



Plagiarism Checker X Originality Report

Similarity Found: 13%

Date: Saturday, September 07, 2024

Statistics: 264 words Plagiarized / 2108 Total words

Remarks: Low Plagiarism Detected - Your Document needs Optional Improvement.

Kombinasi **Algoritma K-Means Dan DBSCAN** Dalam Identifikasi Anomali Pada Data Log Server Rico Puji Irawan Universitas Budidarma, Sisingamangaraja, Sumatera Utara, Indonesia E-Mail : ricopujiirawan@gmail.com Sony Bahagia Sinaga STMIK Mulia Darma, Adam Malik, Labuhan Batu, Indonesia E-Mail : sonybahagia@gmail.com ABSTRACT Detecting anomalies in server **log data is a** crucial element of information system management and security.

This research seeks to develop a method for identifying anomalies by integrating two well-known clustering algorithms: K-Means and **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**. K-Means is effective at partitioning data into clusters based on average distances, while DBSCAN excels at detecting anomalies or noise in datasets without a distinct cluster structure. In this study, K-Means is employed for initial clustering of server log data to reveal general patterns and group similar data.

The results from K-Means clustering are then examined using DBSCAN to detect anomalies more accurately. Combining these two algorithms aims to enhance anomaly detection accuracy by leveraging the strengths of each approach. The research was performed on a server log dataset encompassing various server activities.

The effectiveness of this combined approach was assessed by comparing its anomaly detection performance against the individual K-Means and DBSCAN methods, as well as other anomaly detection techniques. Experimental results indicate that the K-Means and DBSCAN combination successfully improves anomaly detection rates by reducing both false positives and false negatives compared to using each algorithm independently.

Keywords: K-Means, DBSCAN, anomaly detection, clustering, server log data

PENDAHULUAN Data log server adalah catatan yang mendokumentasikan semua aktivitas yang terjadi pada sistem, termasuk interaksi pengguna, akses file, permintaan layanan, dan lainnya. Salah satu tantangan utama dalam menganalisis data log server adalah volume data yang besar dan kompleksitasnya.

Selain itu, pola-pola dalam data log server sering kali sulit diidentifikasi, dan aktivitas mencurigakan bisa tersembunyi di antara data yang tampak normal [1]. Anomali merujuk pada sesuatu yang tidak biasa, tidak normal, atau tidak sesuai dengan pola yang diharapkan dalam konteks tertentu. Anomali bisa disebabkan oleh berbagai faktor, termasuk serangan dari luar seperti peretas.

Penyebab anomali dapat meliputi lonjakan data, noise, pola konstan, dan drift. Anomali dalam data log server bisa menandakan adanya serangan atau pelanggaran keamanan yang memerlukan perhatian dan penanganan segera. Mengingat banyaknya aktivitas yang terjadi, sangat penting untuk dapat membedakan antara aktivitas normal dan yang mencurigakan[2].

Untuk mendeteksi anomali dalam data log server sebagai langkah pencegahan terhadap serangan, penelitian ini mengkombinasikan algoritma K-Means dan DBSCAN. K-Means adalah salah satu algoritma klustering yang paling umum digunakan. Klustering adalah proses membagi sekumpulan data menjadi subset-subset. Setiap subset, atau kluster, terdiri dari objek-objek yang mirip satu sama lain, tetapi berbeda dari objek-objek di kluster lainnya[3][4].

Algoritma DBSCAN dalam melakukan klustering dapat lebih efisien untuk membentuk kluster yang tidak teratur dibandingkan dengan algoritma CLARANS. Proses klustering ini didasarkan pada tingkat kedekatan atau kepadatan jarak antar objek dalam dataset, sehingga termasuk dalam kategori klustering berbasis kepadatan (density-based clustering) [5]. METODE 2.1

Log Server Log server mengawasi aktivitas di berbagai perangkat dan aplikasi dalam jaringan, memberikan wawasan kepada administrator sistem tentang kondisi lingkungan IT mereka. Banyak regulasi dan standar keamanan yang mewajibkan organisasi untuk menyimpan dan menganalisis log aktivitas. Log server membantu organisasi memenuhi kewajiban ini dengan menyimpan log secara terstruktur dan dapat diakses [6]. 2.2

Anomali Anomali adalah penyimpangan yang terjadi dalam model lingkungan. Penyimpangan ini bisa disebabkan oleh faktor internal sistem itu sendiri atau oleh faktor eksternal. Penyimpangan internal biasanya memerlukan perbaikan langsung dari pengembang, sedangkan penyimpangan yang disebabkan oleh lingkungan menjadi

perhatian bagi pengguna.

