awareness, Work scheduling. A Hadoop cluster contains one Master node and Many Slave nodes. The master node consists of Data node,Name node, Job Tracker and Task Tracker where slave node acts as both a TaskTracker and Data node which holds compute only and data only worker node. The Job Tracker manages the job scheduling. Basically Hadoop consists of two Parts. They are Hadoop Distributed File system (HDFS) and Map Reduce[6].HDFS provides Storage of data and Map Reduce provides Analysis of data in clustered environment.The Architecture of Hadoop is represented in figure 3.

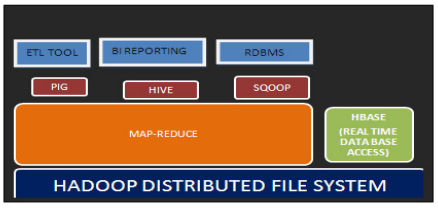

Fig.3 Hadoop Architecture

2.1HDFS Architecture

The HDFS is the Java portable file system which is more scalable, reliable, distributed in the Hadoop framework environment. A Hadoop cluster contains the combination of single Name node and group of Data nodes. Using Commodity Hardware it provides redundant storage of large amounts of data with low latency where it performs the operations like “Write Once, Read Many Times”. The files are stored as Block with default size of 64MB.The communication between the nodes occurs through Remote Procedure calls. Name node stores metadata like the name, replicas,file attributes,locations of each block address and the fast lookup of metadata is stored in Random Access Memory by Metadata. It also reduces the data loss and prevents corruption of the file system.Name node only monitors the number of blocks in data node and if any block lost or failed in the replica of a datanode,the name node creates another replica of the same block.Each block in the data node is maintained with timestamp to identify the current status. If any failure occurs in the node, it need not be repair immediately it can be repaired periodically. HDFS [7] allows more than 1000 nodes by a single Operator. Each block is replicated across many data nodes where original data node is mentioned as rack 1 and replicated node as rack 2 in Hadoop framework and never supports Data[21] Cache [19][20] due to Large set of data. The architecture of HDFS is shown in Figure 4.

i) Security Issues in HDFS

The HDFS is the base layer of Hadoop Architecture contains different classifications of data and it is more sensitive to security issues. It has no appropriate role based access for controlling security problems.Also the risk of data access,theft,and unwanted disclosure takes place when embedded a data in single Hadoop environment.Thereplicated data is also not secure which needs more security for protecting from breaches and vulnerabilities. Mostly Government Sector and Organisations never using Hadoop environment for storing valuable data because of less security concerns inside a Hadoop Technology. They are providing security in outside of Hadoop Environment like firewall and Intrusion Detection System. Some authors represented that the HDFS in

Hadoop environment is prevented with security for avoiding the theft, vulnerabilities only by encrypting the block levels and individual file system in Hadoop Environment.Even though other authors encrypted the block and nodes using encryption technique but no perfect algorithm is mentioned to maintain the security in Hadoop Environment. In order to increase the security some approaches are mentioned below.

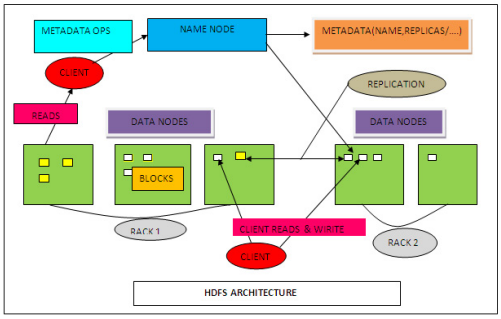

Fig.4 HDFS Architecture

ii) HDFS Security Approaches

The proposed work represents different Approaches for securing data in Hadoop distributed file system. The first approach is based on Kerberos in HDFS.

a. Kerberos Mechanism

Kerberos [10] is the network authentication protocol which allows the node to transfer any file over non secure channel by a tool called ticket to prove their unique identification between them. This Kerberos mechanism is used to enhance the security in HDFS. In HDFS the connection between client and Name node is achieved using *Remote Procedure Call* [11] and the connection from Client (client uses HTTP) to Data node is Achieved using *Block Transfer*. Here the Token or Kerberos is used to authenticate a RPC connection. If the Client needs to obtain a token means, the client makes use of Kerberos Authenticated Connection. *Ticket Granting Ticket (TGT)* or *Service Ticket* are used to authenticate a name node by using Kerberos. Both TGT and ST can be renewed after long running of jobs while Kerberos is renewed, new TGT and ST is also issued and distributed to all task. The Key Distribution Centre (KDC) issues the Kerberos Service Ticket using TGT after getting request from task and network traffic is avoided to the KDC by using Tokens In name node, only the time period is extended but the ticket remain constant. The major advantage is even if the ticket is stolen by the attacker it can't be renewed. We can also use another method for providing security for file access in HDFS.

