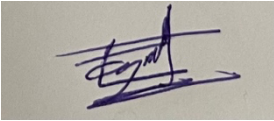




**Faculty of Computing
Department of Data Science
Group Assignment**

INDEX NUMBER		STUDENT NAME	
24960		MTN Gunawardhana	
24077		CS Wickramarachchi	
23144		CD Malaviarachchi	
YEAR OF STUDY AND SEMESTER		3 rd year 1 st semester	
MODULE CODE		MODULE NAME	Data Warehouse
MODULE LECTURER	Ms. Nethmi Weerasingha	SUBMISSION DATE	29 th Apr 2024
Declaration: I certify that I have not plagiarized the work of others or participated in unauthorized collusion when preparing this assignment.			
Signature of the Group Leader:			
Date: 29 th Apr 2024			
For office purpose only:			
GRADE / MARK			
COMMENTS			

CyberGuard Insights by hSenid Software

International using data mining and warehousing techniques



Content



1. Abstract
2. Company Profile
3. Problem Statement
4. General Solution
 - 4.1 Data warehousing and data mining
 - 4.2. Data mining techniques
 - 4.2.1. Classification and prediction
 - 4.2.2. Clustering
 - 4.2.3. Outlier analysis
5. Introduction to intrusion detection systems
6. Data mining for intrusion detection
7. A data architecture for IDS
 - 7.1. A software architecture and data model for intrusion detection
 - 7.2. Data modeling for historical data analysis using STAR schema
 - 7.3. Support for high speed drill down queries and detection of attacks/virus/worms
 - 7.4. Feature extraction from network traffic data
 - 7.5. Help the security officer for forensic analysis
8. System implementation
9. Conclusions

1. Abstract

This paper explores the application of advanced data mining and data warehousing techniques to enhance the efficacy and usability of Intrusion Detection Systems (IDS), focusing on the context of hSenid Software International, a prominent software solutions provider catering to telecom, financial, and enterprise markets globally. The current landscape of IDS systems often lacks robust support for historical data analysis and summarization, posing significant challenges in detecting and mitigating cyber threats effectively. Leveraging a multi-dimensional data model and star schemas, this study proposes innovative approaches to model network traffic and alerts, facilitating comprehensive network security analysis and the detection of denial of service (DoS) attacks.

The proposed data model accommodates heterogeneous data sources, including firewall logs, system calls, and net-flow data, enabling seamless integration and correlation of disparate information for enhanced threat detection capabilities. Through the implementation of advanced data mining algorithms and optimized query response times, our approach promises up to two orders of magnitude improvement in performance compared to existing IDS systems.

Furthermore, the paper showcases the successful deployment of a prototype system utilizing these techniques at Army Research Labs, highlighting its efficacy in detecting intrusions and conducting historical data analysis to generate insightful trend reports. By addressing the limitations of current IDS systems and offering a scalable and agile solution, hSenid aims to bolster its reputation as a trusted provider of cybersecurity solutions, thereby better serving the evolving needs of its diverse clientele across the globe.

Keywords

Data warehouse . OLAP . Data mining and analysis . Computer security . Intrusion detection



2. Company Profile:



hSenid Software International

Founded in 1997, hSenid Software International began its journey as a software company in Sri Lanka. Over the years, it has evolved into a multi-national corporation specializing in providing innovative solutions for the telecom, financial, and enterprise markets. hSenid's core competencies lie in human resource applications, mobile solutions, and outsourcing services, making it a versatile player in the global tech industry.

What They Do:

hSenid is a leading application and service provider, offering a wide range of solutions tailored to meet the needs of its diverse customer base. With a focus on telecom, financial, and enterprise sectors, hSenid develops cutting-edge software applications, mobile solutions, and outsourcing services. Its portfolio encompasses human resource management systems, mobile applications, telecom software solutions, and complex global outsourcing projects.

Customer Base:

hSenid serves over 1200 clients globally across 18 industries in more than 35 countries. Its clientele includes prominent companies such as Dialog, Etisalat, Celinco Life, Huawei, British Council, Keells, Fashion Bug, Brandix, and Arpico, among others. With a diverse customer base spanning various industry, hSenid has established itself as a trusted partner for businesses seeking innovative software solutions.

Operational Model:

Operating from multiple locations worldwide, hSenid has a truly multinational presence. With offices in the USA, Australia, Singapore, Kenya, India, Bangladesh, and its headquarters in Sri Lanka, hSenid operates as a global entity. Its overseas branches enable it to cater to the needs of clients across different regions, while its Sri Lankan roots ensure a strong local presence and understanding of the domestic market.

Product Portfolio:

hSenid offers a comprehensive range of products and services to address the diverse needs of its customers. From human resource management systems to mobile applications, telecom

software solutions, and outsourcing services, hSenid's product portfolio reflects its commitment to innovation and excellence. Its solutions are designed to empower businesses, streamline operations, and drive growth in an increasingly competitive landscape.

Technological Expertise:

Leveraging cutting-edge technologies, hSenid develops innovative solutions that deliver tangible value to its clients. With expertise in areas such as mobile technology, cloud computing, artificial intelligence, and data analytics, hSenid remains at the forefront of technological innovation. By embracing emerging technologies, hSenid ensures that its solutions are future-proof and capable of meeting the evolving needs of its customers.

Corporate Initiatives:

In addition to its core business operations, hSenid is actively involved in corporate social responsibility initiatives and sustainable practices. Through initiatives such as LOFT1024, an incubator for start-ups, and Mount Havana, a sustainable holiday home, hSenid demonstrates its commitment to supporting entrepreneurship and environmental conservation. These initiatives reflect hSenid's ethos of community mindfulness and social responsibility.

3. Problem Statement

hSenid Software International, a prominent provider of software solutions across various sectors including telecom, finance, and enterprise, faces a critical challenge in bolstering the efficiency and usability of its Intrusion Detection Systems (IDS). In today's cybersecurity landscape, where organizations encounter an array of cyber threats, the limitations of existing IDS systems present a significant obstacle. One notable deficiency is the absence of support for historical data analysis and summarization within these systems.

Without the capability to delve into historical data, security analysts at hSenid and its clientele struggle to discern emerging threats and patterns effectively. This deficiency results in delayed responses to potential security breaches and leaves systems vulnerable to exploitation. Despite hSenid's diverse range of cutting-edge products and services, the absence of robust historical data analysis in its IDS solutions undermines the overall cybersecurity posture of the organization and its customers.

Recognizing the paramount importance of cybersecurity in safeguarding sensitive data and infrastructure, hSenid acknowledges the imperative to address these shortcomings. Enhancing the performance and usability of IDS is crucial not only for hSenid's reputation as a trusted provider but also for the protection of its global customer base. As cyber threats continue to evolve and proliferate, the need for advanced IDS systems that can effectively analyze historical data to identify emerging threats becomes increasingly pressing.

By addressing the challenge of historical data analysis and summarization within its IDS solutions, hSenid endeavors to fortify its position as a leader in the cybersecurity domain. Through innovative approaches and robust technological advancements, hSenid aims to empower its customers with resilient and responsive cybersecurity solutions that can adapt to the ever-changing threat landscape. Ultimately, by enhancing the performance and usability of its IDS offerings, hSenid seeks to mitigate risks, safeguard critical assets, and instill confidence in its clientele.