Meskipun deteksi anomali yang melibatkan manusia bisa sangat efektif, prosesnya memerlukan waktu yang lama. Oleh karena itu, dikembangkanlah sistem deteksi anomali otomatis yang tidak memerlukan intervensi manusia. Sistem ini didasarkan pada teknologi machine learning, data mining, dan algoritma statistik[7].

Algoritma K-Means Algoritma K-Means adalah metode dalam data mining yang digunakan untuk mengelompokkan data ke dalam satu atau lebih cluster. Dalam proses ini, data dengan karakteristik yang serupa dikelompokkan bersama dalam satu cluster, sementara data dengan karakteristik yang berbeda dikelompokkan ke cluster lainnya.

Tujuan utama dari klastering K-Means adalah menemukan titik data prototipe untuk setiap cluster; kemudian, semua titik data akan ditetapkan ke prototipe terdekat, membentuk sebuah cluster, langkah-langkah kerja algoritma K-Means [8]: Menentukan Jumlah Centroid Menetapkan Poin Data Menghitung Centroid Baru Mengulang Penetapan dan Penghitungan Centroid Penghentian Algoritma DBSCAN Algoritma DBSCAN (Density-Based Spatial Clustering of Applications with Noise) adalah metode klastering yang berbasis kepadatan, digunakan untuk mengelompokkan data dengan cara menemukan area-area dengan kepadatan tinggi dan memisahkan area-area dengan kepadatan rendah sebagai noise atau anomali. DBSCAN sangat efektif dalam mendeteksi klaster yang tidak berbentuk bulat dan menangani data dengan noise.

Berikut adalah langkah-langkah dasar dalam algoritma DBSCAN [9][10]: Inisialisasi Identifikasi Titik Awal Klasifikasikan Titik Bentuk Klaster Ulangi Proses Hasil Klastering Setelah algoritma selesai, dataset akan terbagi menjadi beberapa klaster, dan beberapa titik data mungkin diklasifikasikan sebagai noise jika tidak memenuhi syarat untuk menjadi bagian dari klaster manapun.

DBSCAN unggul dalam mengidentifikasi bentuk klaster yang tidak teratur dan menangani noise, serta tidak memerlukan jumlah klaster yang harus ditentukan sebelumnya, berbeda dengan algoritma klastering lain seperti K-Means. HASIL DAN PEMBAHASAN Identifikasi anomali dalam data log server adalah proses untuk menemukan aktivitas atau pola yang tidak biasa atau mencurigakan dalam log server, yang mungkin menandakan adanya masalah keamanan, kinerja, atau operasional. Analisis ini sangat penting untuk mendeteksi dan mencegah insiden yang tidak diinginkan.

Langkah pertama dalam analisis anomali adalah mengumpulkan data log dari berbagai sumber, seperti server web, aplikasi, database, dan sistem operasi. Data log ini

mencakup informasi seperti waktu kejadian, alamat IP, pesan kesalahan, dan aktivitas pengguna. Karena data log sering kali tidak terstruktur dan berantakan, preprocessing diperlukan untuk membersihkan dan mengatur data.

Setelah data dibersihkan, langkah berikutnya adalah ekstraksi fitur, yang melibatkan identifikasi atribut atau karakteristik penting dari data log yang dapat digunakan untuk analisis. Untuk mendeteksi anomali dalam data log server, dilakukan kombinasi antara algoritma K-Means dan DBSCAN. K-Means Clustering dan DBSCAN dilakukan pada variabel akses log yang dipilih. Berikut tahapan yang dilewati dalam pengerjaan penelitian ini.

Data preprocessing Data akses log dengan ekstensi log perlu diuraikan terlebih dahulu untuk memudahkan sistem dalam memproses data tersebut. Ekstraksi dan pemilihan fitur Analisis K-Means dan DBSCAN diterapkan pada beberapa variabel yang dipilih melalui tahapan Ekstraksi dan Pemilihan Fitur. Proses ekstraksi fitur ini bertujuan untuk mengubah data menjadi fitur yang sesuai dengan model.

