УНИВЕРСИТЕТ ЗА НАЦИОНАЛНО И СВЕТОВНО СТОПАНСТВО

Факултет „Приложна информатика и статистика“

Катедра „Информационни технологии и комуникации“

„Принципи и методи за проектиране на оперативен център за управление на информационна сигурност за системи с големи данни“

Автореферат на дисертационен труд за придобиване на образователна и научна степен "доктор"

Докторант:
Ивона Велкова

Научен ръководител:
проф. д-р Любен Боянов

София, 2023

Съдържание

Обща характеристика на дисертационния труд

1. Актуалност на проблема

Нарастващата дигитализация на процесите, обектите и дейностите в нашето ежедневие, е съвременна тенденция, породена от технологичния напредък през последните няколко десетилетия. Един от резултатите на тази дигитализация е значителното увеличаване на обема на данните, генерирани от различни източници [1]. С навлизането на дигиталната трансформация в почти всяка област на човешката дейност, данните се превърнаха в един от основните активи за получаване на конкурентно предимство за всеки бизнес. Това е повсеместен процес и засяга на практика всички области на човешка дейност - от здравеопазване, финанси, логистика, транспорт до екология и спорт. От всички тези дейности непрекъснато се събират огромни количества данни и информация. Данните могат да бъдат от различно естество - за пазари, клиенти, конкуренти, различни процеси, количества стоки, състояние на средата, видеонаблюдения, данни от социални медии и т.н. Благодарение на тези данни, хората получават информация, която може да бъде анализирана и да носи полза на бизнеса. Процесът на извличането на знания от събирането на данни и техния анализ е в състояние да оптимизира и носи по-голяма ефективност на всеки бизнес или части от него. С получените знания, компаниите могат да идентифицират какво е важно за тях и какви допълнителни или алтернативни действия могат или трябва да предприемат, за да актуализират или оптимизират целите си за краен успех [2].

Според Statista, общото количество генерирани и консумирани данни в световен мащаб е достигнало 97 зетабайта (ZB) през 2022 г., което е значително увеличение спрямо предходни години [1]. За сравнение, през 2020 г. броят на генерираните данни е 64 ZB [1]. С нарастването на обема на генерираните данни, нарастват и рисковете, свързани с тяхната защита - колкото са по-големи данните, толкова по-сложни са механизмите и подходите, с помощта на които тези данни се събират и анализират [3].

Информационната сигурност е от съществено значение за защитата на информацията в дадена организация. Това се отнася за данни за клиенти, финансови

записи, интелектуална собственост и др., които могат да са подложени на киберзаплахи като зловреден софтуер, фишинг, атаки чрез социално инженерство и др. Информационната сигурност представлява процес, който има за цел да предотврати неоторизиран достъп, да противодейства на различните видове заплахи, да осигури поверителност и да намали риска от унищожаване или модифициране на съхраняваната информация [4].

За да може да бъде ефективно управлявана, информационната сигурност е необходимо да идентифицира, оценява и управлява рисковете на информационните активи на организацията. Това изисква задълбочен и интегриран подход към сигурността, който включва прилагането на политики, стандарти, процедури и технически контрол за защита срещу потенциални заплахи. Поради изброените причини, наличието на оперативен център за управление на информационната сигурност, който може да наблюдава и да реагира на заплахи за сигурността, е от решаващо значение за предпазването на дигиталната информация и осигуряването на подходящо ниво на сигурност за бизнеса [5].

Защитата на данните и информацията при съхранението, обработката и анализа на големи данни в редица случаи се осъществява от облачни и софтуерни решения, които се характеризират с високо ниво на сигурност. Все по-често тези системи включват Изкуствен интелект (ИИ), който повишава ефективността на обработката на големи количества данни, включително и на неструктурирани данни под формата на текст, видео, аудио и уеб съдържание [6].

В настоящия труд се предлага подход, с който се осигурява сигурност на различните нива на данните, на процесите и на комуникациите в системи за големи данни. Данните са извлечени от хетерогенни източници, част от които са видео потоци от камери, данни от уеб сайтове, социални медии и др.. Последните се използват активно от бизнеса, тъй като предлагат набор от възможности - от повишаване на осведомеността за дадена търговска марка, модел или услуга, до информиране на потенциални клиенти и стимулиране на продажбите. Като такива, те са един от важните съвременни дигитални инструмент за още по-голям успех на бизнеса. Според доклад на Datareporta към юли 2022 г., 59% от хората, които използват Интернет в световен мащаб

използват платформи за социални медии [7]. Това показва огромната важност на тези медии, като източник на данни за бизнеса.

Към момента не е известно да е предлаган метод, който обединява различни софтуерни решения за системи за големи данни, които да извличат и обработват данни с ИИ от различни източници, към които да са приложени когнитивна и адаптивна сигурност, и които в същото време да са утвърдени, представени и обособени в единен оперативен център. Всяка система за големи данни има технически специфики, зависещи от особеностите на нейното приложение, и броя на активните потребители, мястото на дейност и технологиите на изпълнение. Представеното изследване може да помогне на специалисти, работещи в сферата на големите данни да повишат нивото на сигурност чрез прилагането на предложените принципи и методи, като изградят единен център за управление, който да осигурява различни нива на сигурност. Представеното решение може да бъде допълнително разработено и разширено с допълнителни функционалности, така че да отговаря на нуждите на различни видове бизнес и потребители.

