КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (23), 2024
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2024.23.274283
UDC 004.67

Viktor Sahaidak
PhD student
State University of Information and Communication Technologies, Kyiv, Ukraine
ORCID 0009-0000-9724-958X
qsagvict@gmail.com

OVERVIEW OF FRAUD DETECTION SYSTEMS AND PERFORMANCE KPI DEVELOPMENT

Abstract. In this article overview was provided on several fraud detection systems, analysis result of common scheme and development of KPIs to detect performance degradation or improvement from business logic point of view. Four different systems were reviewed. Following FMS were developed by Gigamon and Argyle Data cooperation, AWS, Subex, Cvidya Amdocs. Solution developed by Gigamon and Argyle Data consists of Gigamon fabric for information collection/filtering/enrichment and Argyle Data Fraud detection system, which is based on Hadoop technology to store collected data and analysis results by application. AWS Fraud Detection collects NRTRDE flow and process it by using ML technics provided by AWS. Subex fraud management system provides flexible ETL for data collection from different sources with adjustable detection rules and ML for suspicious behavior learning. FraudView by Cvidya Amdocs collects information from varying points like OSS/BSS, CRM customer details, Prepaid platforms, HLR, Switch CDRs, Probe (SS7, VoIP, IP) and process it by different detection engines. Simplified processing FMS processing scheme and KPIs based on different timestamps were made. Following conclusions were made: In reviewed FMS was noticed that instead of using traditional NRTRDE and TAP3 file formats, data can be collected directly from network by using network tap or port mirroring with next data enrichment, cleaning, formatting for fraud detection system to consume. Following real time method can be realized by using probes to perform data preparation or some complex solution described by Gigamon; Detection is performed by rules, provided by vendor or by ML modules, which learns behavior of subscriber in order to create detection rules. Most of systems allow to modify threshold of following rules in order to meet system user demands to check data within specific time (for example fraudster night calls to subscriber) or detect specific number of suspicious sessions, etc; In order speedup fraud detection hotlists, whitelists can be used for enrichment to filter out fraudsters, emergency or business numbers. Geographical location can be used to identify fraudster's location within network and make correlation with other possible fraud sessions; During analysis of each FMS architecture, 3 processing stages were highlighted, which allowed to create simple KPIs for business logic and data arrival check; Developed methodology allows to check data arrival and fraud recognition with used data type to define which information provides better detection or view on rules for detection in order to show, which of them should be adjusted.

Ключові слова: FMS; ML; ETL; Hadoop; AWS; RDBMS; SS7; VoIP; IP.

INTRODUCTION

With telecommunications development a lot of services have evolved. Same can be said for telecommunications fraud detection:

- TAP3 CDR format was developed at 1991, which is delivered within 30 days after call was made;
- NRTRDE CDR format which is mandatory since 2008. Following files delivered within 4 hours after call was made.

ISSN 2663 - 4023

Currently on market we can see a lot of FMS (Fraud management system) and fraud detection. Some of them based on ML, some instead of mentioned above file formats use data from other sources. While part of solution requires dedicated hardware for each module, others prefer virtualized infrastructure.

Based on following differences the most common questions will be:

- Which solution company should choose?
- How to measure KPI of such system?
- How detection depends on data type and technology used by solution?

The purpose of following study is to define technical KPIs for fraud detection system in order to detect performance degradation or improvement of business logic in fraud detection system workflow.

To achieve defined purpose next tasks should be done:

- Analyze module, subsystems used in FMS in order to define stages of each processing.
- Propose simplified scheme with stages of data processing in Fraud detection system.
- Define KPIs, based on timestamps of each processing stage in FMS.

ANALYSIS OF EXISTING FMS CAPABILITIES AND MODULES

To define common parts and features in fraud detection system, let's take a look at some solutions:

The Gigamon and Argyle Data Joint Solution [1] – [2] consists of Gigamon fabric solution and Argyle Data Real-Time Fraud Analytics Hadoop Application. Gigamon fabric solution (Fig. 1) implements real-time data collection by network tap installation between servers with VNF (virtualized network functions), leaf switches, "spine" switches, routers or port mirroring with data masking and filtering, duplication cleaning, source labeling, time stamping, packet slicing, etc.