Dalam penelitian terkait ekstraksi fitur pada data log server web, ditemukan bahwa terdapat 30 fitur yang bisa diambil dari sebuah file log untuk mendeteksi serangan. Pengolahan dan Analisis Data Setelah proses seleksi fitur untuk K-Means dan DBSCAN selesai, data yang telah diproses akan siap untuk dianalisis menggunakan Hadoop dan Spark. Proses ini melibatkan penggunaan dua aplikasi terkait: Hadoop Distributed File System (HDFS) dan Spark.

HDFS berfungsi sebagai media penyimpanan terdistribusi antar komputer, memungkinkan Hadoop dan Spark untuk melakukan pengolahan data secara terdistribusi. Dalam eksperimen ini, digunakan dua komputer: satu sebagai master node dan satu sebagai slave node. Master node bertanggung jawab mengatur pengolahan data secara terdistribusi, sementara slave node berfungsi sebagai komputer pekerja.

Untuk tujuan penelitian ini, evaluasi model dilakukan dengan menggunakan metrik yang sama seperti pada penelitian sebelumnya, yaitu menghitung tingkat false-positive, false-negative, sensitivitas model, spesifisitas model, tingkat klasifikasi, dan presisi hasil model. Tabel 1. Fitur Terpilih Variabel Penelitian

No	Feature Name
1	Panjang dari request line
2	Panjang dari IP Host
3	Panjang header dari user-agent
4	Besarnya ukuran byte setiap request
5	Panjang header dari referer
6	Jumlah okurensi garis miring
7	Jumlah okurensi titik
8	Jumlah okurensi persentase

Clustering Dari fitur yang telah dipilih, dilakukan ekstraksi dengan cara mengukur panjang dari variabel tertentu dan menghitung frekuensi kemunculan karakter tertentu.

Selain itu, variabel-variabel yang tidak diperlukan juga dihapus. / Gambar 1. Alur Kerja Fisik Pengolahan data pada Spark tidak sepenuhnya bergantung pada jumlah komputer, melainkan pada jumlah node yang digunakan. Dalam Spark, terdapat satu master node dan beberapa slave node. Dalam penelitian ini, digunakan dua slave node.

Master node, yang dikendalikan oleh driver, bertanggung jawab untuk mengalokasikan tugas. Driver meneruskan tugas-tugas tersebut ke cluster manager, dalam hal ini adalah Spark sebagai cluster manager. Cluster manager kemudian mendistribusikan tugas dari driver ke slave node untuk diproses.

Hasil dari pengolahan data ini mencakup pusat kluster dan model yang dapat diakses. / Gambar 2. Hasil Running Hasil Pengolahan data Hasil dari pengolahan data diambil dari akses log 8 variabel yang menghasilkan dua pusat cluster yaitu [11.944,147 ; 58,0004 ; 13,257 ; 101,08 ; 25,342 ; 4,685 ; 1,997 ; 0,147] serta [1.729.124,25 ; 57,677 ; 13,38 ; 109,8 ; 39,744 ; 5,123 ; 2,066 ; 1,465] / Gambar 3.

Hasil Running Menggunakan Algoritma Hasil dari menjalankan skrip menunjukkan bahwa dari 5.700.375 catatan log akses dengan 8 variabel, dihasilkan dua pusat kluster sebagai berikut: [751,238 ; 67,983 ; 13,099 ; 89,99 ; 21,475 ; 5,605 ; 1,8864 ; 0,1387] dan [3,256 ; 4,469 ; 13,469 ; 115,906 ; 30,55 ; 3,46 ; 2,144 ; 0,1633].

Untuk menguji efektivitas model, dihitung serangkaian metrik, termasuk false positive rate (FPR), false negative rate (FNR), sensitivitas, presisi, tingkat klasifikasi, dan spesifisitas. Pada algoritma K-Means Clustering, nilai FPR yang diperoleh adalah 0,0191%, sementara pada DBSCAN mencapai 53,58%. Semakin tinggi nilai FPR, semakin sensitif model dalam mendeteksi aktivitas jaringan.

Namun, nilai FPR yang tinggi juga menunjukkan bahwa banyak aktivitas normal mungkin terdeteksi sebagai anomali. Tabel 1. Confusion Matriks _Predicted value _ _True value _ _P _N _ _P _76 _295 _ _N _53.390 _46.238 _ _Hasil dari metode clustering K-Means menghasilkan pembagian data ke dalam tiga kluster, di mana setiap pusat centroid digunakan untuk menentukan kategori nilai risiko serangan. Hasil dari proses clustering dengan metode K-Means berdasarkan dataset dapat dilihat pada Tabel 2. Tabel 2.