2. Обект и предмет на изследването

Обект на изследването са принципи и методи за проектирането на оперативен център за управление на сигурността в системи за големи данни.

Предмет на изследването е извеждане на метод за изграждане на нов тип оперативен център за управление на сигурността, базиран на обработка с ИИ на неструктурирани данни, системи с големи данни и Управление на информацията за сигурността и събитията (Security information and event management - SIEM), интегрирани в обща архитектура.

Изследователският проблем на дисертационния труд, е съвместяването на технологии и предоставяне различни нива на сигурност при изграждането на оперативен център в системи за големи данни.

3. Цел и задачи на дисертационния труд

Целта на дисертационния труд е да се предложат принципи и методи за проектиране и изграждане на оперативен център, който да управлява информационната сигурност, за системи, опериращи с големи данни.

Задачите, които се изпълняват в процеса на реализация на поставената цел на дисертационния труд са:

- Да се проучат видовете данни и предимствата, и недостатъците на съществуващи средства за съхранение и обработка на големи данни;
- След анализ на съществуващи решения и добри практики в световен мащаб в тази област, да се разгледат принципи и методи за управление различните нива на информационна сигурност;
- Да се направи преглед на моделите и подходите при използването на изкуствен интелект за осигуряване по-високо ниво на сигурност;
- Да се направи преглед на оперативните центрове за сигурност (ОЦИС) като се разгледат различните поколения на ОЦИС;
- На база направения анализ да се обособи подход за проектиране на ОЦИС в системи за големи данни за управление на сигурността;
- Да се проектира функционална архитектура с различни нива на сигурност, която да предложи решение на дефинирания изследователски проблем в дисертационния труд и предложената архитектура да е в състояние да използва заложените принципи и методи за проектиране и изграждане на оперативен център за управление на сигурността;
- Да се извлекат данни от хетерогенни източници и да се съхранят в единна система за големи данни и да се установи ниво на сигурност при тяхното извличане и последващо обработване;
- Да се анализира нивото на сигурност в системата за съхранение и обработка на големи данни и да се дефинира ниво на адаптивна сигурност; да се разгледа вътрешно-организираната сигурност на системата за големи данни; да се обособи софтуерна бизнес сигурност;

- Да се предложат подходи за изграждане на оперативен център, който ще управлява сигурността при различните процеси в системи с големи данни;
- За постигане на поставената цел и задачи е използван подхода на проектиране и реализация на софтуерни системи „отдолу-нагоре“. При този подход процесът започва с проектиране на основните компоненти на системата, които след това следва да бъдат обединени, за да се създадат комплексни модули на оперативния център.

4. Работни хипотези на дисертационния труд

Основните работни хипотези за настоящия дисертационен труд са:

Хипотеза I:

Възможно е да се създаде модел, който да опише проектирането на оперативен център за управление на информационната сигурност на различни нива за системи с големи данни и извличането на данни от хетерогенни източници.

Хипотеза II:

Възможно е да се постигне ефективна интеграция на различни технологии и системи за работа с големи данни.

Хипотеза III:

Възможно е да се създаде многослойна архитектура с различни подходи за предоставяне на сигурност.

5. Научни и научно-приложни приноси

1. Изследвана е същността и компонентната структура на „оперативен център за сигурност“ и е дефинирано актуално определение и необходими елементи, спрямо съвременните условия за функциониране на такъв център в среда на големи данни.

2. Предложени са критерии за сравнение на технологии за обработка на големи данни, във връзка с целите на разработката по отношение на събирането, организацията и съхранението на неструктурирани данни, с възможности за прилагане на средства за изкуствен интелект.
3. Изведени са основни принципи и методи за управление на оперативен център за информационна сигурност, обхващащи процеса, функционалностите и нивата на сигурност.
4. За проектирането и изграждането на ОЦИС са предложени актуални принципи, необходими за обхващане на спецификата на управление на сигурността на системите, работещи в среда на големи данни.
5. Дефиниран е метод за проектиране и създаване на функционална архитектура на ОЦИС. Предложената архитектура е с три нива на управление на сигурност, обхващащи мрежово ниво на сигурност, процес по извличане и обработка на данни, удостоверяване на вътрешно-базирана сигурност в среда за големи данни и анализ на получените резултати.
6. Реализиран е прототип, изграден с технологиите на предложената архитектура с цел обхващане на всички необходими функционалности. Прототипът свързва технологиите NiFi, Micro Focus IDOL и Apache Hadoop и е тестван с данни от социални медии, видео поток, данни от уебсайтове и лог файлове.

6. Обем и структура на дисертационния труд

Дисертационният труд е в обем от 170 страници, от които 153 съдържат същинското изследване, без приложенията. Структурата на дисертационния труд се състои от:

- Въведение;
- Три глави, в които се реализира анализ на предметната област, извеждат се основни дефиниции, свързани с изследователския проблем, представят се предпоставките за реализиране на труда и се дефинира метод, с който да се постигне това. Предложеният метод е верифициран с използване на различни технологии;
- Заключение, което включва списък на научните и научно-приложните приноси, Списък на направените публикации по темата на дисертационния труд, Бъдеща

работа, Литература, Списъци на фигурите и на таблиците и Списък на термините и съкращенията.