The primary issue faced by hSenid is the need to enhance the efficiency and effectiveness of its Intrusion Detection Systems (IDS) to better detect and mitigate cyber threats. The current state-of-the-art IDS systems lack the capability to conduct comprehensive historical data analysis. This limitation restricts the system's ability to identify subtle patterns and anomalies that may indicate malicious activities.

For example, without historical data analysis, the IDS may overlook trends or recurring patterns in network traffic that could signify an ongoing attack or unauthorized access. By incorporating historical data analysis capabilities into the IDS, hSenid can improve its ability to identify and respond to potential threats in real-time, thereby enhancing the overall security posture of its clients.

Moreover, hSenid faces the challenge of integrating heterogeneous data sources, including firewall logs, system calls, and net-flow data, into a unified data model for more robust threat

detection. These disparate data sources provide valuable insights into different aspects of network activity and potential security incidents. However, the diverse formats and structures of these data sources make it challenging to correlate and analyze the information effectively.

By integrating these heterogeneous data sources into a unified data model, hSenid can gain a more comprehensive view of network activity and potential security threats. This unified approach enables the IDS to correlate information from multiple sources and identify complex attack patterns that may span across different layers of the network infrastructure. Additionally, it allows hSenid to leverage advanced data mining and machine learning techniques to identify emerging threats and adapt its detection mechanisms accordingly.

Overall, by addressing these challenges and enhancing the efficiency and effectiveness of its IDS, hSenid can better protect its clients' infrastructure and data from cyber threats. This proactive approach to cybersecurity strengthens hSenid's reputation as a trusted provider of innovative and robust security solutions in the rapidly evolving threat landscape.

The critical issue of time-to-market for cybersecurity solutions at hSenid underscores the pressing need for agile and responsive Intrusion Detection Systems (IDS). With cyber threats evolving at an alarming rate, the demand for adaptive IDS systems capable of detecting emerging threats in real-time is paramount. However, existing IDS solutions often fall short in agility and scalability, leading to delayed responses and heightened risk exposure for hSenid's clientele.

To address this challenge, hSenid must focus on developing and implementing advanced data mining and data warehousing techniques within its cybersecurity solutions. By harnessing the power of these techniques, hSenid can overcome the limitations of traditional IDS systems and significantly enhance the performance and usability of its cybersecurity offerings.

Data mining techniques enable hSenid to analyze vast volumes of data from diverse sources, such as network traffic logs, system event logs, and threat intelligence feeds, to identify patterns and anomalies indicative of malicious activities. By leveraging machine learning algorithms and statistical analysis, hSenid can detect and mitigate cyber threats in real-time, empowering its customers to respond swiftly to evolving security threats.

Additionally, data warehousing techniques enable hSenid to store and organize historical data in a structured and accessible manner, facilitating retrospective analysis and trend identification. By building comprehensive data warehouses that integrate disparate data

sources, hSenid can provide its customers with valuable insights into past security incidents and emerging threat trends, enabling proactive threat mitigation strategies.

By addressing the challenge of time-to-market with advanced data mining and data warehousing techniques, hSenid aims to bolster its reputation as a trusted provider of cybersecurity solutions. By delivering agile and responsive IDS systems that adapt to the evolving threat landscape, hSenid can better serve the dynamic needs of its global customer base and stay ahead of the curve in the ever-changing cybersecurity landscape.



4. General Solution:

4.1 Data warehousing and data mining

we focus on harnessing data warehousing and data mining techniques to facilitate knowledge discovery in the realm of cybersecurity. Much like in other domains such as retail, finance, and bio-informatics, these methodologies play a pivotal role in analyzing data collected from distributed information repositories within our network infrastructure.

For instance, imagine a scenario where a telecommunications service provider seeks to gather network usage data to discern usage patterns, detect fraudulent activities, optimize resource allocation, and enhance service quality. Through the application of data warehousing and data mining techniques, our project endeavors to automate the analysis of this data, distill valuable insights, and forecast future trends. By aggregating and harmonizing data from diverse sources, our approach enables the seamless integration of information for comprehensive analysis. This empowers us to unveil hidden patterns, anomalies, and trends within the network data, thereby empowering organizations like ours to proactively address cybersecurity threats, optimize network performance, and enhance overall operational efficiency.

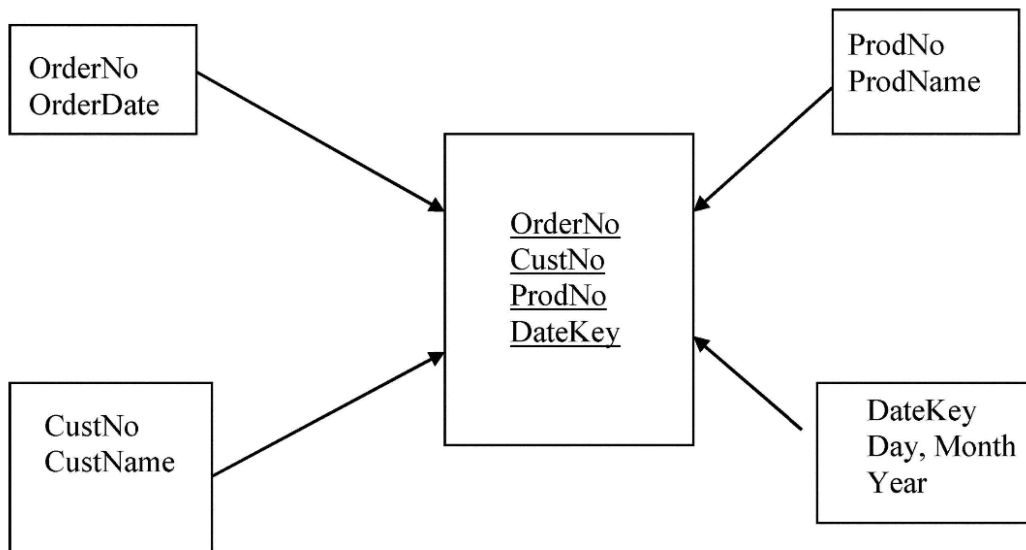


Fig. 1 A star schema for the sales information

Data Warehouse uses a data model that is based on a multidimensional data model. Data cubes, as this concept is frequently called, enable multi-dimensional modelling and viewing of data. Dimensions are the several viewpoints that an organization is interested in for a given entity. To track sales according to various dimensions, including time, branch, and location, a store can, for instance, establish a sales data warehouse. One example of a major theme that the data model is structured around is "sales." A fact table is another name for this main idea. We can think of facts as quantities that we wish to use to analyze relationships between dimensions because they are numerical measures. Facts include things like money sold, units sold, and so forth. The names of the facts and keys to all associated dimension tables are contained in the fact table. Relational database design frequently use the entity-relationship data paradigm. But a data warehouse isn't the right place for that kind of schema. A succinct, topic-oriented schema is necessary for a data warehouse in order to enable online data analysis. A multidimensional model is the most widely used data model for a data warehouse. A star schema is one example of such a scheme. The following makes up the star schema.

1. A large central table (fact table) containing the bulk of data.
2. A set of smaller *dimension tables* one for each dimension.