Hasil Clustering Cluster _Jumlah Data _ _Cluster 1 _306743 _ _Cluster 2 _1685 _ _Cluster 3 _353 _ _Kelihatan bahwa Cluster 1 memiliki jumlah data terbesar dibandingkan kluster lainnya, yaitu total 306.743 data. Sementara itu, Cluster 2 terdiri dari 1.685 data, dan Cluster 3 memiliki jumlah data terkecil, yaitu 353 data. Titik pusat centroid akhir dari hasil clustering ini ditunjukkan dalam Tabel 3. Tabel 3.

Pusat Centroid Akhir Cluster _Prioritas _Prot _Frekuensi _ Cluster 1 _2 _4 _5 _ Cluster 2 _3.0652819 _4 _1 _ Cluster 3 _3.25212465 _1.99150142 _1 _ KESIMPULAN Berdasarkan hasil penelitian dan analisis mengenai pengamanan pesan teks, kesimpulan yang dapat diambil adalah sebagai berikut: Dari 308.781 data yang diklaster, serangan berhasil dikategorikan dengan baik.

Terdapat tiga klaster, di mana Cluster 2 dan Cluster 3 termasuk dalam kategori risiko serangan rendah, dengan total 2.038 data. Sebaliknya, Cluster 1 tergolong dalam kategori risiko serangan menengah, dengan jumlah data sebanyak 306.743. Nilai silhouette score untuk klastering ini adalah 0,999.

Hasil identifikasi menunjukkan protokol-protokol yang sering menjadi target serangan serta kategori risiko serangan yang terdeteksi dalam data log server. Metode K-Means biasanya digunakan dalam berbagai penelitian lain untuk mengelompokkan data log. Wazuh berkontribusi pada pengembangan Dasbor Sistem Pencatatan Log yang lebih optimal, yang terbukti efektif dan bermanfaat untuk pengelolaan log terpusat (CLM).

Informasi log yang dihasilkan dalam format yang mudah dipahami mendukung pengembangan CLM dengan baik. UCAPAN TERIMKASIH Penulis mengucapkan terimakasih ke pembimbing yang telah banyak memberikan masukan dalam penyelesaian penelitian ini. Ucapan terima kasih juga penulis ucapkan ke program studi teknik informatika yang telah memberikan kesempatan untuk melakukan penelitian ini serta ucapan terimakasih juga penulis ucapkan kepada orangtua yang telah memberikan dukungan. DAFTAR PUSTAKA [1] Gustientiedina, et.al, 2019, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan Pada RSUD Pekanbaru", Jurnal Nasional Teknologi dan Sistem Informasi, Vol 05, No.

01 [2] Wahyu Sudrajat, et.al, 2022, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokkan UMKM Menggunakan Rapidminer", Jurnal JUPITER, Vol. 14, No. 1 [3] Rozi Kesuma Dinata, et.al, 2020, "Analisis K-Means Clustering Pada Data Sepeda Motor", Informatic Journal, Vol. 5, No. 1 [4] Devi Fitriana, et.al, 2021, "Implementasi Algoritma DBScan Dalam Pengambilan Data Menggunakan Scatterplot", Jurnal Techno Xplore, Vol. 6, No. 2 [5] Betha Nurina Sari, et.al, 2019, "Penerapan Clustering DBScan Untuk Pertanian Padi di Kabupaten Karawang", Jurnal JIKO, Vol. 4, No. 1 [6] F. Pangestu, et.al, 2023, "Penerapan Algoritma K-Means Untuk Mengklasifikasi Data Obat", Jurnal SISFOKOM, Vol. 12, No.

1 [7] Mustofa, 2019, "Penerapan Algoritma K-Means Clustering Pada Karakter Permainan Multiplayer Online Battle Arena", Jurnal Informatika, Vol. 6, No. 2 [8] B.S. Ashari, et.al,

2019, "Perbandingan Kinerja K-Means dengan DBScan Untuk Metode Clustering data Penjualan Online Retail", Jurnal Siliwangi, Vol. 5, No. 2 [9] T.I. Hermanto, et.al, 2020, "Analisis Data Sebaran Bandwidth Menggunakan Algoritma DBScan Untuk Menentukan Tingkat Kebutuhan Bandwidth Di Kabupaten Purwakarta", Jurnal RABIT, Vol.