Първа глава (въведението) се фокусира върху актуалността на разглеждания проблем, като разглежда обекта, предмета, изследователския проблем и целите на дисертационния труд. Дефинирани са три хипотези, които следва да бъдат доказани.

Втора глава представлява проучвателната част на научното изследване. В нея се поставят теоретичните основи на значението и видовете данни, разглеждат се средствата и подходите за обработка на големи данни и се прави сравнение по избрани критерии, които да помогнат за избирането на подходящите инструменти за дисертационния труд. Във втора глава се разглеждат приложенията на изкуствения интелект (ИИ) в сферата на сигурността, изведени са принципи и методи, които се използват за осигуряване по-високо ниво на сигурност в организациите и е представена дефиницията и компонентната структура на оперативен център за информационна сигурност (ОЦИС).

Трета глава се съсредоточава върху проектирането на ОЦИС. В нея се извеждат принципи и методи, които се прилагат за проектиране на център от такъв вид, разгледани са съществуващите поколения ОЦИС и се изведени предизвикателствата, пред тях. В следствие на тези неща се предлагат актуални принципи и се използват контроли на международния стандарт ISO 27001 за проектиране на ново поколение ОЦИС. В тази глава се предлага метод, който описва новия тип ОЦИС във функционална архитектура на три нива за управление на сигурността. Първо ниво обхваща адаптивна сигурност и интелигентна обработка на данни от хетерогенни източници, второ ниво се съсредоточава върху вътрешно-базираната сигурност на Hadoop системата за големи данни и софтуерната бизнес сигурност на процеси в организацията, докато трето ниво се съсредоточава върху визуализация на анализираната информация от предходните две нива и при наличие на потенциална уязвимост за сигурността - известяване на екипа на ОЦИС за взимане на конкретни действия.

Четвърта глава се фокусира върху доказване функционалността на предложената архитектура като използва различни технологии на всяко от предложените три нива.

Представен е работещ прототип с експериментална цел за верифициране на работещ процес, използвайки предложения ОЦИС.

В пета глава (заключение) се обобщава работата по дисертационния труд, извеждат се научните и научно-приложните приноси, както и се дава насока за бъдеща работа на научния труд.

Кратко изложение на дисертационния труд

1. Въведение

Въведението прави обзор на актуалността на избраната тема, обекта, предмета, дефинира задачите, които следва да бъдат изпълнени, както и формулира три хипотези. Целта на дисертационния труд е да се предложат принципи и методи за проектиране и изграждане на оперативен център, който да управлява информационната сигурност, за системи, опериращи с големи данни.

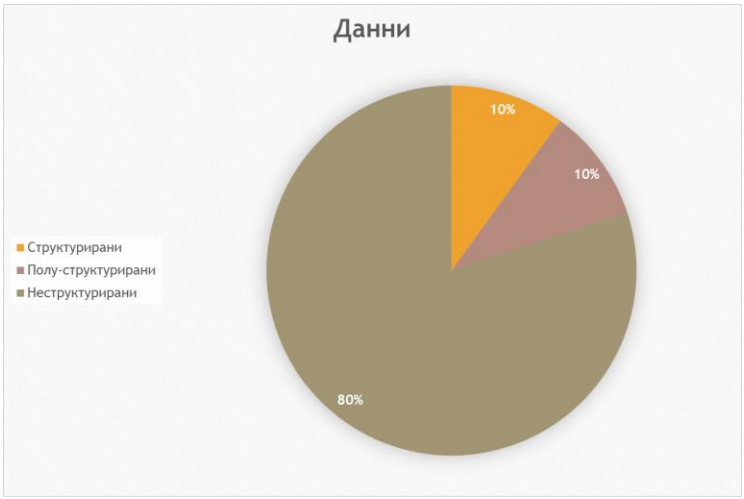
2. Сигурност в среда на големи данни и оперативен център за сигурност

Данните се използват от векове, но в днешно време дигиталните технологии са повсеместни с над 5 милиарда потребители в свързана Интернет среда [8]. Всяко цифрово действие генерира данни, които бизнесът може да използва, за да подобри ефективността и да открие нови възможности. Данните се класифицират в структурирани, полу-структурирани и неструктурирани типове и могат да предоставят ценна информация за вземане на стратегически решения и конкурентно предимство на пазара.

- **Структурираните данни** са организирани количествено измерими данни, често в таблична форма като SQL бази данни. Колоните имат типове данни - текстови или числови, а редовете имат специфични стойности. Тези данни са сигурни благодарение на контролирания достъп и установената технология. Те се използват в чувствителни приложения, като финанси или здравеопазване [9].

- **Полу-структурираните данни** нямат фиксирана схема и позволяват промени, без да се нарушава структурата им. Те са добри за динамични набори от данни и използват метаданни като тагове или атрибути, за да опишат формата, структурата и значението им. Обикновено се извличат в XML, HTML и JSON документи [10].