The schema resembles a star, with the dimension tables displayed in a radial pattern around the central fact table. An example of a sales table and the corresponding star schema is shown in the Fig. 1. Besides *OderNo*, *CustNo*, *ProdNo* and *Date*, the sales table will have an attribute *total sales amount* that corresponds to total sales. For each dimension, the set of associated values can be structured as a hierarchy. For example, cities belong to states and states belong to countries.

Similarly, dates belong to weeks that belong to months and quarters/years. The hierarchies are shown in Fig. 2.

In data warehousing, there is a distinction between a data warehouse and a data mart. A data warehouse collects information about subjects that span the entire organization such as customers, items, sales and personnel. Therefore, the scope of a data warehouse is *enterprise wide*. A data mart on the other hand is a subset of the data warehouse that focuses on selected

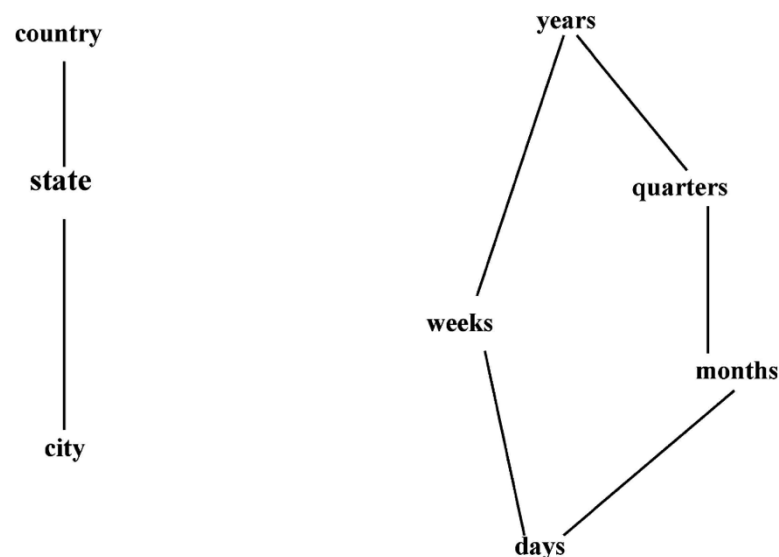


Fig. 2 Concept hierarchy

subjects and is therefore limited in size. For example, there can be a data mart for sales information another data mart for inventory information.

Business executives use the data collected in a data warehouse for data analysis and make strategic business decisions. *Data Mining* supports knowledge discovery by finding hidden patterns and associations and presenting the results using visualization tools. The process of knowledge discovery is illustrated in the Fig. 3 and it consists of the following steps:

- (a) *Data cleaning*: removing invalid data

- (b) *Data integration*: combine data from multiple sources
- (c) *Data transformation*: data is transformed using summary or aggregation operations
- (d) *Data mining*: apply intelligent methods to extract patterns
- (e) *Evaluation and presentation*: use visualization techniques to present the knowledge to the user



DATA MINING

4.2. Data mining techniques

Association Analysis Adaptation for hSenid Software International:

Association analysis involves discovering patterns or rules showing attributes or conditions that frequently co-occur in a given dataset. At hSenid Software International, association analysis can be applied to understand patterns in customer behavior and preferences across their diverse range of services, including telecom, financial, and enterprise solutions.

For example, consider the following adapted rule:

If a customer frequently uses mobile banking services and has subscribed to a particular telecom plan, then they are likely to also use the company's enterprise solutions.

Support = 3%, Confidence = 70%

Here, support represents the percentage of customers who exhibit the specified behavior out of all the transactions analyzed. Confidence indicates the likelihood that a customer who uses mobile banking and subscribes to a particular telecom plan will also use enterprise solutions.

Apriori Method Adaptation for hSenid Software International:

The Apriori method is a popular algorithm for discovering association rules. It utilizes prior knowledge of frequent itemset properties to iteratively explore larger itemsets.

In the context of hSenid Software International, the Apriori method can be adapted to analyze customer usage patterns across different services offered by the company. The algorithm would identify frequent combinations of services utilized by customers, thereby enabling the company to tailor their offerings and marketing strategies more effectively.

Adapting the Apriori method involves:

1. Identifying frequent itemsets: Determine the combinations of services frequently used together by customers, such as mobile banking, telecom plans, and enterprise solutions.
2. Generating association rules: Based on the frequent itemsets, generate rules that highlight the associations between different services. These rules can provide valuable insights into customer behavior and preferences.
3. Evaluating rule interestingness: Assess the support and confidence levels of the generated rules to determine their significance and reliability.

By applying association analysis and the Apriori method, hSenid Software International can gain valuable insights into customer behavior and preferences, ultimately enhancing their service offerings and marketing strategies.

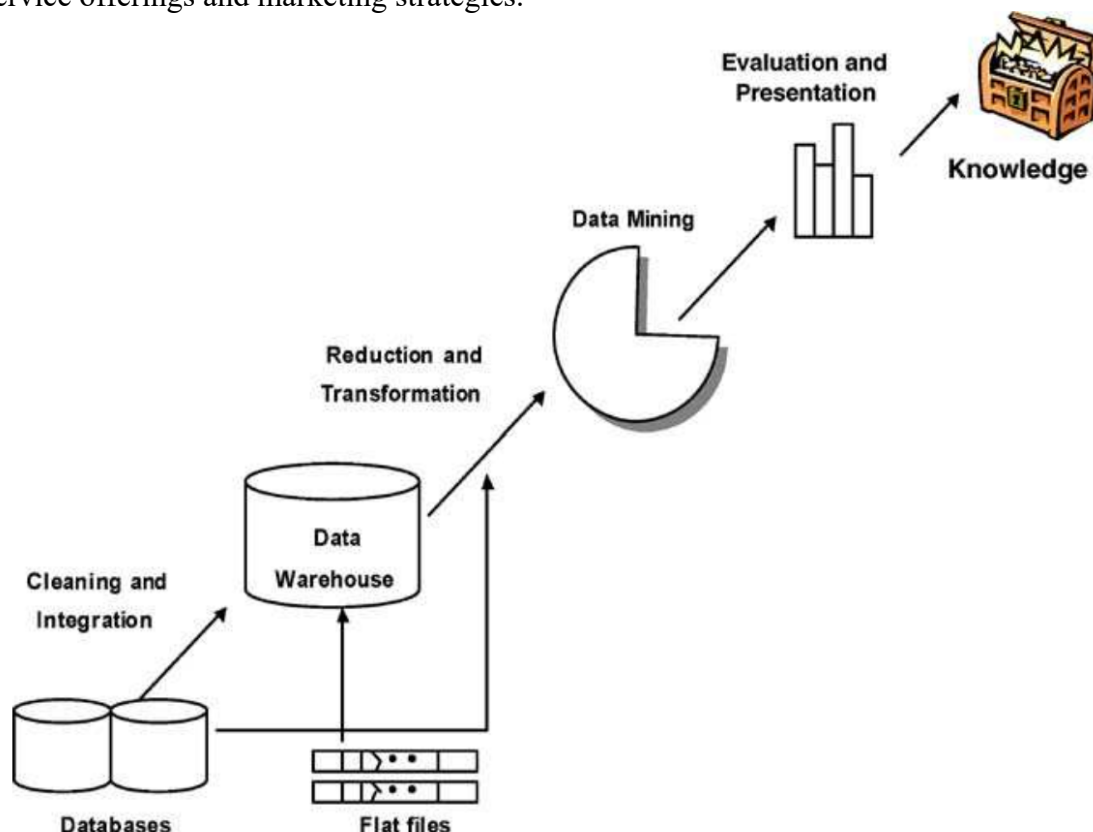


Fig. 3 Architecture of the knowledge discovery process

4.2.1. Classification and prediction

In the context of enhancing Intrusion Detection Systems (IDS) at hSenid Software International, classification and prediction play pivotal roles in extracting models that delineate various types of network intrusions or cyber threats, as well as forecasting future trends in network security. For instance, a classification model could be devised to categorize network activities as either normal or suspicious, while a prediction model could forecast the likelihood of a particular type of cyber attack based on historical data patterns.