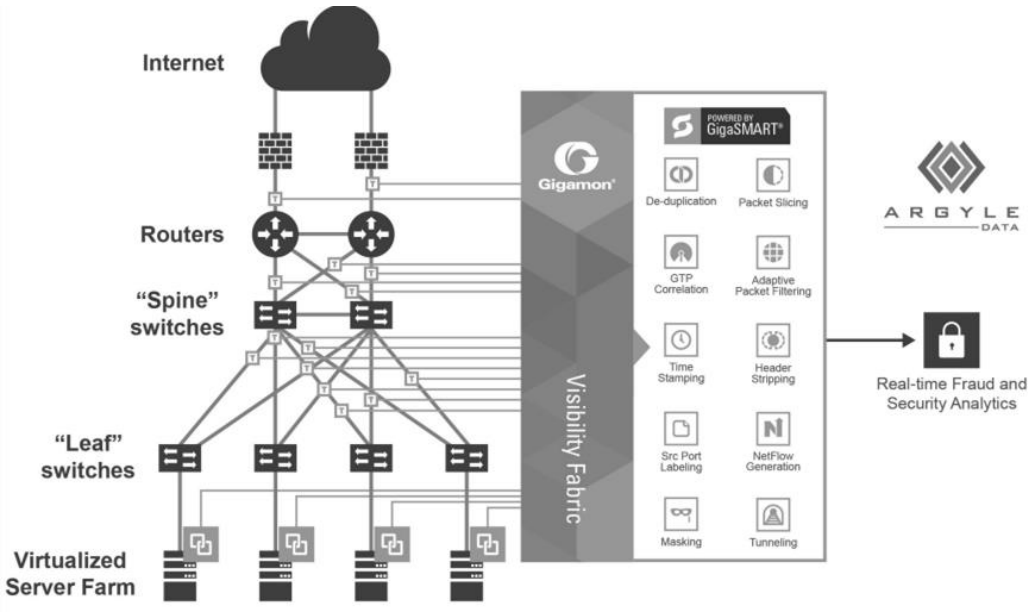

Fig. 1. Gigamon fabric scheme

ISSN 2663 - 4023

Argyle Data Real-Time Fraud Analytics Hadoop Application [3] (Fig. 2) collects information to data access, stores and manages data in Hadoop with help of RDBMS (Relational database management system). In the next step applications like statistical analysis, BI/Reporting, Ad Hoc Analysis provide processing of data from Hadoop and write down results back to it, so user can have access to processing results.

