Implementasi Ketersediaan Elemen Hadoop dan Analisis Berdasarkan Daemon Log Monitoring

Cikal Khairrun Nissa, Akbar Krishnawan, Adjie Budi Negoro*, FerdiAprian, Zalfa Amira, Zaidan Amru Abdillah

*Korespondensi: adjie544@gmail.com

ARTICLE INFO

Article History:
- Received 12 January 2023
- Received in revised form 22 January 2023
- Accepted 22 February 2023
- Available online 30 March 2023

ABSTRAK

Big data saat ini menjadi topik yang sering dibahas dalam industri, terutama seiring dengan perkembangan teknologi informasi yang semakin mengandalkan big data untuk mendukung berbagai aspek, terutama di bidang business intelligence, data analytics, dan machine learning. Hadoop, sebagai kerangka kerja perangkat lunak untuk komputasi yang andal, terukur, paralel, dan terdistribusi, memainkan peran penting dalam pemrosesan, pengorganisasian, dan analisis berbagai jenis data, termasuk data terstruktur, tidak terstruktur, dan semi-terstruktur. Pertumbuhan penggunaan Hadoop sejalan dengan ketersediaan layanan inti dari Hadoop itu sendiri. Untuk memastikan ketersediaan elemen Hadoop selama proses berlangsung, monitoring log dari daemon Hadoop menjadi suatu keharusan. Melalui pemantauan daemon sistem inti Hadoop, kita dapat melakukan pengecekan terhadap ketersediaan elemen-elemen kunci Hadoop selama proses berjalan.

Kata Kunci: Hadoop, Business Intelligence, Data Analytics, Log, Machine Learning

ABSTRACT

Big data is currently a frequently discussed topic in the industry, particularly in relation to the development of information technology that utilizes big data to support various sectors, especially in the fields of business intelligence, data analytics, and machine learning. Hadoop is a reliable, scalable, parallel, and distributed software framework for computation. It is used to process, organize, and analyze various types of data, including structured, unstructured, and semi-structured data. The growth in the use of Hadoop is directly proportional to the availability of core Hadoop services. To achieve this, you need to monitor the logs from the Hadoop daemon. By monitoring the core Hadoop system daemon, you can check the availability of Hadoop elements while the process is running.

Keywords: Hadoop, Intelligence, Analytics, Log, Machine Learning

1. PENDAHULUAN

Big Data pada saat ini adalah isu yang sering dibahas dalam industri, hal ini berkaitan dengan perkembangan teknologi informasi yang menggunakan Big Data untuk mendukung industri khususnya dalam Business Intelligent, Data Analytic, dan Machine Learning. Big Data merupakan sejumlah besar data-data yang penting dan memiliki sejarah yang dikumpulkan, dan merupakan aset paling berharga dari setiap organisasi dan individu yang dimanfaatkan secara cerdas untuk keperluan bisnis agar dapat mendukung sebuah keputusan berdasarkan fakta yang ada daripada persepsi (Bhathal & Singh, 2019) [1]. Berbagai Industri saat ini memanfaatkan Big Data untuk mendukung bisnisnya. Data-data dapat berasal dari berbagai macamsumber seperti data dari sosial media, data dari berbagai sensor remote, data transaksi, serta log sistem yang dikumpulkanuntukmendukung aktivitas bisnis (Bhathal & Singh,

2019) [1]. Untuk mengumpulkan Big Data dan melakukan pengolahan terhadap Big Data, diperlukan sebuah infrastruktur mumpuni yang dapat digunakan untuk menjalankan sistemnya. Salah satu framework pengolahan Big Data yang saat ini sering digunakan oleh industri yaitu Hadoop.