- **Неструктурираните данни** включват данни, генерирани от хора и машини, като публикации в социални медии, изображения и данни от сензори [11]. Той съставлява 80% от глобалните данни, поставяйки предизвикателства пред организациите, които се опитват да извлекат стойност [12] (Фигура 1).



Фигура 1. Обобщени дялове на генерирани типове данни в световен мащаб.

За разлика от структурираните и полу-структурираните данни, където има конкретни методи за обработка, при неструктурираните данни извличането на знание и извършването на анализ представлява проблем, тъй като разнообразието на формати изисква специализирани инструменти за тяхната обработка [13]. Част от тези инструменти са Apache Hadoop, Micro Focus IDOL, IBM Watson, AWS и др..

Анализът на информация, извлечена от неструктурирани данни е важен, защото с подобен подход организациите могат да получат представа за поведението на клиентите, пазарните тенденции и друга важна за бизнеса информация, която би било трудно или невъзможно да се получи по друг начин.

Големи данни

Големите данни са голяма, сложна и бързо нарастваща колекция от структурирани, полуструктурирани и неструктурирани данни, генерирани от различни източници като

социални медии, уеб търсачки и ИнО устройства. Статистически данни Statista за социалните медии показват, че през 2022 г. броят на потребителите е 4,26 милиарда по целия свят – бройката е прогнозирана, че е с нарастващ темп през следващите няколко години [14]. Само за периода октомври 2022 г. до януари 2023 г. има увеличение на броя на активни потребители на платформата Facebook по целия свят с 5 милиона [15]. По-голямата част от глобалния растеж на социалните медии се дължи на нарастващото използване на мобилни устройства, с които потребителите пишат публикации, изразяват мнения, качват снимки или реагират на публикации на други потребители. Освен социалните медии, масово се използват и уеб-търсачките за различни запитвания. Проучване показва, че към края на 2022 г. Google получава 8,5 милиарда заявки дневно, което се равнява на почти 100 хиляди заявки за секунда [16]. Тези данни сочат, че генерираната информация постоянно нараства.

Големите данни могат да бъдат най-добре определени с шестте V's: обем, скорост, разнообразие, достоверност, стойност и променливост (volume, velocity, variety, veracity, value, and variability). Първоначално големите данни често са представяни като 3-те V – обем, скорост и разнообразие, но в литературата може да се срещнат и до 42 V [17].

- ***Обемът*** се отнася до огромното количество данни, генерирано и събрано от лица, организации и машини.
- ***Скоростта*** се отнася до скоростта, с която данните се генерират, обработват и анализират в реално време.
- ***Разнообразието*** се отнася до разнообразието от типове данни, източници и формати, включително структурирани, полу-структурирани и неструктурирани данни.
- ***Достоверност*** се отнася до надеждността, точността и последователността на данните.
- ***Стойността*** се отнася до извличане на знания и ползи, които могат да бъдат получени от анализирането и използването на големи данни.

- ***Променливостта*** се отнася до непредсказуемостта на данните, генерирани и събрани от различни източници, включително социални медии, ИнО устройства и сензори.

Обемите данни, събирани от бизнеса, нарастват експоненциално и с развиването на технологиите се разпространяват в локални, облачни и хибридни системи [18]. Това увеличава както сложността на управлението и достъпа, така и обработката на данните. Поради това се развиват и надграждат системи за големи данни като този процес се очаква да продължава и в бъдеще.

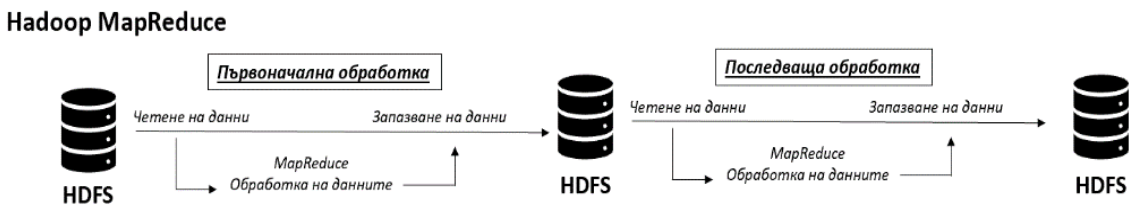
Системи за големи данни и тяхната обработка

За съхраняване и обработка на големи данни са разработени специализирани системи. Тези среди използват разпределени файлови системи и техники за паралелна обработка за съхраняване и обработка на големи обеми от данни. Примери за системи за големи данни са Apache Hadoop, Amazon Redshift, Snowflake и NoSQL бази данни сред които са MongoDB и Apache Cassandra [19], [20].

В труда са разгледани Apache Hadoop, Apache Spark, Apache Hive, Micro Focus IDOL и Apache NiFi.

- ***Apache Hadoop***

Hadoop е система с отворен код за съхранение и обработка на различни типове данни в разпределена среда. Състои се от HDFS за съхраняване на данни и MapReduce за паралелна обработка между възли [21]. За да може да обработи данните, Hadoop системата първо трябва да ги съхрани в директориите си, за което отговаря HDFS (Фигура 2).