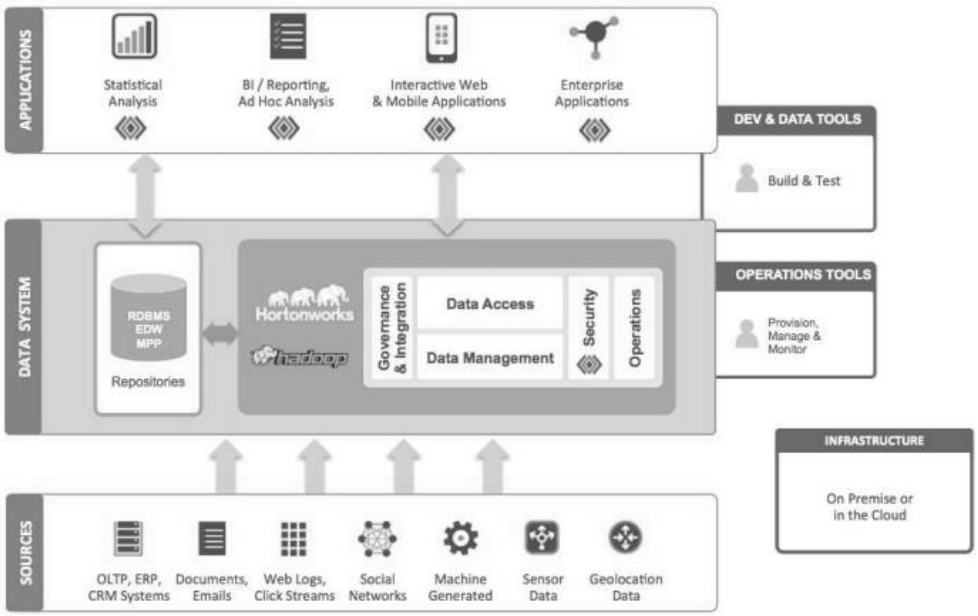

Fig. 2. Argyle Data FMS structure

AWS Fraud Detection [4] provide NRTRDE data processing by using ML, which works in next way (Fig. 3):

1. At training time, telecom data is batch transferred or streamed using Amazon Kinesis, into an Amazon Simple Storage Service (Amazon S3) bucket with help of AWS Glue Data Catalog to catalog the data.
2. Data is feature engineered using Amazon SageMaker Data Wrangler and transformed into features. Data can be sourced direct from Amazon S3 or using Amazon Athena queries. Features are stored in Amazon SageMaker Feature Store.
3. A custom (classification) model for fraud detection is trained using Amazon SageMaker. The model is tested and validated to ensure the model is regularized and performant for real-world use.
4. Trained models are stored in Amazon SageMaker Model Registry to track and manage the model over time.
5. Amazon SageMaker Model Monitor is used to monitor model quality over time, including data and model quality, and bias drift.
6. Once all tests for accuracy and performance have passed, the model is deployed using Amazon SageMaker endpoints to support near real-time inference. Amazon SageMaker endpoints help manage scalability, efficient operation, and reliability.
7. At inference time, data from the telco and partners is streamed in using Amazon Kinesis to an AWS Lambda function. Amazon API Gateway provides features to control access to the model endpoint. The fraud detection result produced by the model can then be consumed by the telco.

ISSN 2663 - 4023

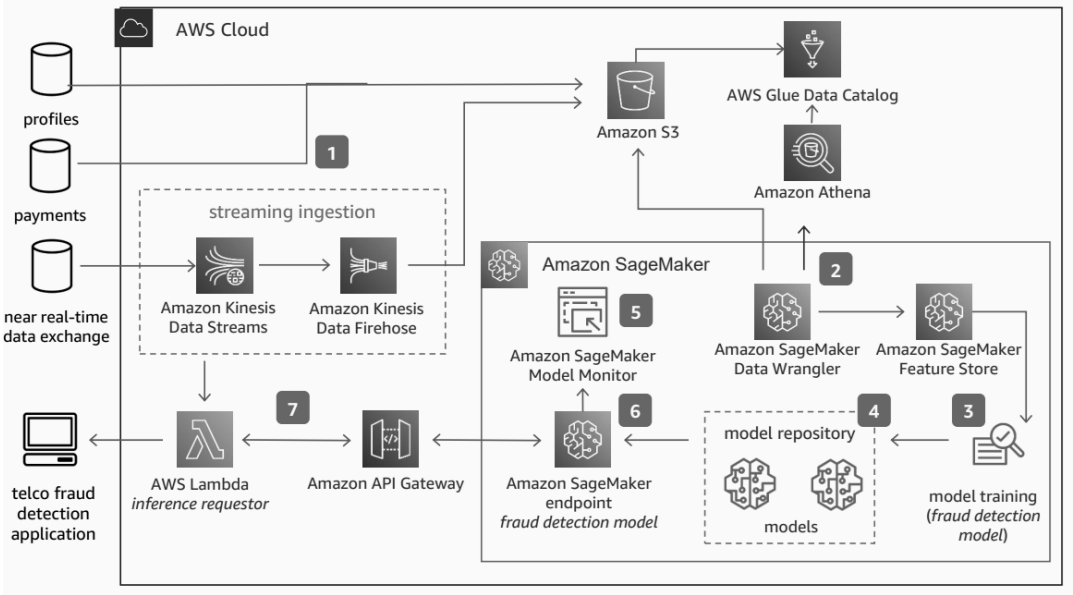

Fig. 3. AWS Fraud Detection flow

Subex fraud management system [5] (Fig. 4) collects data by using flexible and scalable ETL tools in real time from various data sources, transforms and enriches it. In next step data is loaded to DB and gets processed by rules that can be configured on the system include threshold rules, geographic rules, pattern rules, machine learning rules, hotlist rules, SPAM detection rules, intrusion detection rules, and smart pattern rules. And at last alerted data is visualized on dashboards, creates cases.