Some fundamental techniques for data classification in this scenario include anomaly detection, pattern recognition, and behavioral analysis. These methodologies aim to discern abnormal network behaviors indicative of potential security breaches or malicious activities. Techniques such as anomaly detection could identify deviations from established patterns of network traffic, while behavioral analysis could uncover subtle indicators of unauthorized access or data exfiltration.

These techniques culminate in the derivation of models that encapsulate the distinct classes of network threats or anomalies. These models can subsequently be leveraged to predict the classification of network events or to anticipate the occurrence of specific cyber threats in real-time. The resultant models may be represented in various forms, including rule-based systems, decision trees, or statistical algorithms, enabling rapid and accurate identification of security incidents within the network environment at hSenid Software International.

4.2.2. Clustering

clustering involves grouping network traffic data and alerts so that items within a cluster exhibit high similarity but are distinctly dissimilar to items in other clusters. The fundamental principle of clustering is to maximize the similarity within clusters while minimizing the similarity between clusters. In the realm of cybersecurity and network monitoring, clustering can be instrumental in identifying patterns and anomalies in network traffic behavior. By clustering network traffic data and alerts, we can discern distinct groups of network activities, enabling more efficient detection of suspicious behavior and potential security threats.

For hSenid Software International, clustering techniques can be applied to categorize customer usage patterns, thereby facilitating targeted marketing strategies and personalized service offerings. Additionally, clustering can aid in organizing and classifying documents and information on various platforms, enhancing information discovery and retrieval for both internal and external stakeholders. Given the substantial volume of data collected by hSenid across its diverse business verticals, cluster analysis has emerged as a crucial area of focus within the realm of data mining research. As a form of unsupervised learning, clustering allows us to derive insights and patterns from data without predefined labels or categories, making it a valuable tool for learning through observation. While traditional statistical analysis packages have historically focused on distance-based cluster analysis, the application of clustering techniques in machine learning underscores its relevance and efficacy in addressing contemporary data analysis challenges. By leveraging clustering algorithms and methodologies, hSenid can gain deeper insights into customer behavior, network activity, and business operations, ultimately driving informed decision-making and strategic initiatives.

4.2.3. Outlier analysis

Outlier mining within the context of hSenid Software International's cybersecurity solutions can be described as follows: Given a database of network traffic data and intrusion alerts, the objective is to identify data objects that significantly deviate from the expected behavior of the network, potentially indicating malicious activities such as cyberattacks or intrusions. These outliers play a crucial role in applications such as fraud detection and network intrusion detection, enabling organizations to proactively identify and mitigate security threats.

Two primary approaches to outlier detection can be employed:

1. **Statistical-Based Outlier Detection:** In this approach, a probabilistic model is assumed for the network traffic and intrusion alert data. Outliers are identified based on their deviation from this model using statistical tests. The application of these tests necessitates knowledge of key parameters of the data set, such as mean, variance, and the expected number of outliers. By leveraging statistical analysis, hSenid's

cybersecurity solutions can effectively identify anomalies in network behavior that may indicate potential security breaches or suspicious activities.

2. **Distance-Based Outlier Detection:** In the context of network security, distance-based outliers can be interpreted as network entities (e.g., IP addresses, devices) that exhibit insufficient similarity to their neighboring entities. Neighbors are defined based on factors such as communication patterns, traffic volume, and protocol usage. Objects that lack "enough" neighbors or display anomalous communication patterns compared to their surroundings are flagged as outliers. By employing distance-based outlier detection techniques, hSenid's cybersecurity solutions can identify suspicious network entities that may be involved in malicious activities or unauthorized access attempts.

5. Introduction to intrusion detection systems

In today's rapidly evolving digital landscape, the escalating proliferation of cyber threats poses significant challenges to organizations worldwide. With the increasing accessibility to information processing and Internet connectivity, the frequency and sophistication of cyber attacks have surged exponentially. Consequently, safeguarding critical information systems, particularly those serving military and commercial purposes, against such threats has become imperative. Intrusion Detection Systems (IDS) play a pivotal role in fortifying the security infrastructure of modern networked information systems. By vigilantly monitoring networks or systems and scrutinizing audit streams for indicators of malicious activity, IDSs strive to preemptively detect and mitigate intrusions.

IDSs are typically categorized based on the type of information they analyze, leading to the distinction between host-based and network-based IDSs. Host-based IDSs scrutinize host-bound audit sources such as operating system audit trails and system logs, enabling precise identification of targeted host resources during an attack. Meanwhile, network-based IDSs analyze network packets captured within the network, leveraging traffic sensors strategically positioned across the network to detect and report suspicious events to a central location.

Furthermore, IDSs can be classified into different categories based on their detection methodologies. Misuse Detection systems seek intrusions by identifying direct matches to known attack patterns or signatures, offering low false alarm rates but limited efficacy against unknown attacks. Anomaly Detection systems, on the other hand, establish expected network behavior profiles in advance and flag deviations from these norms as potential threats, albeit grappling with high false alarm rates and labor-intensive data analysis. State Transition Based Intrusion Detection employs finite state machines to model IDS states and transitions, promptly flagging security threats when certain events trigger state changes.

Despite the promise of data mining-based IDSs in enhancing threat detection capabilities, they encounter significant challenges. Data collection and preparation processes are often labor-intensive, while performance discrepancies between simulated and real environments result in high false alarm rates and cumbersome data analysis tasks.

To address these challenges and enhance the efficacy of IDSs, there is a pressing need to develop methods and tools that empower security analysts to comprehend, analyze, and interpret the vast volumes of data collected by IDSs. This necessitates leveraging data modeling, data visualization, and data warehousing techniques to streamline intrusion detection processes, monitor network activities in real-time, conduct historical data analysis, and generate insightful trend reports.

In this paper, we present a comprehensive framework that integrates data modeling, data visualization, and data warehousing techniques to augment the performance and usability of IDSs. By delineating a robust data architecture tailored to our organization's needs and challenges, we aim to empower security analysts at hSenid Software International to bolster their intrusion detection capabilities, proactively safeguarding critical information assets against evolving cyber threats.

6. Data mining for intrusion detection

There has been a growing interest in applying data mining techniques to bolster intrusion detection systems. The challenge of intrusion detection can essentially be framed as a data mining problem involving classification. In essence, we are presented with a dataset comprising various classes (normal activities, different types of attacks) and aim to accurately segregate them using a model. This section provides an overview of ongoing research in this domain.

(1) **MADAM ID:** The MADAM ID project at Columbia University [9, 10] has shown how data mining techniques can be used to construct a IDS in a more systematic and automated manner.