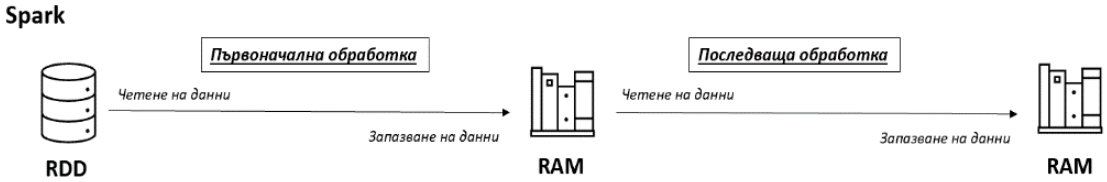


Фигура 2. Процес на обработка с Hadoop.

Hadoop клъстерите могат да се мащабират хоризонтално, за да обработват големи количества данни. MapReduce разделя данните на по-малки части и ги обработва паралелно, като чете от HDFS за всяка стъпка. Резултатите се съхраняват в HDFS или се изпращат за допълнителен анализ. Hadoop може да се интегрира с други инструменти за големи данни като Spark за разширен анализ [22].

- ***Apache Spark***

Spark е софтуерно решение за обработка на големи данни с отворен код, което има висока производителност, гъвкавост и възможност за използване на ИИ. Той предоставя API за популярни езици за програмиране и се състои от Spark Core, Spark SQL, Spark Streaming и MLlib [23]. Spark Core предоставя функционалност за разпределена обработка и използва Resilient Distributed Datasets (RDD) като основна структура на данните (Фигура 3).

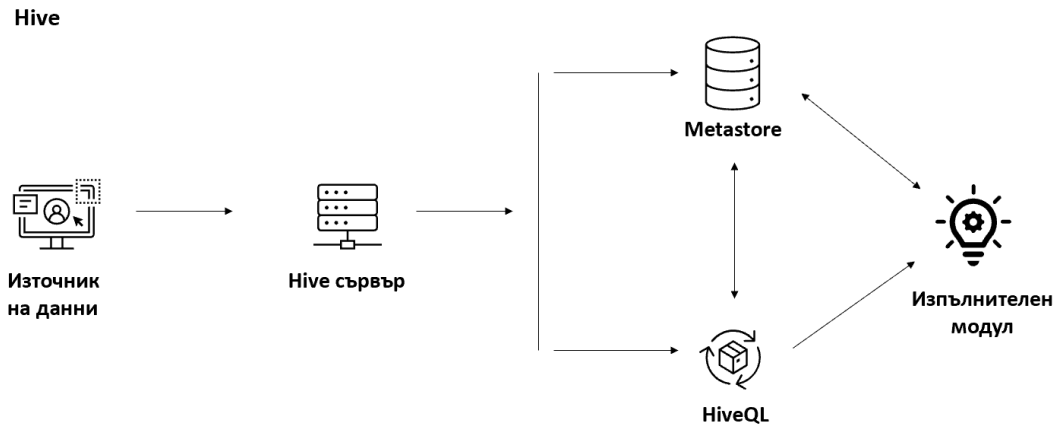


Фигура 3. Обработка на данните със Spark.

RDD осигуряват устойчивост на грешки и паралелизъм и могат да бъдат създадени от различни източници на данни. Spark SQL позволява SQL базирани заявки, Spark Streaming позволява обработка в реално време, а MLlib предоставя възможности за машинно обучение. Spark чете данни от RDD и ги съхранява в RAM за по-бърза обработка и паралелизира процеса в множество възли в клъстер.

- ***Apache Hive***

Hive е инструмент за съхранение на данни за обработка и заявки за големи набори от данни, съхранявани в HDFS на Hadoop или други съвместими файлови системи. Той поддържа различни формати на данни и използва език за заявки, наречен HiveQL, базиран на SQL. Hive се състои от компоненти като процесор за заявки (Фигура 4), Metastore, сървър и изпълнителен модул, които работят заедно за обработка на големи данни [24].

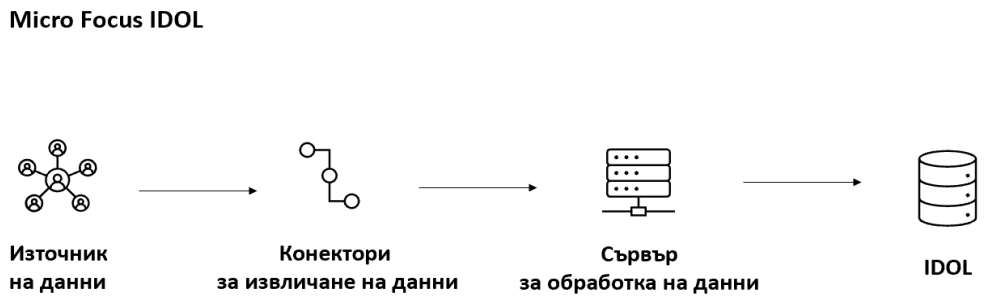


Фигура 4. Обработка на данните с Hive.

Hive не е подходящ за обработка на неструктурирани данни като текст, изображения или видео и е оптимизиран за групова обработка. Може да се използва с други инструменти за разширен анализ.

