IMPLEMENTASI DAN ANALISIS HARDENING DATA CONFIDENTIALITY PADA INFRASTRUKTUR HADOOP DENGAN METODE ENCRYPTION UNTUK PENGAMANAN DATA

IMPLEMENTATION AND ANALYSIS OF HARDENING DATA CONFIDENTIALITY IN HADOOP INFRASTRUCTURE WITH ENCRYPTION METHOD FOR DATA SECURITY

Raisul Amin Zulfikri[1], Adityas Widjajarto[2], Ahmad Almaarif[3]

[1,2,3] Universitas Telkom, Bandung
[1]**aminzulfikri@student.telkomuniversity.ac.id**, [2]**adtwjrt@telkomuniveristy.co.id**,
[3]**ahmadalmaarif@telkomuniversity.ac.id**

Abstrak

***Big data* perlu sebuah infrastruktur untuk menjalankan sistemnya beserta salah satu *framework opensource* pengolahan *big data* yaitu Hadoop. Hadoop berfungsi untuk memproses, mengatur serta menganalisis dari berbagai macam tipe data. *Vulnerabilities* awalnya tidak terlalu diperhatikan pada infrastruktur Hadoop. Namun disaat Hadoop memiliki *issue* mengenai keamanan yang sedang menjadi prioritas dalam menghadapi distribusi data dan proses data secara paralel dan oleh karena itu *vulnerabilities* semakin diperhatikan dalam perkembangan *big data*, pada saat ini, masalah utama *big data* yang paling menantang yaitu mengenai keamanan dan privasi.**

Dalam mengatasi masalah mengenai keamanan pada infrastruktur Hadoop, dibutuhkan *vulnerability assesment tool* yang dapat melakukan *vulnerability scan* yang dilanjutkan dengan melakukan *hardening data confidentiality*. Eksperimen ini dilakukan untuk melihat apakah metode *hardening* dapat membantu dalam memperkuat keamanan yang ada pada infrastuktur Hadoop.

Sehingga eksperimen *vulnerability assessment* menggunakan *tool* Greenbone Security Manager untuk pengujian kerentanan pada infrastruktur Hadoop yang dilakukan akan menghasilkan kerentanan yang ditemukan. Hasil kerentanan yang ditemukan akan dianalisis kemudian dapat dilakukan eksperimen *hardening data confidentiality* pada infrastruktur Hadoop. Hasil dari *hardening* yang telah dilakukan pada infrastruktur Hadoop dapat dianalisis sehingga diketahui apakah *hardening data confidetiality* dapat mengatasi kerentanan pada infrastruktur Hadoop

Kata kunci : ***big data, hadoop, vulnerability assessment, hardening, data confidentiality.***

Abstract

Big data needs an infrastructure to run the system along with one of the open source big data processing frameworks, namely Hadoop. Hadoop works to process, organize and analyze various types of data. The vulnerability was not noticed initially in Hadoop infrastructure. However, when facing security issues that are becoming a priority in dealing with data distribution and data processing in parallel and therefore vulnerabilities need to be considered in the development of big data, at this time, the main big data problems that are the most challenging are security and privacy.

In overcoming security problems in Hadoop infrastructure, a vulnerability assessment tool is needed that can perform vulnerability scanning followed by hardening data confidentiality. This experiment was conducted to see if the hardening method can help in strengthening the existing security on the Hadoop infrastructure.

So that vulnerability assessment experiments using the Greenbone Security Manager tool for vulnerability testing on Hadoop infrastructure will result in vulnerabilities being found. The results of the vulnerabilities that will be analyzed can then be experimented with hardening data confidentiality on the Hadoop infrastructure. The results of the hardening that have been carried out on the Hadoop

infrastructure can be analyzed so that it is known whether the confidentiality of the hardening data can overcome the vulnerabilities in the Hadoop infrastructure.

Keywords: ***big data, hadoop, vulnerability assessment, hardening, data confidentiality*****.**

1. Pendahuluan

Vulnerabilities pada infrastruktur Hadoop tidak terlalu diperhatikan pada awalnya. Namun ketika Hadoop memiliki *issue* mengenai keamanan yang tidak menjadi prioritas dalam menghadapi distribusi data dan proses data secara paralel maka dari itu masalah *vulnerabilities* semakin diperhatikan [1]. Pada awalnya, dalam perkembangan *big data*, *developer* tidak mempertimbangkan isu yang muncul dari segi keamanan dan privasi. Namun, pada saat ini, masalah utama *big data* yang paling menantang yaitu mengenai keamanan dan privasi [2].