If the client wants to access a block from the data node it must first contact the name node in order to identify which data node holds the files of the blocks. Because of name node only authorize access to file permission and issues a token called *Block Token* where data node verifies the token. The data node also issues a token called *Name Token* where it allows the Name node to enforce permission for correct control access on its data blocks. Block Token allows the data node to identify whether the client is authorized access to access data blocks. These block token and Name Token is sent back to client who contains data block respective locations and you're the authorized person to access the location**.** These two methods are used to increase security by preventing from unauthorized client must read and write in data blocks. The figure 5 shows the design view of Kerberos key distribution centre.

b. Bull Eye Algorithm Approach

In big data the sensitive data are credit card numbers, passwords, account numbers, personal details are stored in a large technology called Hadoop. In order to increase the security in Hadoop base layer the new approach is introduced for securing sensitive information which is called "*Bull Eye Approach*". This approach is introduced on Hadoop module to view all sensitive information in 360° to find whether all the secured information are stored

without any risk, and allows the authorized person to preserve the personal information in a right way. Recently this approach is using in companies like Dateguise's DGsecure[8] and Amazon Elastic Map Reduce[9]. The DGsecure Company which is famous for providing a Data centric security and Governance solutions also involves in providing a security for Hadoop in the cloud. The data guise company is decided to maintain and provide security in Hadoop wherever it is located in cloud. Now a days the Companies are storing a more sensitive data in cloud because of more breaches taking place in traditional on premise data store. To increase the security in Hadoop base layers, the Bull eye Approach also used in HDFS to provide security in 360° from node to node. This approach is implemented in Data node of rack 1, where it checks the sensitive data are stored properly in block without any risk and allows only the particular client to store in required blocks. It also bridges a gap between a data driven from original data node and replicated data node. When the client wants to retrieve any data from replicating data nodes it also maintained by "Bull Eye Approach" and it checks where there is a proper relation between two racks. This Algorithm allows the data nodes to be more secure, only the authorized person read or write about it. The algorithm can be implemented below the data node where the client read or writes the data to store in blocks. It is not only implemented in the rack 1 similarly it is implemented in Rack 2 in order to increase the security of the blocks inside the data nodes in 360°. It checks for any attacks, breaches or theft of data taking place in the blocks of the data node. Sometimes data are encrypted for protection in data mode. These types of encrypted data also protected using this Algorithm in order to main order security. The Algorithm travels from less terabyte to multi-petabytes of semi-structured, structured and unstructured data stored in HDFS layer in all angles. Mostly encryption and Wrapping of data occurs at the block levels of Hadoop rather than entire file level. This algorithm scans before the data is allowed to enter into the blocks and also after enters both rack 1 and rack 2. Thus, this Algorithm concentrates only on the sensitive data that matters about the information stored in the data nodes. In our work, we mentioned this new Algorithm to enhance more security in the data nodes of HDFS.

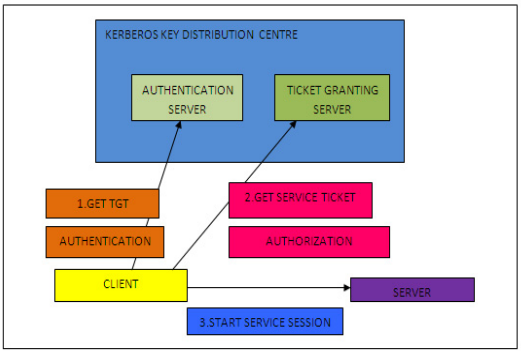

Fig.5 Kerberos Key Distribution Centre

c. Namenode Approach

In HDFS if there is any problem in Name node event and becomes unavailable, it makes the group of system service and data stored in the HDFS make unavailable so it is not easy to access the data in secure way from this critical situation. In order to increase the security in data availability, it is achieved by using two Namenode. These two Name nodes servers are allowed to run successfully in the same cluster. These two redundant name nodes are provided by Name Node Security Enhance (NNSE), which holds Bull Eye Algorithm. It allows the Hadoop administrator to run the options for two nodes. From these name node one acts as Master and other acts as a slave in order to reduce an unnecessary or unexpected server crash and allows predicting from natural disasters. If the Master Name node crashes, the administrator needs to ask permission from Name Node Security Enhance to provide a data from a slave node in order to cover a time lagging and data unavailability in secure manner. Without getting permission from NNSE admin never retrieves the data from slave node to reduce the complex retrieval issue. If both Name node acts as a master there is a continuous risk occurs, reduces a secure data availability and bottleneck in performance over a local area network or Wide Area Network. Thus in future we can also increase security by using *Vital configuration* that provides and ensures data is available in secured way to client by replicating many Name node by Name Node Security Enhance in HDFS blocks between many data centres and clusters.