- ***Micro Focus IDOL***

IDOL (Intelligent Data Operating Layer) използва техники като обработка на естествен език, ИИ и семантичен анализ, за да извлече знания от различни типове данни. Системата има няколко компонента като конектори, индексатор, манипулатор на разпределен индекс, анализатор на съдържание и сървър за заявки (Фигура 5) [25].



Фигура 5. Обработка на данни с IDOL.

IDOL използва машинно обучение и семантичен анализ, за да прави прогнози, да класифицира данни и да идентифицира връзките между различни данни. Обработените данни и резултати се съхраняват в хранилището на IDOL за големи данни.

Сравнение на средите за големи данни по избрани критерии

На база представените технологични решения, следва обобщение на техните възможности по избрани критерии в Таблица 1. Тези критерии са важни, защото гарантират, че данните могат да се управляват ефективно и се използват за стимулиране на бизнес успеха [26], [27].

Представените критерии в Таблица 1 са подбрани в зависимост от конкретните цели и задачи в текущия дисертационен труд. Съпоставени са представените технологични продукти спрямо мястото за съхранение и обработка на големи данни, както и е насочено вниманието към възможността за обработка на неструктурирани данни, тъй като те играят ключова роля за успеха на бизнеса. За да може да се вземе решение кои инструменти да се използват, е представено сравнение на типа обработка на данните. Това е важно условие, тъй като някои от продуктите не поддържат обработка в реално време, което би било от важно значение в критични моменти за организацията. На следващо място е обърнато внимание на мястото за съхранение на данните по време на обработка, тъй като това може да повлияе на ефективността и скоростта на обработката на данни, както и на общите разходи за съхранение на данни. За да се гарантира сигурността на данните при потенциална уязвимост или криптиране на данните на някой от сървърите, е важно да се представи критерия за устойчивост на грешки. За анализа на данните от различните типове с правилно разбиране на смисъла на данните, е подходящо използването на ИИ. Той би допринесъл за подобрена скорост при анализа на данни и извличане на зависимости, тенденции и модели.

Таблица 1. Сравнение между технологичните среди за големи данни.

Критерий	**Hadoop**	**Spark**	**Hive**	**IDOL**
Съхранение и обработка на големи данни	Файлова система HDFS за съхранение и MapReduce за групова обработка на данни	Файлова система HDFS за съхранение и устойчив разпределен набор от данни RDD за обработка на данни	Файлова система HDFS за съхранение и използване на HiveQL за заявки в Hadoop	Конектори за извличане на данни и IDOL сървър за обработка и анализ
Обработка на неструктури-рани данни	Да	Да	Не	Да
Тип обработка	Обемна пакетна обработка на исторически данни	Обработка в реално време и пакетна обработка на данни	Пакетна обработка на структурирани данни исторически данни	Обработка в реално време и пакетна обработка на данни
Съхранение на данните по време на обработка	На диска на клъстер в HDFS системата	В RAM паметта	На диска на клъстера в HDFS системата	Съхранява в RAM паметта или на диска на клъстер
Устойчивост към грешки	Да, чрез репликиране на данните с HDFS върху възлите от клъстера	Да, чрез абстракция на RDD с репликиране върху възлите от клъстера	Да, чрез репликиране на данните с HDFS върху възлите от клъстера	Да, чрез конфигуриране с излишни възли за наличност и автоматично прехвърляне на възел

Използване на ИИ	Не	Да	Не	Да

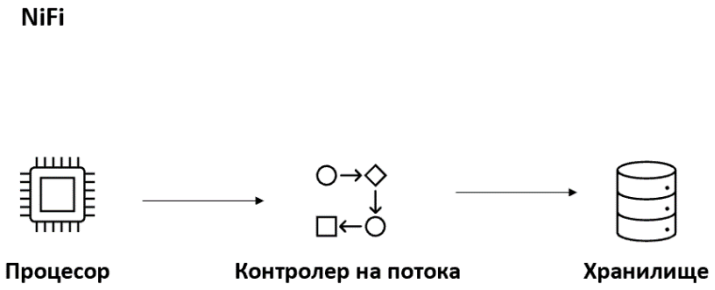
Таблица 1 сравнява силните и слабите страни на четири инструмента за обработка на големи данни: Hadoop, Spark, Hive и IDOL. Hadoop е подходящ за групова обработка, но може да бъде бавен за обработка в реално време. Spark е по-бърз, но по-сложен и изисква повече памет от Hadoop. Hive има лесен за използване интерфейс, но не е подходящ за обработка в реално време или неструктурирани данни. IDOL обработва неструктурирани данни с AI и има толерантност към грешки. И четирите инструмента репликират данни за бързо възстановяване в случай на пробиви в системата.

На база получените изведените характеристики, се взима решение кои технологични решения ще бъдат използвани и кои от тях биха могли да бъдат комбинирани за реализиране на крайния резултат на текущия труд.

За интеграция на инструментите за обработка и анализ на данните, се използва системата NiFi.