Keamanan sistem merupakan hal yang perlu dipikirkan dalam komunikasi antar mesin, baik secara offline maupun online. Hal mendasar yang dapat dijadikan acuan dalam keamanan sistem yaitu Thriad of CIA (Confidentiality, Integrity, dan Availablity). Jika dijelaskan secara singkat, Confidentiality atau kerahasiaan mengacu kepada upaya agar informasi tidak dapat diakses oleh pihak yang tidak berwenang. Integrity mengacu pada sebuah metode atau cara untuk menjaga agar data atau informasi tidak dapat dimanipulasi atau diubah oleh pihak yang tidak berwenang. Dan konsep terakhir yaitu Availability, menjelaskan mengenai sebuah sistem harusterjaga ketersediaannya. Untuk itu penilitian ini bertujuan untuk melakukan implementasi dan analisis hadoop element availability berdasarkan daemon log monitoring, untuk melakukan langkah teknis mengenai bagaimana sebuah log monitoring system dapat menjaga ketersediaan atau availability dari Hadoop core services.

2. TINJAUAN PUSTAKA

2.1 Big Data

Big Data merupakan sejumlah besar data-data yang penting danmemiliki sejarah yang dikumpulkan, dan merupakan aset paling berharga dari setiap organisasi dan individu yang dimanfaatkan secara cerdas untuk keperluan bisnisagar dapat mendukung sebuahkeputusanyang berdasarkan fakta yangada daripada persepsi (Bhathal & Singh, 2019). Berdasarkan definisi, Big Data memiliki 3 karakteristik yang dapat menggambarkan sebuah Big Data, yaitu Volume, Velocity, dan Variety. Volume adalah suatu Big Data dapat dikenali berdasarkan ukuran datanya, Big Data memiliki ukuran data yang besar dari rata-rata ukuran data. Velocity adalahukurankecepatan daripertukaran data yang dialami oleh kumpulan data-data tersebut. Jadi dapat dikatakan Big Data selain memiliki Volume yang besartapikecepatan daripertukaran ataupun peningkatan datanya itu memiliki kecepatan diatas rata-rata. Kemudian yang ketigayaitu Variety, yaitu Big Data memiliki berbagai macam jenis mulai dari file berupa audio, video, text, log, data transaksi, data dari sensor, dan sebagainya.

2.2 Hadoop

Hadoop adalah software framework open-source yang digunakan untutk menyimpan, mengolah, memproses, dan menganalisis sejumlah besar data dari berbagai tipe data, yaitu data terstruktur, tidak terstruktur, dan semi-terstruktur (Yadav et al., 2019). Hadoop framework juga digunakan oleh berbagai macam perusahaan terkenal seperti Google, Yahoo, Amazon dan IBM untuk mendukung aplikasinya dalam menghadapi sejumlah besar data (Samet et al., 2019). Hadoop menyediakan sebuah file sistem terdistribusi yang disebut Hadoop Distributed File System (HDFS) yang menyimpan data di tiap node pada cluster. Selain itu Hadoop mengimplementasikan model pola komputasi yang disebut MapReduce, di mana aplikasi dibagi menjadi banyak pecahan di mana tiap pecahan dapat diproses di tiap node pada sebuah cluster. Framework Hadoop memberi ketersediaan dan pergerakan data dalam aplikasi serta bandwidth tinggi pada cluster. MapReduce dan HDFS dirancang agar kerusakan pada node ditangani secara otomatis oleh framework (Agarwal & Maurya, 2016).

2.3 HDFS

HDFS adalah singkatan dari Hadoop Distributed File System, yang membawa kemampuan tingkat tinggi untuk menyimpandata dengan volume yang besar. HDFS membagi-bagi unit data menjadi beberapa unit kecil data yang disebut dengan istilah block, dan