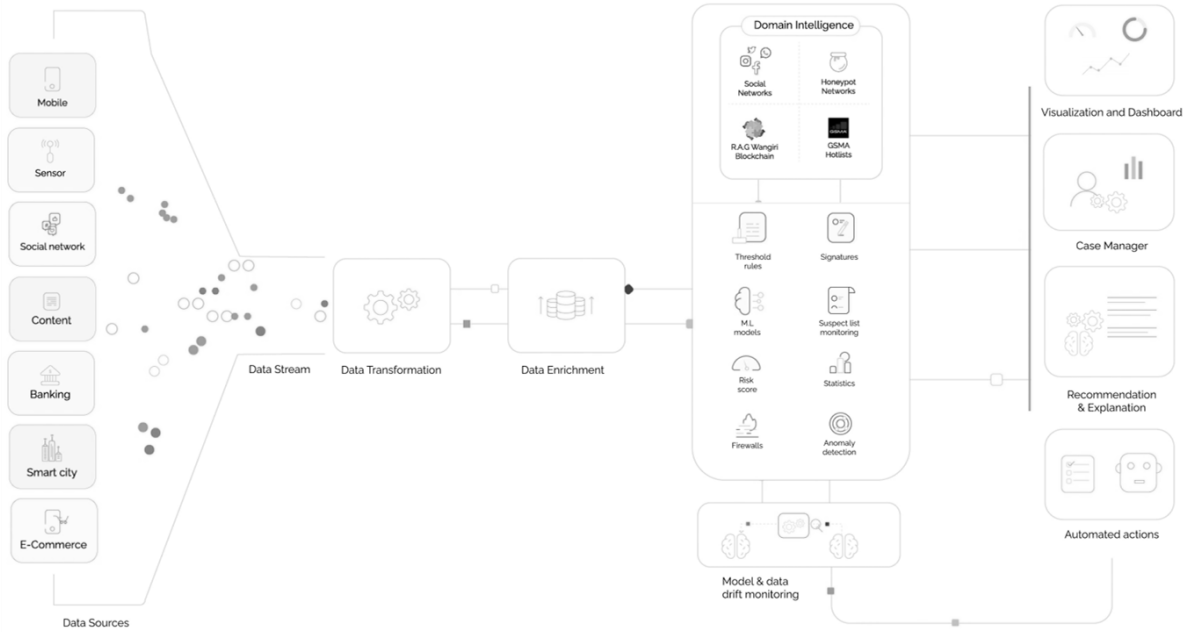

Fig. 4. Subex Fraud Detection

FraudView by Cvidya Amdocs [6] – [7] FMS, that collects data in real time by Probes (SS7, IP, VoIP), OSS/BSS, CRM customer details, Prepaid platforms, HLR, Switch CDRs and other data for which processing and collection can be developed (Fig. 5).