5, No. 2 [10] Mustika Putri, et.al, 2021, "Komparasi DBSCAN Dan K-Means Clustering Pada Pengelompokan Status Desa di Jawa Tengah Tahun 2020", Jurnal Matematika, Statistika, Vol. 17, No. 3 [11] D. P. Isnarwaty, et.al, 2019, "Text Clustering Pada Akun Twitter Layanan Ekspedisi JNE, J&T dan Pos Indonesia Menggunakan Metode Density-Based Spatial Clustering Of Applications With Noise (DBSCAN) Dan K-Means

INTERNET SOURCES:

<1% - <https://santika.upnjatim.ac.id/submissions/index.php/santika/article/view/195>
<1% - <https://netman.aiops.org/wp-content/uploads/2021/09/logstudy.pdf>
1% - <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>
<1% - <https://ieeexplore.ieee.org/document/10409931>
<1% -
<https://blog.kazee.id/potensi-data-dalam-big-data-tantangan-dan-peluang-data-analys>
t
<1% - <https://rifqimulyawan.com/kamus/anomaly-detection/>
<1% -
<https://www.ad-ins.com/id/our-story/kisah-adins/memahami-kebocoran-data-jenis-penyebab-dan-dampaknya/>
<1% -
<https://www.telkomsel.com/enterprise/insight/blog/mengenal-sistem-deteksi-dan-respons>
ons
<1% - <https://medium.com/@myskill.id/clustering-analysis-26bc2e73169d>
<1% - <http://repository.unimus.ac.id/3870/3/13.%20BAB%20II.pdf>
<1% - <https://ejournal.akakom.ac.id/index.php/jiko/article/download/178/pdf>
<1% - <https://www.asdf.id/sistem-informasi-manajemen-fungsi-contoh-dan-manfaat/>
<1% - <https://statsidea.com/id/metode-k-means-clustering-dalam-analisis-data/>
<1% - <https://www.trivusi.web.id/2022/06/algoritma-kmeans-clustering.html>
<1% - <https://www.baeldung.com/cs/dbscan-algorithm>
<1% - <https://dqlab.id/machine-learning-keamanan-data-dengan-deteksi-anomali>
<1% -
<https://akademia.co.id/panduan-metode-analisis-data-jenis-jenis-langkah-dan-tantangan/>
<1% - <https://www.ibm.com/id-id/products/z-anomaly-analytics>
1% -

https://www.academia.edu/70878545/Deteksi_Intrusi_Jaringan_dengan_K_Means_Clustering_pada_Akses_Log_dengan_Teknik_Pengolahan_Big_Data

<1% -

<https://pdfs.semanticscholar.org/8a50/a60661e234fed92d23b298d3efd5a3b050c0.pdf>

<1% - <https://www.ibm.com/think/insights/hadoop-vs-spark>

<1% - <https://sismatik.nusaputra.ac.id/index.php/sismatik/article/download/32/29/161>

<1% - <https://prosiding.stis.ac.id/index.php/semnasoffstat/article/download/706/110>

<1% -

https://www.academia.edu/81193209/Analisis_Data_Log_Honeypot_Menggunakan_Metode_K_Means_Clustering

<1% -

<http://digilib.unimed.ac.id/id/eprint/21492/10/13.%20NIM%201123113028%20BAB%20V.pdf>

<1% -

<https://sumsel.tribunnews.com/2023/12/14/4-contoh-ucapan-terima-kasih-dalam-jurnal-skripsi-singkat-dan-penuh-rasa-syukur?page=all>

<1% - <https://ejurnal.co.id/contoh-ucapan-terimakasih-pada-jurnal>

1% - <https://scholar.google.com/citations?user=9fCM3tYAAAAJ>

<1% - <https://jurnal.unej.ac.id/index.php/INFORMAL/issue/view/947>

<1% - <https://ejurnal.stmik-budidarma.ac.id/index.php/mib/article/view/1784>

<1% - <https://sinta.kemdikbud.go.id/journals/profile/3712>

<1% - <https://journal.lppmunindra.ac.id/index.php/STRING/article/view/11094>

<1% - <https://ejournal.akakom.ac.id/index.php/jiko/article/view/796>

1% - <http://ejurnal.tunasbangsa.ac.id/index.php/jsakti/article/view/501>

<1% -

https://www.academia.edu/84222162/Comparison_of_DB_Scan_and_K_means_clustering_for_grouping_the_village_status_in_central_Java_2020

1% - <https://repository.its.ac.id/64455/>

1% - <https://ejournal.uin-suska.ac.id/index.php/sitekin/article/download/12469/6410>