Maka dari itu untuk mendapatkan informasi mengenai kerentanan sistem pada Hadoop yang dapat menjadi acuan untuk membangun sistem Hadoop yang memiliki keamanan dan privasi yang lebih baik, perlu dilakukan *vulnerability assessment* menggunakan *tool* yang sesuai pada infrastruktur Hadoop sehingga akan menghasilkan informasi mengenai *vulnerability* yang ada. Dari hasil yang ditemukan, dapat dilakukan *hardening* seperti pada aspek *confidentiality* untuk menjaga kerahasiaan data yang ada dengan menggunakan metode yang sesuai sehingga tidak ada kebocoran data penting dan hanya orang yang berwenang yang dapat mengakses data tersebut. Hal ini tentu meningkatkan keamanan pada infrastruktur Hadoop serta dapat membantu mengurangi kerentanan yang ditemukan pada infrastruktur Hadoop.

2. Dasar Teori

2.1 *Big Data*

Big Data merupakan kombinasi dari berbagai faktor, seperti waktu dan tipe data. *Big data* terdiri dari data berkecepatan tinggi dengan volume yang besar, kompleks, dan variabel, yang membutuhkan metode dan teknologi canggih untuk menangkap, menyimpan, mendistribusikan, mengelola, dan menganalisis informasi data tersebut [3]. *Big data* berisi data yang memiliki banyak *variety* (variasi) yang datang dalam *volume* yang terus meningkat dan dengan *velocity (*kecepatan) yang semakin tinggi. Ini dikenal sebagai tiga V [4]. Sederhananya, *big data* adalah kumpulan data yang lebih besar dan lebih kompleks, terutama dari *resources* data baru. Kumpulan data ini sangat banyak sehingga *software* pemrosesan data tradisional tidak dapat mengelolanya. Tetapi data dalam jumlah besar ini dapat digunakan untuk mengatasi masalah bisnis terhadap data yang sebelumnya tidak dapat diatasi.

2.2 Hadoop

Hadoop adalah kerangka kerja pemrosesan terdistribusi (distributed processing framework) yang bersifat *open-source* yang mengelola pemrosesan dan penyimpanan data untuk aplikasi Big Data dalam cluster server komputer yang bersifat scalable. Hadoop juga sebagai pusat ekosistem teknologi Big Data yang terutama digunakan untuk mendukung inisiatif analitik tingkat lanjut, termasuk analitik prediktif, penambangan data (data mining), dan machine learning. Sistem Hadoop dapat menangani berbagai bentuk data terstruktur dan tidak terstruktur, memberikan pengguna lebih banyak fleksibilitas untuk mengumpulkan, memproses, menganalisis, dan mengelola data daripada yang disediakan oleh database relasional dan gudang data.

2.3 *Data Security*

Elemen inti dari keamanan data adalah kerahasiaan (*confidentiality*), integritas (*integrity*), dan ketersediaan (*availability*). Juga dikenal sebagai CIA, ini adalah model dan panduan keamanan bagi organisasi untuk menjaga data sensitif mereka agar terlindungi dari akses dari individu yang tidak berwenang (*unauthorized access*) dan eksfiltrasi data (*data exfiltration*). Kerahasiaan (*confidentiality*) memastikan bahwa data hanya diakses oleh individu yang berwenang. Integritas (*integrity*) memastikan bahwa informasi dapat diandalkan serta akurat. Ketersediaan (*availibility*) memastikan bahwa data tersedia dan dapat diakses untuk memenuhi kebutuhan bisnis [5].

2.4 Greenbone Security Manager

GSM (Greenbone Security Manager) adalah *tool vulnerability assessment* dengan solusi *feature-rich* yang menyediakan kemampuan yang diperlukan untuk integrasi ke dalam arsitektur keamanan secara keseluruhan, bahkan untuk jaringan dengan keamanan tinggi yang memerlukan pendekatan *air-gap*. Dibuat untuk penggunaan profesional di perusahaan dan administrators, digunakan sebagai alat *turn-key*.