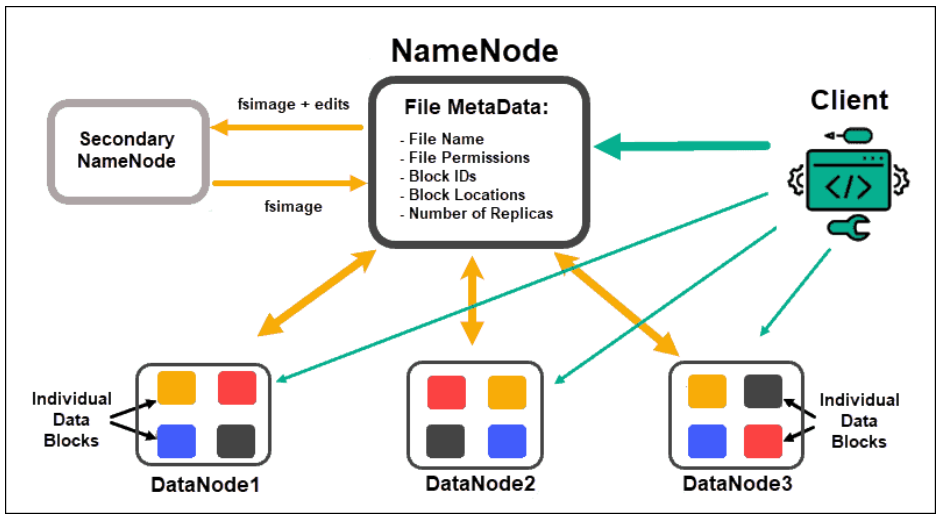
menyimpannya secara terdistribusi (Mishra, 2020). Hadoop Distributed File System (HDFS) merupakan file system berbasis Java yang terdistribusi pada Hadoop. Sebagai file system terdistribusi, HDFS berguna untuk menangani data dalam jumlah besar yang disimpan dan tersebar didalam banyak komputer yang berhubungan yang biasa disebut dengan cluster. HDFS pada Hadoop dapat diartikan sebagai file system yang menyimpan data tidak dalam satu Hard Disk Drive (HDD) atau media penyimpanan lainnya, tetapi data dipecah-pecah (file dipecah dalam bentuk block dengan ukuran 64 MB - bisa dikonfigurasi besarnya) dan disimpan tersebar dalam suatu cluster yang terdiri daribeberapa komputer.

2.3.1. Komponen HDFS

Sebagai file system terdistribusi HDFS memiliki komponen-komponen utama berupa NameNode, DataNode, dan Secondary NameNode.

a. NameNode

NameNode terdapat pada komputer yang bertindak sebagai master yang mengkoordinasikan DataNode untuk melakukan beberapa tugas (jobs). NameNode ini adalah pusat dari sistem berkas pada HDFS. Gambaran NameNode yang berada pada master sebagai pusat sistem berkas HDFS dapat dilihat pada Gambar dibawah ini:



Gambar 1. Namenode HDFS

NameNode membuat sistem direktori dari semua file yang ada di dalam sistem dan dapat mengetahui bagaimana file tersebut di pecah-pecah menjadi beberapa blocks data serta mengetahui nodes yang menyimpan blocks data tersebut.

b. DataNode

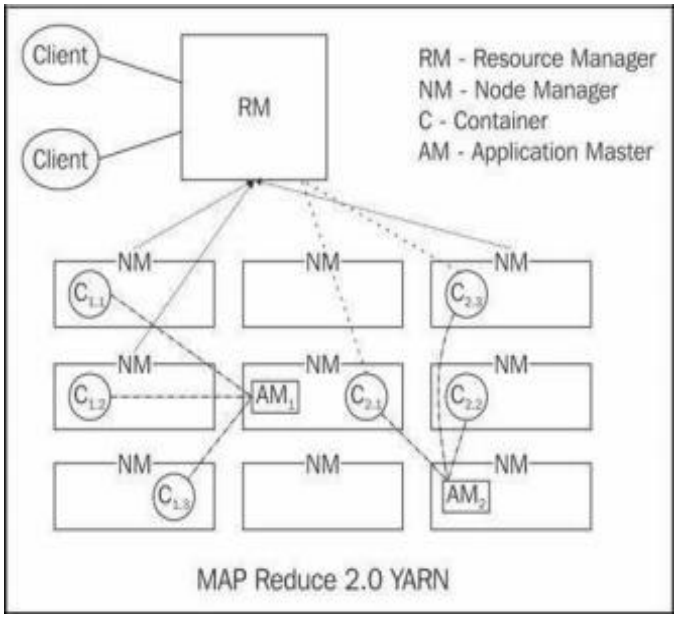
DataNode adalah salah satu komponen dari HDFS yang berfungsi untuk menyimpan dan mengambil kembali data pada slave node pada setiap permintaan yang dilakukan oleh NameNode. DataNode berada pada setiap slave node pada sebuah cluster yang telah dibuat. DataNode juga berfungsi untuk membaca dan menulis block pada HDFS ke file yang sebenarnya pada local filesystem. Sebagai contoh apabila user ingin membaca atau menulis file ke HDFS, file tersebut akan dipecah menjadi beberapa block, kemudian NameNode akan memberitahu dimana blocks tersebut berada sehingga DataNode dapat membaca dan menulis blocks tersebut ke file yang sebenarnya pada filesystem.