- ***Apache NiFi***

NiFi е платформа с отворен код за автоматизиране на потока от данни между различни системи. Той има различни компоненти като процесор, контролер на потока, файлово хранилище и връзка, които работят заедно, за да предоставят възможности за интегриране на данни (Фигура 6) [28].



Фигура 6. Процес на работа на NiFi.

NiFi може да насочва и разпространява данни въз основа на различни критерии, да трансформира данни по различни начини и да валидира качеството на данните. Важно е да се интегрират различните системи за максимална полза. Трите основни модела на изкуствен интелект (обработка на естествен език, машинно обучение и дълбоко обучение) също се изследват за откриване на уязвимости и пробиви в системата.

Всички изброени до тук системи имат различни приложения. За максимална полза от тях е подходящо интегрирането им в единна система, която да удовлетворява бизнес нуждите. За да се вземе решение кои от вече разгледаните системи да бъдат използвани за постигане на резултатност, се разглеждат трите основни модела на изкуствения интелект – обработка на естествен език, машинно и дълбоко обучение и тяхното приложение в областта на сигурността, за да се придобие представа относно предоставените възможности за откриване на уязвимости и пробиви в системата.

Приложение на изкуствен интелект за сигурност при големи данни

Развитието на изкуствения интелект е едно от най-значимите технологични постижения в последно време. ИИ започва да се прилага на много места и употребата му в различни области на приложение нарасна осезаемо през последните години. Примери за използването му в ежедневието ни са смартфоните, където чрез разпознаване на лица или пръстови отпечатъци се отключват. Друг пример е откриването на измами с кредитни карти, при които алгоритмите на ИИ анализират модели и аномалии в транзакционните данни. Това означава, че ако дадена транзакция е значително по-голяма от обикновено или ако клиентът започне да прави покупки, които не са характерни за неговото поведение, ИИ може да я маркира за по-нататъшно разглеждане като извести администратора на системата за потенциална заплаха [29].

ИИ се базира на идеята за симулация на човешкия интелект и прилагането му в дигитални устройства, които са програмирани да изпълняват задачи, обикновено изискващи човешки интелект, който от своя страна се асоциира с процеси като учене, решаване на проблеми, вземане на решения и визуално възприемане на генерираната информация [30]. Недостатък при използването на ИИ е ако бъде обучен от данни, които съдържат пристрастия, може да се доведе до неточни резултати, което от своя страна да

бъде сериозно предизвикателство да се идентифицират и коригират грешки [31]. За да се преодолее този недостатък е необходимо прилагането на техники за предварителна обработка като почистване на данни и нормализиране на данните за намаляване на отклонението в данните.

ИИ се използва в сигурността на големи данни за защита срещу кибератаки чрез анализиране на модели на поведение на потребителите, откриване на аномалии в данните и разработване на прогнозни модели на заплахи. Тези ИИ приложения могат да маркират потенциални заплахи и да помогнат за предотвратяване на бъдещи атаки. ИИ моделите се използват и в информационната сигурност за подобряване на точността и ефективността на системите за сигурност.

Разгледани са три модела на използване на ИИ в сигурността: обработка на естествен език (natural language processing), машинно обучение (machine learning) и дълбоко обучение (deep learning), които предоставят различни методи за управление на сигурността.

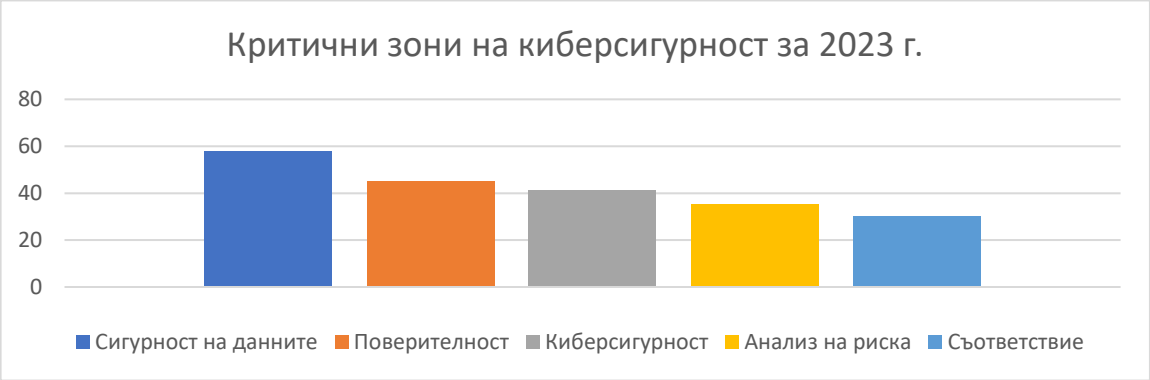
Принципи и подходи за управление на информационната сигурност

Информационната сигурност е проблем с важно значение за бизнеса, тъй като все повече данни се съхраняват и обработват онлайн или в системи, които са свързани онлайн. Въпреки това, сигурността се отнася не само до защитата на данните и процесите по обработка и анализ, в онлайн, но и във физическа среда.