2.5 *Hardening*

Hardening adalah kumpulan alat, teknik, dan praktik terbaik untuk mengurangi kerentanan dalam aplikasi teknologi, sistem, infrastruktur, *firmware*, dan area lainnya. Tujuan pengerasan sistem adalah untuk mengurangi risiko keamanan dengan menghilangkan potensi serangan dan menguatkan suatu sistem dari serangan.

2.6 *Encryption* dan *Decryption*

Encryption adalah metode untuk mengamankan informasi atau data sehingga menjadi pola atau sesuatu yang tidak mudah dipahami tanpa adanya kunci untuk memecahkan pola tersebut. Sedangkan *decryption* merupakan kebalikan dari *encryption*, yaitu proses memecahkan suatu pola tertentu agar informasi atau data yang telah diamankan atau di enkripsi dapat dilihat atau dibaca kembali.

2.7 Hadoop *Data at Rest Encryption*

Enkripsi *data at rest* dilakukan dengan tujuan pengamanan data pada saat data sedang tersimpan dalam *DataNode.* Ada beberapa metode yang dapat untuk mengenkripsi *Data at Rest* (Data yang tersimpan) pada Hadoop, yaitu *Filesystem encryption, Application level encryption,* HDFS *Data at Rest encryption.*

2.8 Hadoop *Data-in-transit Encryption*

Enkripsi *data-in-transit* dilakukan dengan tujuan pengamanan data pada saat data sedang berpindah, contohnya perpindahan data antar *DataNode* pada infrastruktur Hadoop. Ada beberapa metode yang dapat digunakan untuk mengenkripsi *Data-in-transit* (Data yang berpindah) pada Hadoop, yaitu Hadoop RPC *Encryption*, HDFS *data transfer protocol*, dan Hadoop HTTP *encryption*.

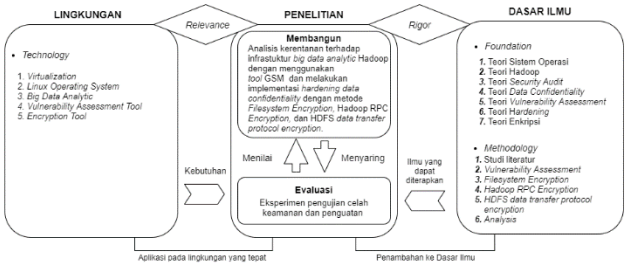
2.9 eCryptfs

ECryptfs adalah sistem file kriptografi berdasarkan PGP yang dibuat oleh Philip Zimmerman pada tahun 1991. Hal yang menonjol tentang eCryptfs dibandingkan dengan sistem file enkripsi lainnya, seperti TrueCrypt, adalah bahwa kita tidak perlu mengalokasikan sejumlah ruang *disk* yang kita miliki sebelumnya ketika ingin mengenkripsi

3. Metodologi Penelitian

3.1 Model Konseptual

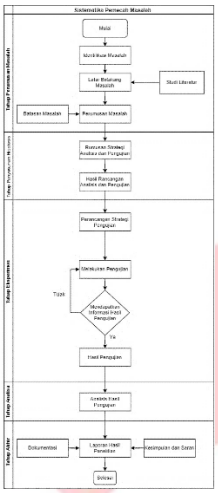
Model konseptual ini mengarah kepada kerangka penelitian tugas akhir untuk menganalisis kerentanan terhadap infrastruktur Hadoop menggunakan *vulnerability assessment tool* yaitu GSM (Greenbone Security Manager) yang bertujuan untuk mengetahui apa saja *vulnerabilities* yang terdapat pada infrastruktur Hadoop dan melakukan penguatan kerahasiaan data (*data confidentiality*) terhadap infrastruktur Hadoop.



Gambar 3.1 Model Konseptual

3.2 Sistematika Penelitian

Pada penelitian ini tahapan yang menggunakan lima tahapan yang dapat dilihat pada gambar dibawah ini. Terdapat lima tahapan yang harus dilakukan untuk menerapkan metode ini, yaitu tahap perumusan masalah, tahap penyusunan hipotesis, tahap eksperimen, tahap analisia, dan tahap akhir.