c. SecondaryNode

Secondary NameNode adalah daemon yang berfungsi melakukan monitoring keadaan dari cluster HDFS. Sama seperti NameNode, pada setiap cluster yang ada terdapat satu Secondary NameNode, yang berada pada master node. Secondary NameNode ini juga berfungsi untuk membantu dalam meminimalkan downtime dan hilangnya data yang terjadi pada HDFS. Secondary NameNode ini sering menimbulkan kesalahpahaman pengertian bahwa apabila NameNode down maka akan langsung digantikan oleh Secondary NameNode padahal Secondary NameNode ini hanya menyimpan informasi terbaru dari struktur direktoripada NameNode. Jadijikaterjadikegagalan yang dilakukan oleh NameNode maka dibutuhkan konfigurasi yang dilakukan oleh user untuk menjadikan Secondary NameNode sebagaiNameNode yang utama.

2.3.2. YARN

YARN (Yet Another Resource Negotiator) adalah resource manager multi guna yang dapat menjadwalkan dan mengatur siklus CPU dari cluster Hadoop ke aplikasi yang akan dijalankan. YARN merupakan peningkatan dari MapReduce yang dikenalkan pada Hadoop versi 2.0 yang dikenal juga sebagai MapReduce 2.0. Dalam Hadoop YARN dideskrisipsikan sebagai alat yang memungkinkan framework pemrosesan data lain dijalankan di Hadoop, sehingga Hadoop tidak hanya dapat menjalankan MapReduce. (deRoos et al., 2014).

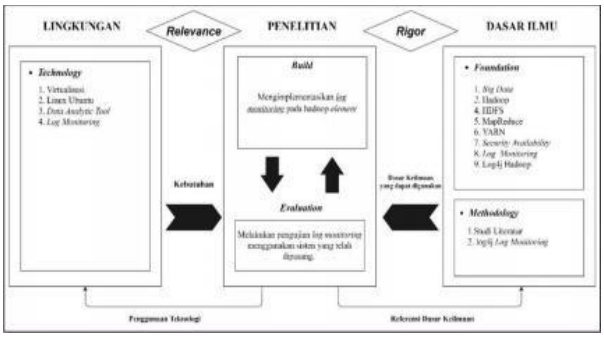


Gambar 2. Arsitektur YARN

3. METODELOGI PENELITIAN

3.1 Model Konseptual

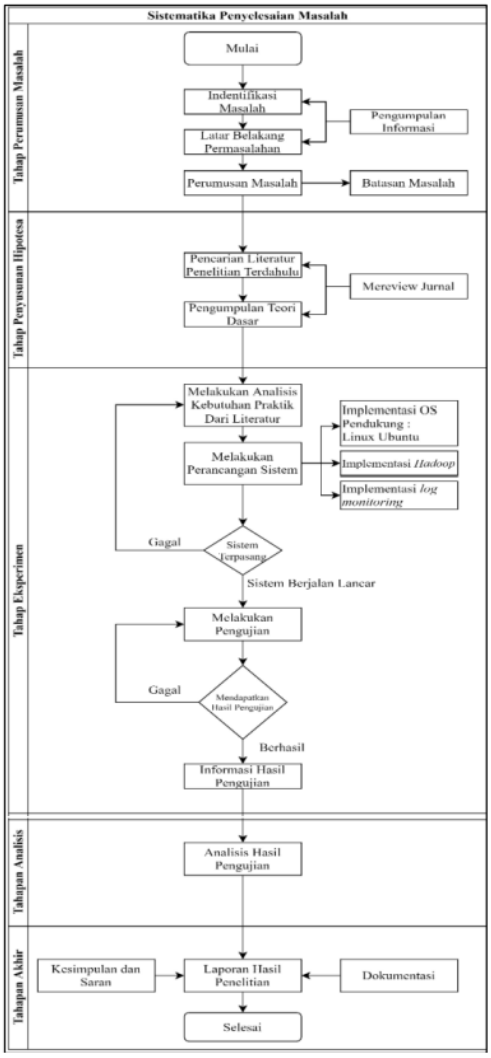
Model konseptual merupakan suatu gambaran logis dari suatu masalah yang digambarkan dalam rangkaiankonsep berdasarkan aspek hipotesis danteoritis, dan dapat digunakanuntukmembantu peneliti dalam merumuskan pemecahan masalah, perumusan solusi, serta evaluasu dari permasalahan yang ada. Model konseptual juga dapat membantu memperjelas masalah yang ada dan dapat mengidentifikasikan faktor-faktor masalah yang relevan, serta dulakukan dengan lebih mudah.