3. Discussion

The proposed work represents different Approaches for securing data in Hadoop distributed file system.The first approach is based on Kerberos in HDFS, it is used to access a data blocks correctly and also only by an authorised user. Here Ticket Granting Ticket and Service Ticket playing a major role in providing a security in name node.The Second approach is based on Bull Eye Algorithm Approach explains about the security method from node to node and also scan the nodes in all the angles to prevent from attacks. The third approach is based on Name node where the security is achieved by replicating [17] a name node to reduce the server crashes for future references.

4. Conclusion

This paper 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. We can improve security in big data by using any one of the approach or by combining these three approaches in Hadoop Distributed File System which is the base layer in Hadoop, where it contains large number of blocks. These approaches are introduced to overcome certain issues occurs in the name node and also in Data node. In Future these approaches are also implemented in other layers of Hadoop Technology.

References

[1] Prof. Dr. Philippe Cudré-Mauroux, “An Introduction to BIG DATA”, June 6, 2013 Alliance EPFL, http://exascale.info/
[2] Fremont Rider, “The future of the Research Library”, http://crl.acrl.org/content/50/1/48.html
[3]RobPegler,“Introduction to big data, analytics knowledge and skill approach with various techniques,http://www.snia.org/sites/default/files2/ABDS2012/Tutorials/RobPeglar_IntroductionAnalytics%20_Big%20Data_Hadoop.pdf
[4] Gartner, http://www.gartner.com/newsroom/id/2848718, STAMFORD, Conn., September 17, 2014
[5] “Leveraging Massively Parallel Processing in an Oracle Environment for Big Data”, An Oracle White Paper, November 2010.
[6]Jeffrey Dean and Sanjay Ghemawat, “Map Reduce: Simplified Data Processing on Large Clusters”, Google, Inc.
[7] “Hadoop and HDFS:Storage for Next Generation Data Management”, Cloudera, Inc, 2014.
[8]Data guise protect, http://www.dataguise.com/?q=dataguise-dgsecure-platform
[9] Parviz Deyhim, “Best Practices for Amazon EMR”, August 2013.
[10] Al-Janabi, Rasheed, M.A.-S., “Public-Key Cryptography Enabled Kerberos Authentication”, IEEE, Developments in E-systems Engineering (DeSE), 2011
[11] Heindel L.E, “Highly reliable synchronous and asynchronous remote procedure calls”, Conference Proceedings of the IEEE Fifteenth Annual International Phoenix Conference on computers and communications,1996.
[12] The journey to big data, EMC2 Publications.
[13] Horton Technical Preview for Apache Spark, Horton works Inc.
[14] Shay Chen, “Application Denial of Service”, Hack tics Ltd, 2007.
[15] Daniel J. Bernstein, “Understanding brute force”, National Science Foundation, Chicago.
[16] Introduction to Pig, Cloud era, 2009.
[17] P. Victer Paul, N. Saravanan, S.K.V. Jayakumar, P. Dhavachelvan and R. Baskaran, “QoS enhancements for global replication management in peer to peer networks”, Future Generation Computer Systems, Elsevier, Volume 28, Issue 3, March 2012, Pages 573–582. ISSN: 0167-739X.
[18]Ashish Thusoo, Joydeep Sen Sarma, “Hive –A Petabyte Scale Data Warehouse Using Hadoop, Facebook Data Infrastructure Team.
[19]P. Victer Paul, D. Rajaguru, N. Saravanan, R. Baskaran and P. Dhavachelvan, "Efficient service cache management in mobile P2P networks", Future Generation Computer Systems, Elsevier, Volume 29, Issue 6, August 2013, Pages 1505–1521. ISSN: 0167-739X.
[20]N. Saravanan, R. Baskaran, M. Shanmugam, M.S. SaleemBasha and P. Victer Paul, "An Effective Model for QoS Assessment in Data Caching in MANET Environments", International Journal of Wireless and Mobile Computing, Inderscience, Vol.6, No.5, 2013, pp.515-527. ISSN: 1741-1092.
[21] R. Baskaran, P. Victer Paul and P. Dhavachelvan, "Ant Colony Optimization for Data Cache Technique in MANET", International Conference on Advances in Computing (ICADC 2012), Advances in Intelligent and Soft Computing" series, Volume 174, Springer, June 2012, pp 873-878, ISBN: 978-81-322-0739-9.
[22] https://www.ida.gov.sg/~/media/Files/Infocomm%20Landscape/Technology/TechnologyRoadmap/BigData.pdf
[23] Philip Russom, “Managing Big Data”, TDWI research, Fourth Quarter 2013.
[24] Changqing, “Big Data Processing in Cloud Computing Environments”, International Symposium on Pervasive Systems, Algorithms and Networks, 2012
[25] Young-Sae Song, “Storing Big Data- The rise of the Storage Cloud” , December, 2012.