Gambar 3.2 Sistematika Penelitian

4. Rancangan Sistem dan Implementasi

4.1 Hadoop *multi-node*

Penelitian ini mengunakan hadoop *multi-node cluster*, yang memiliki *node master* yaitu hadoop-master dan dua *slave* atau worker *node* yaitu hadoop-slave1 dan hadoop-slave2. *Node master* dalam *cluster* hadoop bertanggung jawab untuk menyimpan data dalam HDFS dan mengeksekusi komputasi paralel data yang disimpan menggunakan MapReduce.

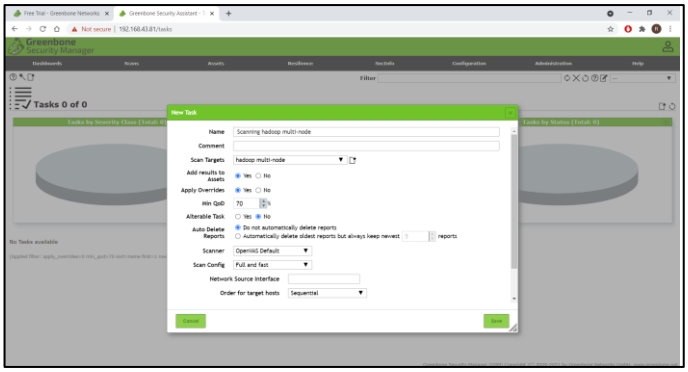
Tabel 4. 1 IP *Address* Hadoop *Multi-node*

Nama	IP *Address*
hadoop-master	192.168.197.137
hadoop-slave1	192.168.197.138
hadoop-slave2	192.168.197.139

Hadoop-slave memiliki *DataNode*, dalam Arsitektur Hadoop HDFS, *DataNode* menyimpan data aktual dalam HDFS. *DataNode* juga bertanggung jawab untuk menjalankan *read-write operation* sesuai dengan permintaan, dan mengirimkan informasi ke NameNode tentang file dan blok yang disimpan di *node* tersebut dan merespons *NameNode* untuk semua *filesystem operation*.

4.2 Implementasi GSM Trial

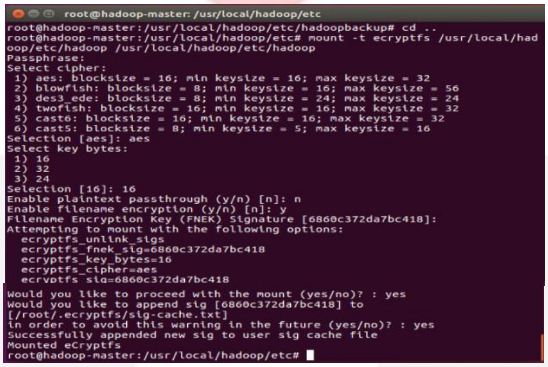
Penggunaan GSM Trial ditujukan untuk melakukan *vulnerability assessment* pada infrastruktur Hadoop *multi-node* yang telah dibuat. GSM trial akan melakukan *scan vulnerability* secara otomatis, menyeluruh dan secara *real*, sehingga nantinya kita dapat melakukan analisis *vulnerability*, dan kemudian dapat dilakukan proses *hardening*. Gambar 4.1 adalah implementasi GSM Trial pada saat pembuatan *task* untuk melakukan *vulnerability scanning* terhadap infrastruktur Hadoop.



Gambar 4.1 Implementasi GSM Trial

4.3 Implementasi *Filesystem Encryption*

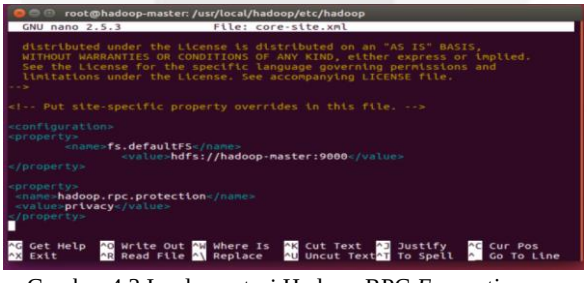
Pada implementasi Hadoop *Data at Rest Encryption* ini menggunakan metode *Filesystem Encryption* guna melindungi *data* konfigurasi dan meningkatkan aspek *data confidentiality* pada infrastruktur Hadoop. Implementasi *hardening* dengan metode *filesystem encryption* menggunakan eCryptfs dilakukan pada direktori *file configuration* /usr/local/hadoop/etc/hadoop.