(2) **ADAM:** The ADAM project [11, 12] is a network-based anomaly detection system. ADAM learns normal network behavior from attack-free training data and represents it as a set of association rules, the so called profile. At run time, the connection records of past delta seconds are continuously mined for new association rules that are not contained in the profile.

(3) **MINDS:** The MINDS project [13, 14] at University of Minnesota uses a suite of data mining techniques to automatically detect attacks against computer networks and systems. Their system uses an anomaly detection technique to assign a score to each connection to determine how anomalous the connection is compared to normal network traffic. Their experiments have shown that anomaly detection algorithms can be successful in detecting numerous novel intrusions that could not be identified using widely popular tools such as SNORT [25].

(4) **Clustering of Unlabeled ID:** Traditional anomaly detection systems require “clean” training data in order to learn the model of normal behavior. A major drawback of these systems is that clean training data is not easily available. To overcome these weakness, recent research has investigated the possibility of training anomaly detection systems over noisy data [15].

(5) **IDDM:** The IDDM system [16] uses anomaly detection techniques for intrusion detection.

(6) **eBayes:** The eBayes [17] system also uses anomaly detection for intrusion detection.

(7) **Alert Correlation:** [18, 19] use correlation techniques to construct “attack scenarios” from low level alerts. [20] also describes a language for modeling alert correlation. [29, 30] describe probabilistic alert correlation. [31] describes use of attack graphs to correlate intrusion event.

7. A data architecture for IDS

we introduce a data architecture tailored to enhance the efficacy of intrusion detection systems within the framework of hSenid Software International. Our focus lies on devising techniques to significantly boost performance, particularly in the realm of multi-dimensional data modeling, which serves as a cornerstone for representing alerts and identifying emerging cyber threats. Furthermore, we present methodologies for extracting relevant features from network traffic data and correlating alerts, all geared towards fortifying our cybersecurity infrastructure.

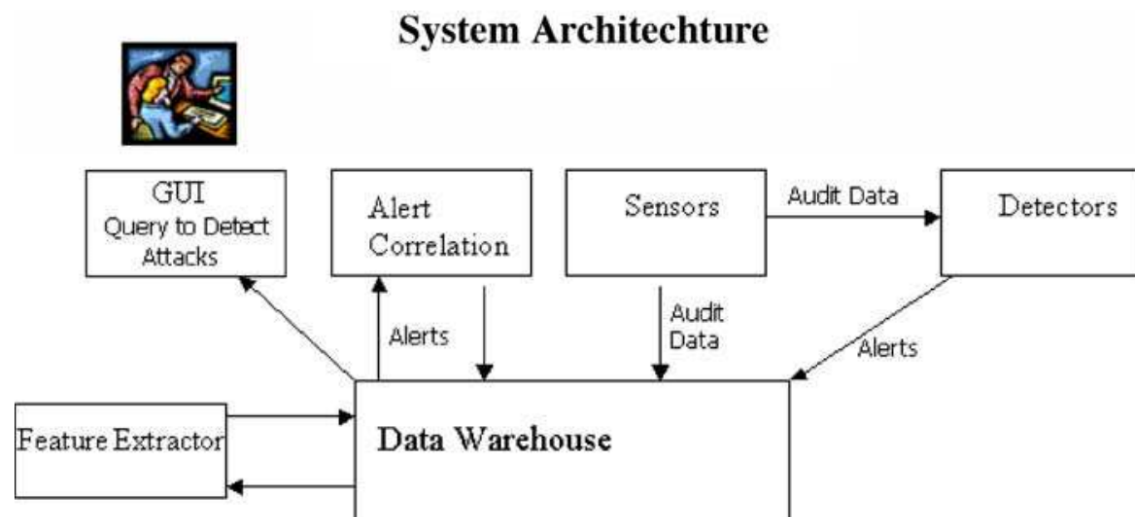


Fig. 4 Data architecture for intrusion detection system

7.1. A software architecture and data model for intrusion detection

Our system architecture, illustrated in Fig 4, addresses the complexities of intrusion detection within the unique context of hSenid Software International. In a diverse network environment, various audit streams contribute valuable data for intrusion detection, including network packets, system logs, and system calls. These data sources possess distinct properties, and detection models may vary from signature-based systems to data mining approaches. It is essential to devise an architecture capable of accommodating diverse data types and detection methodologies. Our architecture encompasses the following key components:

1. **Real-Time Components:** This encompasses sensors and detectors, responsible for monitoring network activity and identifying potential intrusions in real-time.
2. **Data Warehouse Component:** A centralized repository designed to efficiently store vast volumes of audit data, ensuring accessibility and scalability for future analysis.
3. **Feature Extraction Component:** This component extracts relevant features from the audit data stored in the data warehouse, computes aggregates, and enriches the information available for analysts. These extracted features are instrumental in detecting and analyzing potential cyber threats.
4. **Visualization Engine:** An intuitive interface that presents pertinent information to security analysts, facilitating informed decision-making and rapid response to detected threats.

Our proposed architecture offers several advantages tailored to the needs of hSenid Software International:

1. **Modularity:** By consolidating all data within a centralized repository, our architecture enables seamless querying by security analysts and intrusion detection applications, promoting operational efficiency.
2. **Support for Multiple Detectors:** We advocate for the separation of sensor and detector components, allowing for the integration of diverse detection engines, including signature-based and data mining approaches, leveraging the same repository of audit data.

3. **Correlation of Data from Multiple Sources:** Centralizing data from multiple sensors facilitates effortless correlation and analysis, empowering detection engines to access and analyze data across various sources through streamlined database queries.
4. **Reusability:** The centralized storage of extracted features promotes reusability across multiple applications, empowering different detection engines to leverage the same feature set for detecting and mitigating cyber threats effectively.

7.2. Data modeling for historical data analysis using STAR schema

In order to help the security officer or an analyst to decide whether an alert needs further investigation we plan to support the capability of querying and browsing a historical database. We propose to model the alert data as a multidimensional dataset and borrow the model used in On Line Analytical Processing (OLAP). A popular abstraction for multidimensional data that is widely used in OLAP is the data cube. A cube is simply a multidimensional structure that contains at each point an aggregate value, i.e. the result of applying an aggregate function to an underlying relation.

In our case, the underlying relation is the alerts that are generated from an IDS. The alerts can be viewed as a multidimensional data. This schema is known as the star schema. In it, the main table is called the fact table. The attributes are the dimensions of the data. Examples of dimensions are Time&Date, Sdinfo, Service, Attack. Time&Date contains information of date and time when the attack was staged. Sdinfo describes the Source/Destination IP addresses and destination port information. This dimension encompasses a hierarchy which shows how this information can be aggregated to produce different views. Both, the source and destination IP addresses are composed of 4 bytes Sip1Sip2Sip3Sip4 and Dip1Dip2Dip3Dip4. Dropping one or more of these fields produces a higher level view of the address. For example, Sip1Sip2 corresponds to a series of domain of IP addresses each characterized by the first 2 bytes of the address. The Service dimension table contains the service (or protocol e.g. ftp, http) name that was attacked and the class of service (e.g. TCP,

UDP). The hierarchy for these dimensions are also shown. For example, the service ftp and http belong to the TCP class. Similarly, the dimension table contains Attack contains both the name of the attack and its type (e.g. DOS, Probe). The dimension Time&Date presents different views of timing information. Finally, the attribute Duration contains the length of the attack. This can also be viewed as long, medium or short. Figure 5 shows the STAR schema. Figure 6 shows the dimension hierarchy for IP address. Figure 7 shows the dimension hierarchy for time and service/protocol.