ISSN 2663 - 4023

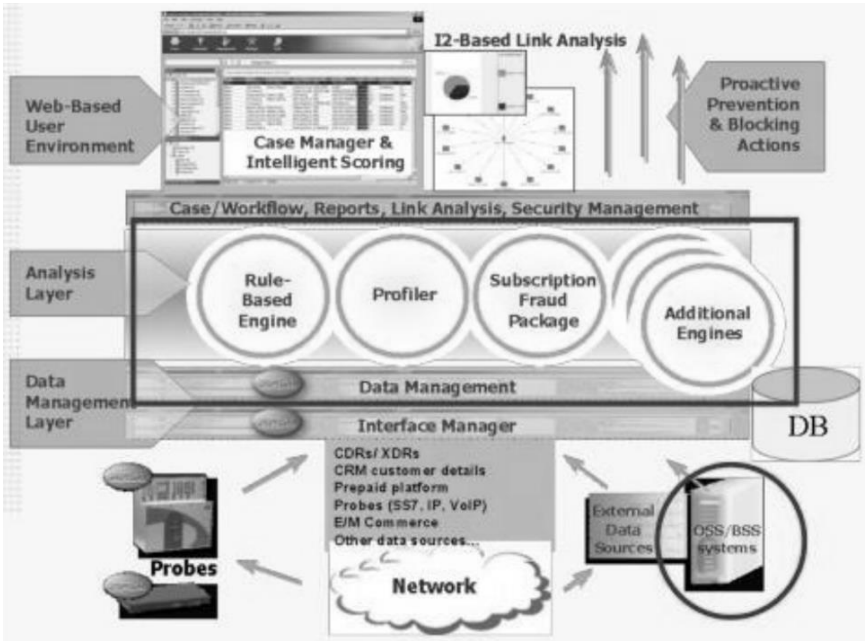

Fig. 5. FraudView Structure

System works in next approach (Fig. 6):

1. FMS collects data from sources on interface manager and on data management load it in DB.
2. Information from DB provided to engines, which process data in different ways.
3. After one of engines finished processing, alerts are created and added to cases.
4. If another engine has detected fraud for suspected entity, which was detected earlier, data is added to case.
5. In case analysis stage FMS provides CDR (Event) Analysis, client data (Billing data), Behavioral analysis, Link analysis to find fraudsters accomplice, Fraud Pattern analysis, Fraud scheme and Historical analysis (past calls and payment analysis).

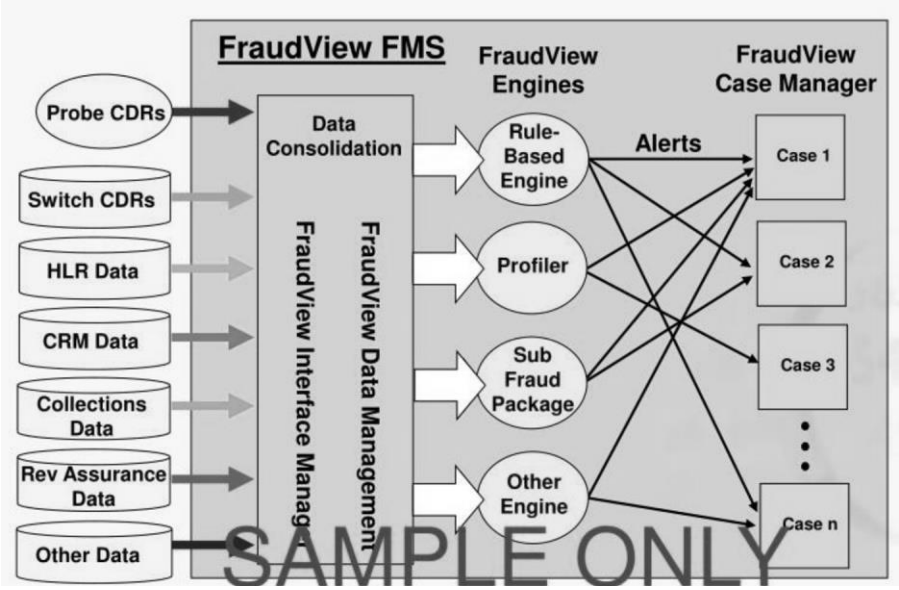

Fig. 6. FraudView Data processing flow

ISSN 2663 - 4023

RESEARCH RESULT

All researched fraud detection systems have 3 stages – data loading/collection, data analysis and alert/case creation (Fig. 7).