Gambar 3. Model Konseptual

3.2 Sistematika Penyelesaian Masalah

Dalam menyelesaikan permasalahan pada peneliatan ilmiah ini, dilakukan sebuah Langkah kerja ataupuntahapanyang terstruktur dan sistematisuntuk memecahkan permasalahan tersebut. Dengantahap ini dapat mempermudah dalam melakukan penelitian untuk memecahkan masalah yang telah ditemukan. Berikut tahapan yang dilakukan pada pelaksanaan penelitian ini dimulai dengan tahapan identifikasi masalah, kemudian melakukan review literatur, tahap perancangan, kemudian pengujian, lalu melakukan tahap analisis dantahap akhir penelitian.



Gambar 4. Sistematika penyelesaian masalah

3.3 Tahap Identifikasi Masalah

Pada tahap ini dilakukan pencarian masalah untuk membuat sebuah topik penelitian berdasarkan informasi-informasi yang telah dikumpulkan. Kemudian ketika masalahnya telah diemukan mencoba mencari latar belakang daripermasalahan yang telah ditemukan, kemudian merumuskan poin-poin permasalahanyang ingin selesaikan dan menentukan batasan-batasan masalahyang akan dibahas.

3.4 Tahap Penyusun Hipotesa

Pada tahapan ini dilakukan pencarian-pencarian jurnal ataupun sumber literatur lainnya mengenai topik terkait, untuk mendapatkan informasi mengenai penelitian-penenlitian terdahulu dan mengumpulkan dasar- dasarteorikeilmuanyang dibutuhkanuntuk menyusunhipotesa.

3.5 Tahap Eksperimen

Pada tahapan ini sudah mulai mempelajari sistem yang akan dibangun dan juga melakukan pemasangan sistem sesuai denganyang dibutuhkan. Dalamtahap perancangan initerdapatiterasiaktivitas jika sistem yang dalam tahap pemasangan masih gagal atau sumber referensi yang digunakan untuk penerapan sistem masih belum cukup. Tahapan perancangan baru dapat dikatakan selesai ketika sistem yang di pasang sudah berhasil fungsionalitasnya sesuai dengan yang diinginkan. selanjutnya melakukan pengujian untuk medapatkan hasil dari penelitian. Pada tahap ini baru pengujian dapat dikatakan selesai, ketika mendapatkan informasi yang dibutuhkanuntuk dilakukan sebuah analisa.

3.6 Tahap Analisis

Pada tahapan ini setelah mendapatkan data-data atau informasi dari hasil pengujian dan melakukan sebuah analisis yang sekiranya dengan hasil analisis tersebut dapat menyelesaikan permasalahan yang sebelumnya ditemukan pada tahap identifikasi masalah, dan tentunya sesuai dengan batasan masalah yang sebelumnyatelah ditentukan.

3.7 Tahap Akhir

Ditahap ini telah melakukan analisa kemudian analisa tersebut ditulis didalam laporan hasil penelitian dan membuat sebuah kesimpulan dan saran terkait masalah yang ada dan hasil temuan pada penelitianyang ditemukan.