Gambar 4.2 Implementasi *Filesystem Encryption*

4.4 Implementasi Hadoop RPC *Encryption*

Remote Procedure Call (RPC) adalah protokol yang dapat digunakan pada satu program untuk berkomunikasi dengan program yang terletak di komputer atau *client* yang lain di suatu jaringan tanpa harus memahami detail jaringan. RPC digunakan untuk memanggil proses lain pada sistem jarak jauh seperti sistem lokal. Untuk melakukan konfigurasi RPC *encryption*, masukkan *property* hadoop.rpc.protection dengan *value privacy* pada *file* core-site.xml.



Gambar 4.3 Implementasi Hadoop RPC *Encryption*

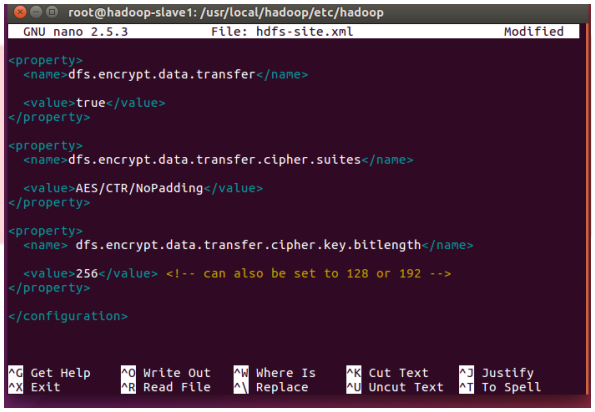
Implementasi RPC Hadoop mendukung SASL (*Simple Authentication and Security Layer*), yang mendukung otentikasi dan enkripsi. Hadoop menggunakan implementasi Java SASL, yang mendukung mode berikut:

- auth, digunakan untuk otentikasi antara klien dan server.

- auth-int, digunakan untuk otentikasi dan integritas.
- auth-conf, digunakan untuk otentikasi (*authentication*), integritas (*integrity*) dan kerahasiaan (*confidentiality*)

4.5 Implementasi HDFS Data Transfer Protocol Encryption

Ketika data HDFS ditransfer dari satu *DataNode* ke yang lain atau antara *DataNode* dan *client*, soket TCP/IP langsung digunakan dalam protokol yang dikenal sebagai HDFS *data transfer protocol*. Protokol Hadoop RPC digunakan untuk menukar kunci enkripsi untuk digunakan dalam *data transfer protocol* pada saat enkripsi transfer data diaktifkan. Gambar 4.4 adalah implementasi HDFS *Data Transfer Protocol Encryption* dengan mengkonfigurasi file hdfs-site.xml pada infrastruktur Hadoop.



Gambar 4.4 Implementasi HDFS *Data Transfer Protocol Encryption*

5. Analisis Hasil Implementasi

5.1 Hasil *scan* GSM

Gambar 5.1 adalah hasil *scan* GSM yang menunjukkan bahwa terdapat *vulnerability* dengan CVSS (*Common Vulnerability Scoring System*) *score* 10.0 (high) pada infrastruktur Hadoop. CVSS (*Common Vulnerability Scoring System*) adalah standar untuk menggambarkan tingkat keparahan risiko keamanan dalam sistem komputer. Nilai 10 berarti tingkat keparahan risiko keamanan adalah *High* (tinggi).


2.1.2 High 9870/tcp

High (CVSS: 10.0)
NVT: Apache Hadoop 'Secure Mode' Disabled

Summary
The host is installed with Apache Hadoop and has 'Secure Mode' disabled.

Vulnerability Detection Result
Vulnerability was detected according to the Vulnerability Detection Method.

Impact
Successful exploitation might allow a remote attacker to gain unauthenticated access to data saved within this Hadoop instance.

Solution:
Solution type: Mitigation
Configure 'Secure Mode' by following the Apache Hadoop documentation.

Affected Software/OS
Apache Hadoop instances with 'Secure Mode' disabled.