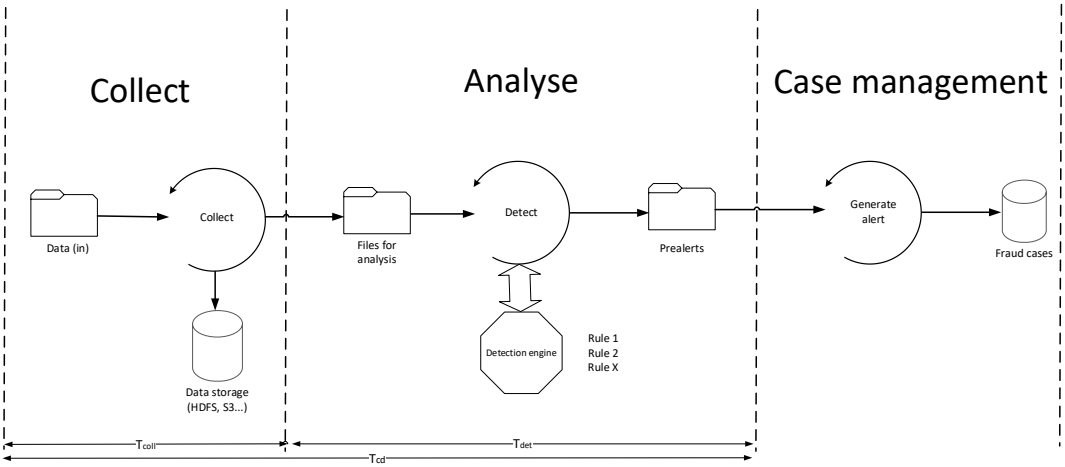

Fig. 7. Common simplified FMS scheme

To estimate each stage performance for collection stage T_{coll} can be used, for analysis stage T_{det} can be used and to estimate alert/case creation T_{cd} can be used. It is recommended to collect statistics for 1 day to check detection at specific time like low traffic period (night time), during business hours (network high usage time) and after business hours (evening time). Following approach can be used for each data source separately or combined together to check how it will improve fraud detection.

T_{coll} is time difference between call/session end time ($T_{session\_end\_time}$) and time, when CDR was loaded to Data storage ($T_{insertion\_time}$).

$$T_{coll} = T_{session\_end\_time} - T_{insertion\_time} \quad (1)$$

T_{det} is time difference between time of alert generation for suspected entity ($T_{alert\_creation\_time}$) and insertion time of CDR ($T_{suspect\_insertion\_time}$) for following suspected entity

$$T_{det} = T_{alert\_creation\_time} - T_{suspect\_insertion\_time} \quad (2)$$

T_{cd} is time difference between call/session end time ($T_{session\_end\_time}$) and time, when alert was generated for suspected entity ($T_{alert\_creation\_time}$).

$$T_{cd} = T_{alert\_creation\_time} - T_{session\_end\_time} \quad (3)$$

In order to scale and visualize the time required to for each stage, it is possible to divide the obtained values into intervals of 5 minutes for more accurate statistics. Intervals for 10 or more minutes can be used depending on calculation result. To see what interval is most crucial, weighted average time can be used.

$$\frac{\sum_{i=1}^{n} T_{coll_i} a_i}{\sum_{i=1}^{n} a_i} \quad (4)$$

$$\frac{\sum_{i=1}^{n} T_{det_i} a_i}{\sum_{i=1}^{n} a_i} \quad (5)$$

$$\frac{\sum_{i=1}^{n} T_{cd_i} a_i}{\sum_{i=1}^{n} a_i} \quad (6)$$

Where a is a number of values within each time interval, n — is number of time intervals, T_{coll_i} — time difference between call/session end time and time, when CDR was loaded to Data storage for each time interval, T_{det_i} — time difference between time of alert generation for suspected entity and insertion time of CDR for suspected entity each time interval, T_{cd_i} — time difference between call/session end time and alert generation time for suspected entity in each time interval.