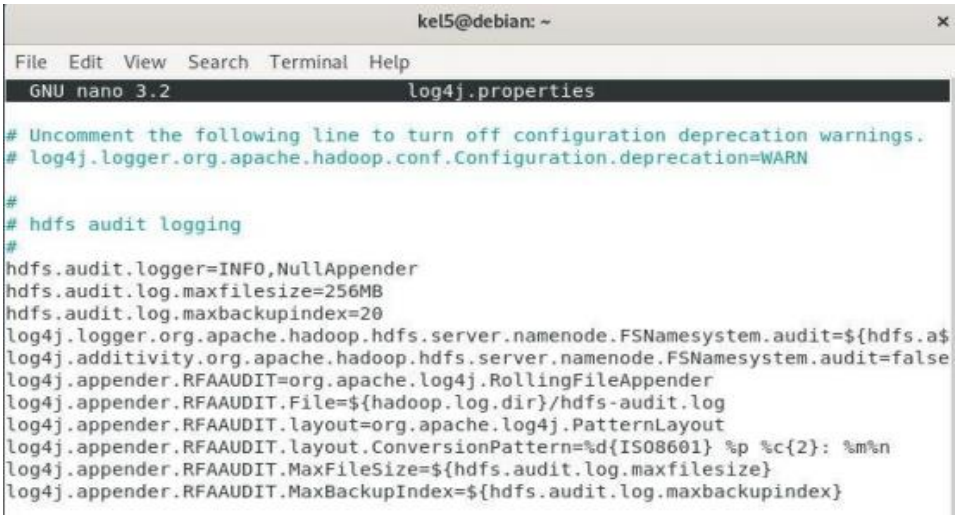
4. HASIL PENELITIAN

4.1 Hasil Implementasi

Implementasi yang dilakukan pada tahapan ini bereferensi ke beberapa buku, yaitu "Hadoop Security: Protecting Your Big Data Platform" ole Ben Spivey & Joey Echeverria, dan "Practical Hadoop Security" oleh Bhusan Lakhe.

4.1.1. Konfigurasi Log4j Monitoring HDFS

Untuk menjalankan pencatatan log, perlumemasukkan beberapa syntax konfigurasipada file, yaitu sebagai berikut:

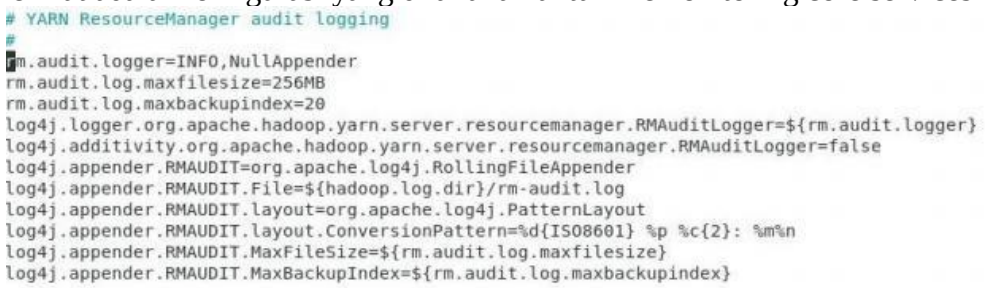


Gambar 5. Konfigurasi Log4j Monitoring HDFS

Konfigurasi juga perlu dilakukan untuk mencatat log dari YARN. Ada 2 file yang perlu dikonfigurasi untuk memonitoring log dari YARN yaitu sebagai berikut:

a. Resource Manager

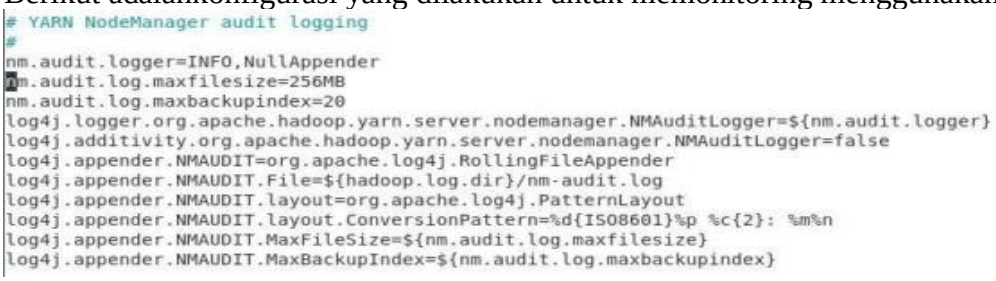
Berikut adalahkonfigurasi yang dilakukan untuk memonitoring core services YARN