Using this schema, a corresponding cube would be a five dimensional structure in which cell contains aggregates of the operations measures. For instance, a cell could correspond to short duration attacks over the ftp service in the period 1 pm to 2 pm on Oct. 20th 1998. Data cubes can be constructed by using SQL aggregation functions (COUNT, SUM, MIN, MAX). Cubes can be organized in a hierarchical manner. At the base of the hierarchy are the aggregates computed from the fact table. We call this base data. As data is consolidated -

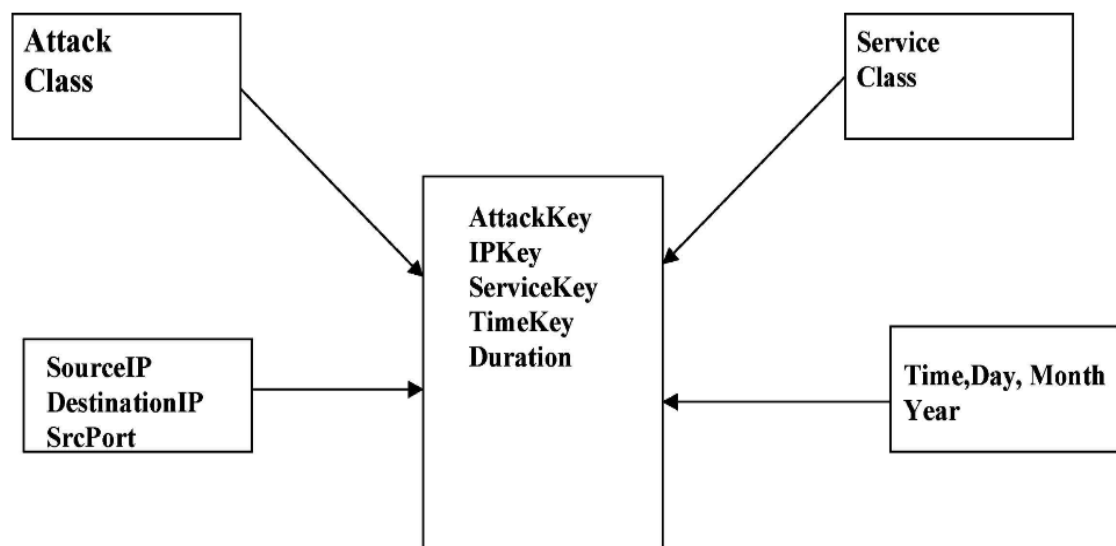


Fig. 5 A star schema for the IDS data warehouse

Fig. 6 Dimension hierarchy for IP address

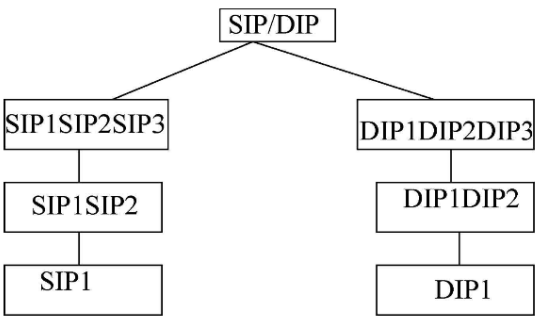
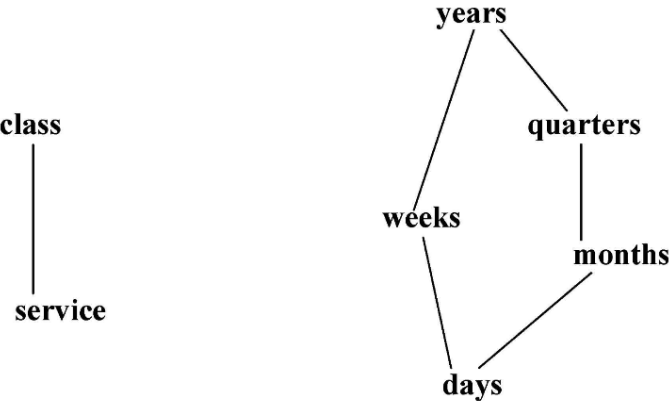


Fig. 7 Dimension hierarchy for time and services



-into higher levels it is called consolidated data. For example, in our data cube, the base data could be cells that contain aggregates of measures per user, operation, time period and date. Higher levels of hierarchy can be specified in terms of classes of users (users in division W), coarser time periods (e.g. morning) and date consolidation (e.g. Sept. 2000).

7.3. Support for high speed drill down queries and detection of attacks/virus/worms

When an alert is triggered by our Intrusion Detection System (IDS), analysts often need to "drill down" to examine the corresponding raw network traffic data to validate the alert. This involves clicking on an alert, prompting the system to retrieve the raw traffic data records associated with that alert. However, given the substantial volume of network traffic data we handle (typically reaching a terabyte for just one week of data), this process can be time-consuming.

To address this challenge, we've developed techniques to organize our raw network traffic data efficiently using STAR schemas. These schemas are optimized to facilitate swift query processing and seamless linkage between the raw network traffic data and the corresponding output alerts. Leveraging techniques such as bitmap indexing and join indexing further accelerates query processing, enhancing the overall efficiency of our system.

Our queries for security analysis of network traffic data encompass various scenarios, including scanning activity detection, worm detection, and identification of denial-of-service (DoS) attacks. For instance, to detect scanning activity, we analyze one-hour data segments and identify flows where the SYN flag is set while ACK/FIN flags are not set. Similarly, we've crafted queries tailored to detect specific threats, such as the Sasser worm, which scanned port 445. By narrowing our search to machines with destination port 445, we can pinpoint potentially infected internal systems.

Moreover, our queries enable us to detect network-based DoS attacks like SYN flooding by identifying source-destination IP pairs with an excessive number of SYN packets.

Additionally, we monitor for anomalies indicative of worm activity, such as the MyDoom worm, which prompted increased scanning for specific backdoor ports. Our SQL queries generate reports detailing fluctuations in flow counts and bytes transferred over specified time intervals, aiding in the timely detection and mitigation of emerging threats.

Incorporating these advanced querying techniques into our IDS infrastructure enhances our ability to swiftly analyze and respond to security incidents, thereby reinforcing our commitment to safeguarding our clients' networks and data assets.

7.4. Feature extraction from network traffic data

In the realm of cybersecurity, numerous Intrusion Detection Systems (IDS) applications, including those pioneered by hSenid Software International, necessitate preprocessing of network traffic data prior to conducting comprehensive analysis.

7.5. Help the security officer for forensic analysis

In the realm of cybersecurity, one crucial aspect is forensic analysis, which presently relies heavily on manual processes. Cybersecurity experts are tasked with sifting through vast volumes of data, often millions of records, individually to identify suspicious activities. This manual approach is not only highly inefficient but also incurs substantial costs. Given our capability to store extensive historical data, encompassing net-flow data, system calls, and firewall logs, within our data warehouse infrastructure, we can revolutionize the forensic analysis landscape. By leveraging our robust data warehouse, security officers gain streamlined access to all potentially suspicious records, aiding in the swift identification of intrusions or anomalous behaviors. Utilizing SQL statements, we can automate the labeling process, distinguishing between normal and anomalous activities, and subsequently updating and storing the flagged records back into our database. Our versatile database platform serves as the foundation for crafting tailored Digital Forensics tools, specifically attuned to the intricacies of Information Warfare, thereby delivering real-time performance and enhancing our cybersecurity capabilities.