CONCLUSIONS

1. In reviewed FMS we can see that instead of using traditional NRTRDE and TAP3 file formats, data can be collected directly from network by using network tap or port mirroring with next data enrichment, cleaning, formatting for fraud detection system to consume. Following real time method can be realized by using probes to perform data preparation or some complex solution described by Gigamon.
2. Detection is performed by rules, provided by vendor or by ML modules, which learns behavior of subscriber in order to create rules. Most of systems allow to modify threshold of following rules in order to meet system user demands to collect data within specific time (for example fraudster night calls to subscriber) or detect specific number of suspicious sessions, etc.
3. In order to enrich and speedup fraud detection hotlists, whitelists can be used to filter out fraudsters, emergency or business numbers. Geographical location can be used help to identify fraudster's location within network and make correlation with other possible fraud sessions.
4. During analysis of each FMS architecture, 3 processing stages were highlighted, which allowed to create simple KPIs for business logic and data arrival check.
5. Developed methodology allows to check data arrival and fraud recognition with used data type to define which information provides better detection or view on rules for detection in order to show, which of them should be adjusted.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Real-Time Fraud Detection and Analytics using Hadoop and Machine Learning. (2015). Network-Level Intelligence for Observability Tools | Gigamon. URL: http://surl.li/tkbjz
2. Argyle Data and Gigamon to deliver real-time fraud detection and analytics for communications service providers | VanillaPlus - The global voice of Telecoms IT. (2015, March 4). VanillaPlus - The global voice of Telecoms IT. URL: http://surl.li/tkbkq
3. Real-Time Fraud Analytics Hadoop Application. (2014, November). Cloudera | The hybrid data company. URL: http://surl.li/tkbmp
4. Intelligent Fraud Monitoring | AWS Solutions for Telecommunications | AWS Solutions Library. (n.d.). Amazon Web Services, Inc. URL: http://surl.li/tkbmu
5. Telecom Fraud Management | Telecom Fraud Detection | Telco Risk. (n.d.). Subex. URL: http://surl.li/tkbmz
6. PPT - Fraud Management and Operations Training PowerPoint Presentation - ID:1050298. (n.d.). SlideServe. URL: http://surl.li/tkbne
7. CVidya Launches FraudView® Version 9. (2010, October 13). Newswire | Press Release Distribution | Media Outreach Platform. URL: http://surl.li/tkbnm
8. Sahaidak V. A., Lysenko M. M., Senkov O. V. (2022). Telecom fraud and it's impact on mobile carrier business. *Connectivity, 160(6),* 17–20. https://doi.org/10.31673/2412-9070.2022.061720

9. Aravamuthan, S. (2021). Revenue Assurance and Fraud Detection for Telecom Operators – Combating Bypass Fraud. *International Journal for Research in Applied Science and Engineering Technology, 9(VII),* 2843–2851. https://doi.org/10.22214/ijraset.2021.37011
10. Pollard, C. (2005). Telecom fraud: The cost of doing nothing just went up. *Computers & Security, 24(6),* 437–439. https://doi.org/10.1016/j.cose.2005.07.006

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (23), 2024
ISSN 2663 - 4023

Сагайдак Віктор Анатолійович
Аспірант
Державний університет інформаційно-комунікаційних технологій, Київ, Україна
ORCID 0009-0000-9724-958X
qsagvict@gmail.com