Vulnerability Insight
The flaw exists due to a disabled 'Secure Mode' which doesn't require authentication for users.

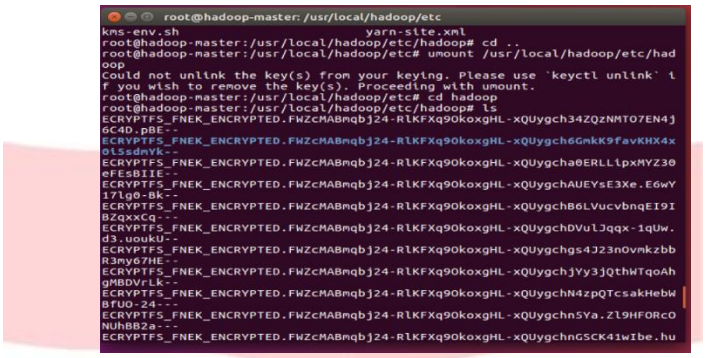
Vulnerability Detection Method
Check the status page of Apache Hadoop if 'Secure Mode' is enabled or not.
Details: Apache Hadoop 'Secure Mode' Disabled
OID:1.3.6.1.4.1.25623.1.0.108173

References
url: https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/Sec
↪ureMode.html
url: https://blog.shodan.io/the-hdfs-juggernaut/

Gambar 5.1 Hasil *Scan* GSM

5.2 Hasil implemetasi Hardening Encryption

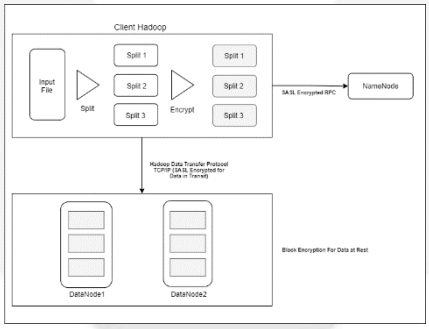
Pada Gambar 5.2 merupakan hasil *Filesystem Encryption* yang berada pada direktori file konfigurasi infrastruktur Hadoop yaitu direktori /usr/local/hadoop/etc/hadoop. Penguatan ini dapat menambahkan aspek *data confidentiality* pada Infrastruktur Hadoop. Selain memperkuat aspek *confidentiality*, penggunaan *encryption* pada direktori *file* konfigurasi juga menjadi salah satu solusi terhadap *vulnerability* Hadoop *secure mode* yang ter-*disable* karena penguatan dilakukan pada salah satu aspek *secure mode* yaitu *data confidentiality*.



Gambar 5.2 Hasil Implementasi *Filesystem Encryption*

File yang di enkripsi merupakan *file* konfigurasi pada Hadoop. Enkripsi dilakukan dengan menggunakan *cipher* AES (*Advanced Encryption Standard*) 128 bit. Pada jurnal "*Data protection on hadoop distributed file system by using encryption algorithms: a systematic literature review*" ditemukan bahwa AES berkinerja lebih baik daripada DES dalam parameter enkripsi, waktu dekripsi, dan ukuran *buffer* [6].

Ketika *file* konfigurasi terenkripsi maka *file* tersebut tidak dapat diakses oleh penyerang maupun *user* yang tidak memiliki akses terhadap *file* tersebut. Kemudian ketika file konfigurasi terenkripsi dan Infrastruktur Hadoop dinyalakan oleh penyusup atau orang yang tidak berwenang, *core service* pada infrastruktur Hadoop tidak dapat berjalan dengan semestinya. Oleh karena itu, file konfigurasi harus terdekripsi terlebih dahulu sebelum *core service* pada infrastruktur Hadoop dinyalakan. Dapat disimpulkan bahwa aspek *confidentiality* yang menjadi tujuan *hardening* pada infrastruktur Hadoop telah terealisasi.



Gambar 5.3 Kerangka Sistem *Three Dimensional Security*

Pada Gambar 5.3 adalah kerangka *sistem three dimensional security* pada Infrastruktur Hadoop dengan 3 metode. Metode yang pertama pada kerangka *sistem three dimensional security* mengacu ke dalam *Data-in-Transit Encryption* pada Infrastruktur Hadoop, yaitu metode RPC *encryption*. Metode yang kedua yaitu metode HDFS *data transfer protocol encryption*. Metode yang ketiga pada kerangka sistem *three dimensional security* mengacu ke dalam *Data at Rest Encryption* yaitu dengan menggunakan metode *Block Encryption*.