8. System implementation

Fig 8 depicts the architecture diagram of a prototype data warehouse system tailored specifically to address the challenges of intrusion detection within the domain of hSenid Software International.

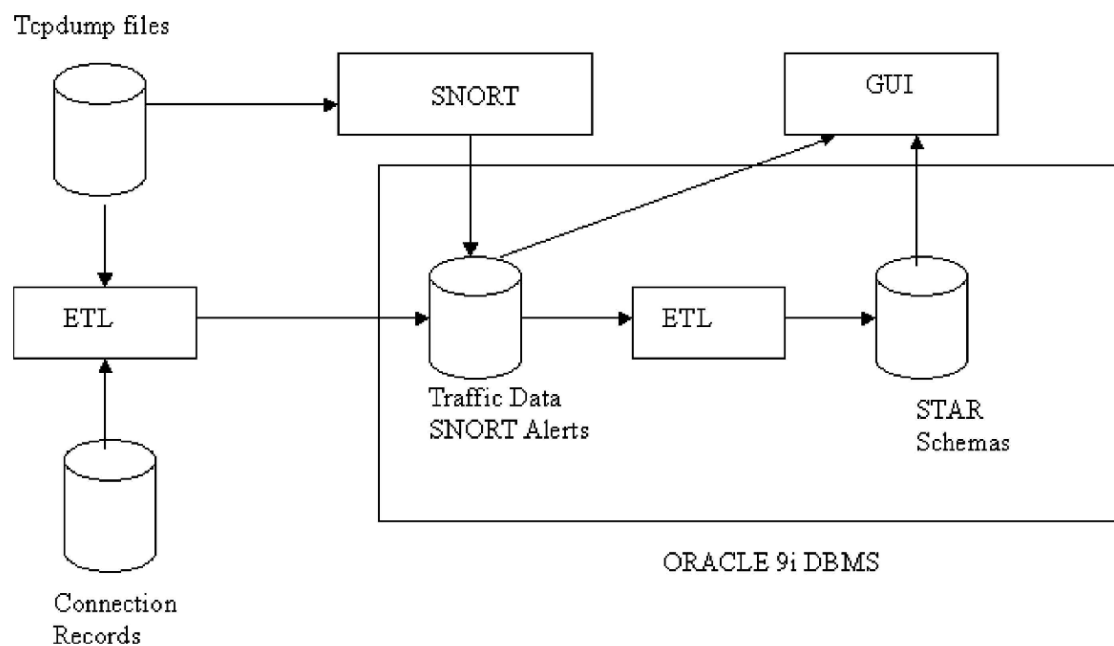


Fig. 8 Architecture of the data warehouse system for IDS

the alerts into the database. A data warehouse based on ORACLE 9i is the center piece of our architecture. We store the following kinds of data in the data warehouse.

Alert data

1. We created tables to store the alerts from SNORT (SNORT can be used to monitor the traffic that goes in and out of a network. It will monitor traffic in real time and issue alerts to users when it discovers potentially malicious packets or threats on Internet Protocol (IP) networks.). Some of these tables are event, sensor, signature and detail.

Network traffic data and extracted features

2. Within the framework of hSenid Software International's cybersecurity initiatives, we devised a database schema tailored to store essential network packet information derived from tcpdump data. This endeavor involved the creation of five distinct tables, each meticulously aligned with specific network packet headers. These tables encapsulate the Ethernet header, IP header, TCP header, UDP header, and ICMP header, respectively. By structuring the database schema in this manner, we ensure a systematic and comprehensive repository of network packet data, laying a solid foundation for advanced data mining and intrusion detection capabilities.

3. The data utilized for our project was sourced from hSenid's internal network infrastructure, spanning a period of approximately one month. This dataset comprises tcpdump data captured by network packet sniffers, encompassing the transmission details of every packet exchanged among devices within and outside hSenid's network environment.

4. We created loaders using Java/JDBC to load network traffic data into the ORACLE tables. We also designed programs to extract features and do data cleaning before loading it into the data warehouse.

5. We have also created a schema for storing *netflow* data and we have designing programs using database queries that do security analysis (e.g. detection of slow port scans) using netflow data. The slow port scans cannot be detected by SNORT because of time window limitations.

Aggregated data, STAR schemas and data summarization reports

6. At hSenid Software International, we devised STAR schemas to house the intricate data generated by our cybersecurity solutions, particularly the alerts generated by our Intrusion Detection Systems (IDS). Complementing this, we engineered an extract/transform/load

(ETL) program tailored to seamlessly extract data from our IDS database and seamlessly load it into the STAR schemas.

Table 1 Detection time for DOS attacks

Intrusion type	Intrusion name	Detection time (s)
Denial of service	Smurf	95
Denial of service	SYN flood	64

By leveraging the inherent data aggregation capabilities of STAR schemas, we effectively streamlined data storage requirements, especially for historical data, optimizing resource utilization while maintaining the integrity and accessibility of critical cybersecurity insights.

7. We also created STAR schemas to store the net-flow data. Since STAR schemas use data aggregation they reduced the amount of storage usage for past history. We wrote several queries to determine how the STAR schemas helped in improving the performance of detecting attacks such as Smurf and SYN flood. This schema also helped in creating Data Summarization Reports such as Top N Lists and Black List IP. One of the reports is to sort the results by the number of flow records and then only return the first N, from the list to present the results with the most traffic. This feature can be used by the user specifying either source or destination ip address. It can also be used on ports rather than IP address. If the user wishes, he can also look at combination of items such as top source/destination ip address pairs or source/destination port pairs.

8. Another useful report is to provide the number of bytes, packets and flows seen in the netflow data broken into user specified time intervals (five minutes, one hour, one day and so on). This allows a user to glance at a report and determine if there was any sudden spike in the activity.

Detection of DOS attacks

Since raw network traffic is stored in the database, we wrote queries that can detect certain kind of attacks such as Smurf and SYN flood.

a. Smurf attack is a scenario when there are a large number of replies to a particular machine from many different machines, but no “echo request” originated from that victim machine. b. SYN flood is a scenario where there are a number of SYN packets coming to a particular machine from an unreachable host.

Fig 9 shows the code for a sliding window algorithm to detect the SMURF attack. Table 1 also gives the time to detect these attacks for a ORACLE 9i database of 10 million records running on a SUN Sparc Station.

Security analysis

Intruders frequently conduct reconnaissance scans on networks before launching actual attacks. This reconnaissance activity typically manifests as SYN scans, wherein the attacker initiates connections with various hosts using SYN packets and waits for responses. To identify such probing behavior, we have developed queries tailored to analyze one-hour data segments, searching for network flows characterized by the presence of SYN flags without corresponding ACK and FIN flags. This approach allows us to pinpoint potential reconnaissance activities where attackers are probing network systems for vulnerabilities without completing full connection handshakes.