Gambar 6. Resource Manager

b. Node Manager

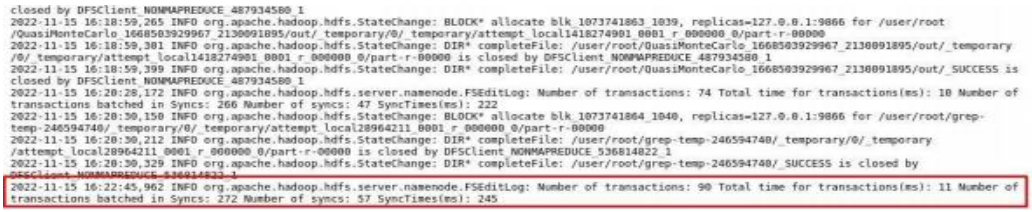
Berikut adalahkonfigurasi yang dilakukan untuk memonitoring menggunakan YARN



Gambar 7. Node Manager

4.2 Tampilan Hasil Monitoring log HDFS

a. NameNodeLog



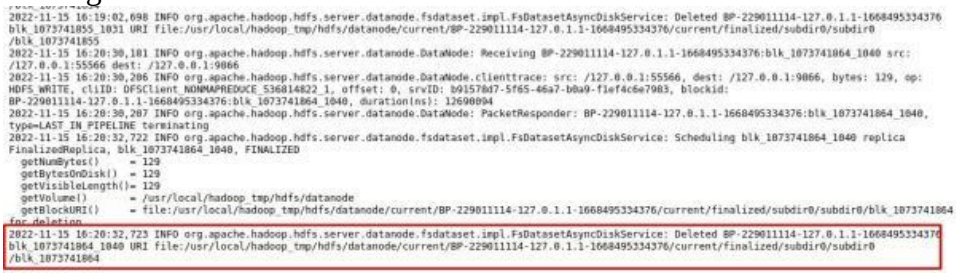
Gambar 8. Hasil Log HDFS 1

Pada gambar 8. ditampilkan mengenai informasi pencatatan log yang telah terimplementasai pada namenode. Dapat dilihat pada keterangan terakhir log tersebut menampilkan informasi mengenai:

- Time: 2022- 11- 15 | 16:22:45,962
- Level Log: INFO
- Job: Number of Transaction
- Time Total: Done Intermediate

Dari informasi diatas dapat diperoleh informasi bahwa job terakhir yaitu Number of transaction yang telahdilakukan sudah selesaipadajam 16:22:45,962 di tanggal 15 November 2022, dantelah ditutup prosesnya oleh client.

b. NameDataLog



Gambar 9. Hasil Log HDFS 2

Pada gambar 9 ditampilkan mengenai informasi pencatatan log yang telah terimplementasai pada Datanode. Dapat dilihat pada keterangan terakhir log tersebut menampilkan informasi mengenai:

- Time: 2022- 11- 15 | 16:20:32,723
- Level Log: INFO
- FsDatasetAsyncDisk Service: Deleted
- File: blk_ 1073741958

Dari informasi diatas dapat diperoleh informasi mengenai service DataNode yang pada waktu 2022-11-15 | 16:20:32,723 lognya tercatat telah melakukan aktivitas penghapusan block dari hasilproses terakhir.

4.3 Tampilan Hasil Monitoring log YARN

a. Resource Manager Log

Gambar 10. Hasil Log YARN 1

Pada gambar 10. ditampilkan mengenai informasi pencatatan log yang telah terimplementasai. Informasi log yang didapatkan telah sesuai dengan konfigurasi yang diberikan. Dapat dilihat pada keterangan terakhir log tersebut menampilkan informasi mengenai:

- Time: 2022-11-15 | 16:16:17,062
- Level Log: INFO
- RMNodeImpl: debian:42167
- Status: Running

Dari poin-poin tersebut didapatkan informasi mengenai YARN node yang berubah proses dari node yang baru jadi dan di running yang dijalankan pada waktu 2022-11-15 | 16:16:17,062.