Dari perbandingan pada sistem diatas, perbedaan implementasi yang telah dilakukan terdapat pada penggunaan *Filesystem Encryption* sebagai metode untuk enkripsi pada saat data disimpan. Kelebihan penggunaan *filesystem encryption* adalah metode *filesystem encryption* mudah diterapkan dan tetap menjaga aspek *confidentiality data*.

6. Kesimpulan dan Saran

6.1 Kesimpulan

Berdasarkan analisis dari pengujian kerentanan dan implementasi *hardening* pada infrastruktur Hadoop yang telah dilakukan, maka dapat diambil kesimpulan sebagai berikut:

1. *Vulnerability Assessment tool* yang dapat dengan lengkap dalam mencari kerentanan pada infrastruktur Hadoop yang ditemukan adalah GSM, GSM dapat memberikan hasil *scan vulnerability* yang cukup lengkap, selain informasi-informasi yang diperlukan, juga diberikan referensi solusi yang dapat dilakukan untuk mengatasi *vulnerability* yang ditemukan.

2. Dalam mengendalikan kerentanan yang ada pada infrastruktur Hadoop dapat dengan melakukan implementasi Hadoop *Data at Rest Encryption* (*Filesystem Encryption*) yang memiliki fungsi untuk melakukan enkripsi pada saat data tersimpan, dan Hadoop *Data-in-Transit Encryption* (Hadoop RPC *Encryption* dan HDFS *data transfer protocol encryption*) yang memiliki fungsi untuk melakukan enkripsi pada saat pengiriman data antar *Node*. Dua implementasi yang telah dilakukan dapat memberikan *confidentiality* sehingga data tetap aman dan tidak dapat diakses oleh orang yang tidak berwenang.

6.2 Saran

Berdasarkan analisis dari pengujian kerentanan dan implementasi hardening pada infrastruktur Hadoop yang telah dilakukan, maka berikut merupakan saran yang dapat dilakukan:

1. Bagi penelitian selanjutnya, apabila menggunakan implementasi *Filesystem Encryption* akan lebih optimal jika dilakukan pada direktori *DataNode* karena dapat mengenkripsi *data* yang tersimpan pada infrastruktur Hadoop.

2. Bagi penelitian selanjutnya, apabila melakukan *hardening* pada Hadoop dapat melakukan kombinasi *hardening* pada sisi *Authentication* dan *Data Confidentiality* sesuai dengan dokumentasi Apache Hadoop *Secure Mode* karena dapat meningkatkan keamanan infrastruktur Hadoop secara lebih baik.

Referensi:

[1] Abouelmehdi, K., Beni-Hssane, A., Khaloufi, H., & Saadi, M. (2017). *Big data emerging issues: Hadoop security and privacy*. *International Conference on Multimedia Computing and Systems - Proceedings*

[2] Yadav, D., Maheshwari, D. H., & Chandra, D. U. (2019). *Big Data Hadoop: Security and Privacy*. *SSRN Electronic Journal*.

[3] Amalina, F., Targio Hashem, I. A., Azizul, Z. H., Fong, A. T., Firdaus, A., Imran, M., & Anuar, N. B. (2020). *Blending Big Data Analytics: Review on Challenges and a Recent Study*. IEEE Access, 8, 3629–3645.

[4] Beyer, M., & Laney, D. (2012). *The Importance of "Big Data": A Definition*. Gartner, G00235055.

[5] Buckbee, M. (2020, March 29). *Data Security: Definition, Explanation and Guide*. [Online] *Available at*: https://www.varonis.com/blog/data-security/ [*Accessed* 10 *February* 2021].

[6] Naisuty, M., Nizar Hidayanto, A., Clydea Harahap, N., Rosyiq, A., Suhanto, A., & Michael Samuel Hartono, G. (2020). *Data protection on hadoop distributed file system by using encryption algorithms: A systematic literature review*. *Journal of Physics: Conference Series*, *1444*(1).