Query benchmarks

Several queries were written to determine how the STAR schemas helped for improving the performance. Table 2 give examples of the sample queries and a performance comparison of executing these queries on the STAR schema as compared to running them directly on the SNORT database. The time is reported in CPU seconds and the experiments were performed

on a SUN SparcStation running ORACLE 9i. The performance of queries on the STAR schema is much better as compared to that on the SNORT database.

```
SOURCE CODE
DECLARE
    timel TIMESTAMP;
    time2 TIMESTAMP;
    ddip NUMBER;
    dsip NUMBER;
    mea NUMBER;
BEGIN
    timel := CURRENT_TIMESTAMP; -- Get current timestamp
    time2 := timel + INTERVAL '2' MINUTE;

    LOOP
        EXIT WHEN timel >= CURRENT_TIMESTAMP + INTERVAL '2' DAY; -- Loop for 2 days

        SELECT COUNT(DISTINCT IP_Dest), COUNT(DISTINCT IP_Src) INTO ddip, dsip
        FROM ether_header e, ip_header i
        WHERE e.Connection_ID = i.Connection_ID
        AND time BETWEEN timel AND time2;

        IF dsip <> 0 THEN
            mea := dsip / ddip;
            INSERT INTO T1 VALUES(timel, time2, ddip, dsip, mea);
        END IF;

        timel := timel + INTERVAL '1' MINUTE;
        time2 := timel + INTERVAL '2' MINUTE;
    END LOOP;
END;
/
```

Fig. 9 A sliding window algorithm to detect smurf attacks

The intuitive reason the query time is much better for the STAR schema is that the aggregate value of COUNT is already pre-calculated.

- a. Count of all the alerts in the morning of the current date:
- b. Count of all alerts over the ftp service on the current date between 1 pm to 2 pm, grouped by destination IP address:

c. Count of all the alerts in the morning of the current date, grouped by subnet:

Alert visualization

In the context of hSenid Software International's cybersecurity initiatives, visualizing alerts generated by intrusion detection systems (IDS) is crucial for effective threat analysis and decision-making. While existing tools like ACID offer valuable features for alert presentation, they may lack the capability to provide a comprehensive summary of alerts through visual representations such as bar charts or graphs. To address this limitation, our team has developed a user-friendly graphical user interface (GUI) using Java and JDBC technologies.

Table 2 Performance comparison of query execution time

Execution time	SNORT database	Star schema	Speed up
Query a	1.56	0.22	7
Query b	40.28	0.45	90
Query c	58.01	1.08	55

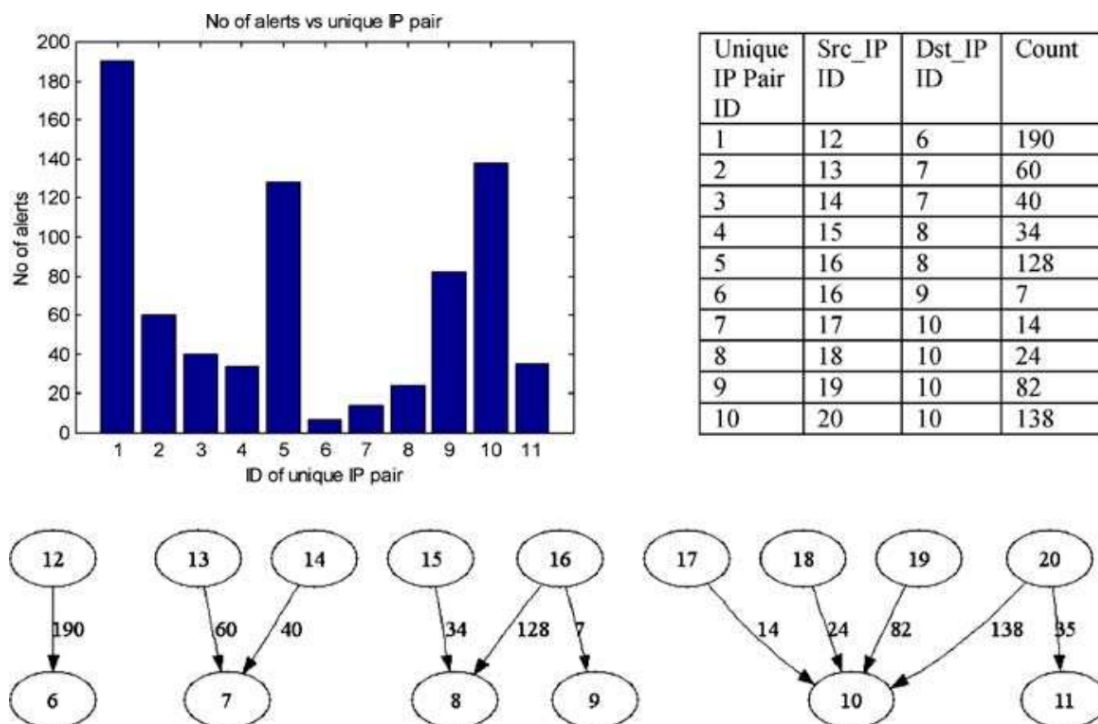


Fig. 10 Visualization of alerts

This GUI enables users to intuitively display and sort alerts based on various criteria, including time, source IP, and destination IP. Additionally, our solution allows for the visualization of alerts in graphical formats, such as graphs where nodes represent IP addresses using tools like GraphViz. By leveraging bar charts, users can gain insights into the frequency of alerts over specific time intervals, with IP addresses plotted along the x-axis and the number of alerts depicted on the y-axis.

Figure 10 show the results as bar charts and graphs.

Incorporating visual representations such as bar charts and graphs of the detected patterns will empower users to discern noteworthy trends and anomalies more effectively, enabling them to guide the system in deeper knowledge exploration. Users should have the flexibility to select the preferred format for presenting these patterns, facilitating a more intuitive understanding of the data. Additionally, we are exploring methods of implementing concept hierarchies to enable drill-down and roll-up functionalities, thereby allowing security personnel to examine identified patterns across various levels of abstraction.

9. Conclusions

This report outlines groundbreaking data modeling and data warehousing techniques tailored to significantly enhance the efficacy and usability of Intrusion Detection Systems (IDS), aligning with the objectives of hSenid Software International. Presently, IDS systems often lack robust support for historical data analysis and summarization, posing notable challenges in effectively detecting and mitigating cyber threats. By introducing innovative techniques to model network traffic and alerts utilizing a multi-dimensional data model and star schemas, this study addresses these limitations head-on.

The proposed data model serves as a robust foundation for comprehensive network security analysis and the detection of denial of service (DoS) attacks, enabling seamless integration and correlation of heterogeneous data sources such as firewall logs, system calls, and net-

flow data. Leveraging advanced data mining algorithms and optimized query response times, our approach promises up to two orders of magnitude improvement in performance compared to existing IDS systems, thus empowering hSenid to offer cutting-edge cybersecurity solutions to its clientele.

Moreover, the successful deployment of a prototype system at Army Research Labs underscores its efficacy in detecting intrusions and conducting historical data analysis to generate insightful trend reports, aligning perfectly with hSenid's commitment to delivering impactful solutions. Looking ahead, hSenid plans to further enhance the correlation capabilities of its system, leveraging data from multiple sources or sensors to gain deeper insights into network activity. Additionally, exploration of main memory database techniques for real-time alert correlation and visualization of attack scenarios will further solidify hSenid's position as a pioneer in cybersecurity innovation.