ОГЛЯД СИСТЕМ РОЗПІЗНАННЯ ШАХРАЙСТВА ТА РОЗРОБКА КОЕФІЦІЄНТІВ ДЛЯ ВИЗНАЧЕННЯ ЇХ ЕФЕКТИВНОСТІ

Анотація. В цій статі було наведено опис декількох систем виявлення шахрайства, спрощена загальна схема, що було створена на базі огляду схем, та розроблено ключові показники ефективності для відстеження покращення або погіршення продуктивності з точки зору бізнес логіки. Були розглянуті 4 системи. Ці FMS були розроблені кооперацією компаній Gigamon та Argyle Data, AWS, Subex, Cvidya Amdocs. Комплекс розроблений Gigamon та Argyle Data складається з Gigamon fabric для збору, фільтрування, доповнення інформації та системи розпізнання шахрайства Argyle Data, що побудована на технології Hadoop для зберігання зібраних даних та результати аналізу додатку. AWS Fraud Detection збирає потік NRTRDE та оброблює за допомогою машинного навчання AWS. Система розпізнання шахрайством Subex надає гнучкий ETL для збору даних з різних джерел, правила виявлення з можливістю редагування та машинне навчання для вивчення підозрілої поведінки. FraudView від Cvidya Amdocs збирає інформацію з різних точок як OSS/BSS, CRM, білінгових платформ, HLR, CDR з комутаторів, Probe (SS7, VoIP, IP) та обробляє її різними механізмами виявлення. Були створені спрощена схема обробки FMS та ключові показники ефективності на основі різних часових позначок. Були зроблені наступні висновки: у розглянутих системах розпізнання шахрайства було виявлено, що замість використання традиційних форматів файлів NRTRDE і TAP3, дані можна збирати безпосередньо з мережі за допомогою мережевого відгалужувача або віддзеркалення порту з наступним збагаченням, очищенням, форматуванням даних для використання системою виявлення шахрайства. Описаний метод може бути реалізований за допомогою зондів для підготовки даних або деякого комплексу, описаного Gigamon; Виявлення виконується за правилами, наданими постачальником, або модулями машинного навчання, які вивчають поведінку абонента для створення правил для розпізнання. Більшість систем дозволяють змінювати налаштування правил, щоб задовольнити вимоги користувача системи щодо перевірки даних протягом певного часу (наприклад, нічні дзвінки які здійснюють шахраї на абонента) або виявлення певної кількість підозрілих сеансів та тому подібне; Щоб прискорити виявлення шахрайства списки номерів можна використовувати для фільтрації номерів шахраїв, екстрених служб чи бізнесу. Географічне розташування може бути використане для визначення місцезнаходження шахрая в мережі та встановити взаємозв'язок з іншими можливими сеансами шахрайства; Під час аналізу кожної архітектури FMS було виділено 3 етапи обробки, що дозволило створити прості ключові показники ефективності для бізнес-логіки та перевірки надходження даних; Розроблена методологія дозволяє перевірити надходження трафіку та розпізнавання шахрайства з використовуваним типом даних, щоб визначити, яка інформація забезпечує краще виявлення, або переглянути правила виявлення, щоб показати, які з них слід відкоригувати.

Ключові слова: FMS; машинне навчання; ETL; Hadoop; AWS; RDBMS; SS7; VoIP; IP.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Real-Time Fraud Detection and Analytics using Hadoop and Machine Learning. (2015). Network-Level Intelligence for Observability Tools | Gigamon. URL: http://surl.li/tkbjz
2. Argyle Data and Gigamon to deliver real-time fraud detection and analytics for communications service providers | VanillaPlus - The global voice of Telecoms IT. (2015, March 4). VanillaPlus - The global voice of Telecoms IT. URL: http://surl.li/tkbkq

3. Real-Time Fraud Analytics Hadoop Application. (2014, November). Cloudera | The hybrid data company. URL: http://surl.li/tkbmp
4. Intelligent Fraud Monitoring | AWS Solutions for Telecommunications | AWS Solutions Library. (n.d.). Amazon Web Services, Inc. URL: http://surl.li/tkbmu
5. Telecom Fraud Management | Telecom Fraud Detection | Telco Risk. (n.d.). Subex. URL: http://surl.li/tkbmz
6. PPT - Fraud Management and Operations Training PowerPoint Presentation - ID:1050298. (n.d.). SlideServe. URL: http://surl.li/tkbne
7. CVidya Launches FraudView® Version 9. (2010, October 13). Newswire | Press Release Distribution | Media Outreach Platform. URL: http://surl.li/tkbnm
8. Sahaidak, V. A., Lysenko, M. M., Senkov, O. V. (2022). Telecom fraud and it's impact on mobile carrier business. *Connectivity, 160(6),* 17–20. https://doi.org/10.31673/2412-9070.2022.061720
9. Aravamuthan, S. (2021). Revenue Assurance and Fraud Detection for Telecom Operators – Combating Bypass Fraud. *International Journal for Research in Applied Science and Engineering Technology, 9(VII),* 2843–2851. https://doi.org/10.22214/ijraset.2021.37011
10. Pollard, C. (2005). Telecom fraud: The cost of doing nothing just went up. *Computers & Security, 24(6),* 437–439. https://doi.org/10.1016/j.cose.2005.07.006