b. Node Manager Log

Gambar 11. Hasil Log YARN 2

Pada gambar 11. ditampilkan mengenai informasi pencatatan log yang telah terimplementasai untuk monitoring log Node Manager. Informasi log yang didapatkan telah sesuai dengan konfigurasi yang diberikan. Dapat dilihat pada keterangan terakhir log tersebut menampilkan informasi mengenai:

- Time: 2022-11-15 | 16:26:10,624
- Level Log: INFO
- Service: ResourceLocalaization
- Cache Size Before Clean: 0
- Total Deleted: 0
- Public Deleted: 0
- Private Deleted: 0

Dari poin-poin tersebut didapatkan informasi mengenai Resource Localization yang membersihkan container yang telah selesai digunakan, dan dalam posisi standby untuk proses selanjutnya.

5. KESIMPULAN

Dari percobaanyang telah kami lakukan dan dari hasil penelitianyang sudah didapatkan, dapat kami simpulkan sebagai berikut:

1. Dengan mengimplementasikan monitoring log menggunakan log4j dapat memantau bagaimana Hadoop element bekerjaketika sistem Hadoop sedang melakukan proses.
2. Availability dari Hadoop Core system dapat dipantau dari daemon log saat menjalankan sebuah proses.

DAFTAR PUSTAKA

[1]. Dharminder Yadav, D. H. (2019). "Big Data Hadoop: Security and Privacy." INTERNATIONAL CONFERENCE ON ADVANCED COMPUTING AND SOFTWARE ENGINEERING.
[2]. Gurjit Singh Bhathal, A. S. (2019). "Big Data: Hadoop Framework Vurnerability, Security Issues and Attacks."
[3]. Karim Abouelmehdi, A. B.-H. (2016). "Big Data Emerging Issues: Hadoop Security."
[4]. Mohd Rehan Ghazi, D. G. (2015). "Hadoop, MapReduce and HDFS: A Developers Perspective." International Conference on Intelligent Computing, Communication & Convergence.
[5]. Refik Samet, A. A. (2019). "Big Data Security Problem Based on Hadoop." 4rd International Conference on Computer Science and Engineering, 2.
[6]. Goel, J. N., & Mehtre, B. M. (2015). "Vulnerability Assessment & Penetration Testing as a Cyber Defence Technology." Procedia Computer Science, 57, 710-715. https://doi.org/10.1016/j.procs.2015.07.458
[7]. Jain, J., & Ram Pal, P. (2017). "A Recent Study over Cyber Security and its Elements." International Journal of Advanced Research in Computer Science, 8(3), 791-793. https://search.proquest.com/docview/2082950830
[8]. Kumar, S. (2014). "Eco-Friendly Hadoop." 16(4), 52-56.
[9]. Mehta, T., & Mangla, N. (2016). "A Survey Paper on Big Data Analytics using Map Reduce and Hive on Hadoop Framework." National Conference on Recent Innovations in Science, Technology & Management (NCRISTM), February, 112-118.
[10]. Mishra, B. (2020). "Big Data Analysis Using Hadoop Map Reduce." International Research Journal of Computer Science, 07(05), 114-122. https://doi.org/10.26562/irjcs.2020.v0705.005
[11]. Shinde, P. S., & Ardhapurkar, S. B. (2016). "Cyber security analysis using vulnerability assessment and penetration testing." IEEE WCTFTR 2016 - Proceedings of 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare. https://doi.org/10.1109/STARTUP.2016.7583912
[12]. Tandel, M. K. B., & M. A. G. M. H. (2019). "A Review Paper on Big Data and Hadoop for Data Science." International Journal of Trend in Scientific Research and Development, 4(1), 1216-1221. https://www.ijtsrd.com/papers/ijtsrd29816.pdf%0Ahttps://www.ijtsrd.com/computer-

science/datamiining/29816/a-review-paper-on-big-data-and-hadoop-for-data-science/mr-ketan-bagade

[13]. Umar, M. A., & Zhanfang, C. (2019). "A Study of Automated Software Testing: Automation Tools and Frameworks." International Journal of Computer Science Engineering (IJCSE), 8(06), 217-225. (Subagya et al., n.d.)