Според направено проучване и данни на Astra, всеки ден се реализират около 2200 хакерски атаки, което прави по една хакерска атака в онлайн пространството на всеки 39 секунди [32]. Хакерската атака е опит на неупълномощено лице да получи достъп или да повреди компютърна система, мрежа или данни. Защитата срещу тези атаки е чрез прилагане на силни мерки за сигурност като защитни стени, системи за откриване и предотвратяване на проникване, контрол на достъпа и обучение на служителите, за да разпознават и избягват атаки като социално инженерство [32].

Според проучване на Statista от края на 2022 г., петте най-критични области в киберсигурността за 2023 са сигурността на данните, последвана от поверителност на личните данни, киберсигурност, анализ на риска и съответствие [33]. Последното гарантира, че мерките за сигурност на организацията са достатъчни за защита на

чувствителна информация и предотвратяване на неоторизиран достъп. Сигурността на данните е най-високо оценената зона като критична в киберпространството - Фигура 7. Този вид сигурност трябва да се разглежда като непрекъснат процес, част от културата на организация, а не като еднократно събитие.



Фигура 7. Зони на киберсигурност.

Значението на управлението на сигурността на данните не може да бъде надценено, тъй като при нарушение, при използването им, може да има тежки последици за всички засегнати лица. Изтичането на бизнес данни е един от важните и разпространени проблеми на сигурността на данните, тъй като наред с корпоративните данни, компаниите пазят и чувствителни данни за своите клиенти.

За да може сигурността да се управлява, трябва да бъдат спазени няколко принципа, които помагат на организациите да проектират и прилагат ефективни мерки, за да гарантират поверителността, целостта и наличността на тяхната информация [34]:

- *Поверителност*: Този принцип гарантира, че чувствителната информация се пази в тайна и е достъпна само за упълномощени лица. Това може да се постигне чрез методи като криптиране на данни, контрол на достъпа и редовни проверки на сигурността.
- *Цялостност*: Този принцип гарантира, че данните остават точни и непроменени, както по време на пренос, така и в покой. Това може да се постигне чрез методи като архивиране на данни, хеширане, контрол на версиите и сигурни практики за кодиране.

- *Наличност*: Този принцип гарантира, че данните и ресурсите са достъпни за упълномощени лица, когато имат нужда от тях. Това може да се постигне чрез методи като резервни системи, планиране на възстановяване след бедствие и споразумения за ниво на обслужване.

- *Удостоверяване*: Този принцип гарантира, че потребителите са тези, за които се представят, преди да им предостави достъп до ресурси. Това се постига чрез методи като многофакторно или многостепенно удостоверяване като изпращане на съобщение по електронната поща или по мобилното устройство, въвеждане на код от картинка, въвеждане на втора парола и т.н..

- *Упълномощаване*: Принципът на упълномощаване гарантира, че потребителите имат подходящото ниво на достъп до ресурси въз основа на тяхната роля или ниво на правомощия. Чрез осигуряване на оторизация организациите могат да ограничат потенциалните щети, които могат да бъдат причинени от пробив в сигурността или атака.

- *Неотричане*: Този принцип гарантира, че автентичността на данните или транзакциите не може да бъде отречена от участващите страни. Това се постига чрез методи като цифрови подписи и протоколи за защитена комуникация.

Докато принципите на сигурността осигуряват рамка за ефективна информационна сигурност, предотвратяването, откриването и реагирането са специфични действия, които организациите предприемат, за да приложат тези принципи и да защитят своята чувствителна информация.

След извеждането на принципи за ефективна програма за информационна сигурност, следва описание на важни методи, които се използват за постигане на целите на информационната сигурност. Има много различни методи за сигурност, използвани за защита на системи, данни и мрежи от неоторизиран достъп, кражба или повреда [3]. Избрани са конкретни методи, които са широко признати и приети като ефективни начини за справяне със специфични рискове и уязвимости за сигурността.

- **Сигурност на процеса**: Внедряване на контроли и процедури, за да се гарантира, че бизнес процесите са защитени и не са уязвими на атаки или пробиви,

включително политики за поверителна информация, обучение за най-добри практики за сигурност и одити на сигурността.

- **Нулево доверие**: Разширяване на мерките за сигурност към всички устройства, приложения и потребители в една организация, изискване на удостоверяване и оторизация, преди да бъде предоставен достъп, и фокусиране върху защитата на отделни устройства и активи с данни чрез многофакторно удостоверяване, контрол на достъпа и непрекъснат мониторинг.

- **Многофакторно удостоверяване**: Изискване от потребителите да предоставят две или повече форми на удостоверяване, за да потвърдят своята самоличност, преди да бъде предоставен достъп до системи или данни, намалявайки риска от неоторизиран достъп и атаки за разкриване на парола.

- **Сигурност на софтуера**: Използване на свързани със сигурността програмни практики за изграждане, тестване и корекции на системата, включително проектиране и кодиране на софтуер за предотвратяване на често срещани уязвимости като препълване на буфер и атаки чрез инжектиране, тестване за уязвимости и прилагане на актуализации и корекции за адресиране на уязвимости.

    - **Многостепенна сигурност:** Прилагането на многостепенната сигурност (Multilevel security - MLS) е начин защита на информацията в компютърни системи, която съдържа данни с различни нива на чувствителност или класификации. Многостепенната сигурност предлага по-високо ниво на защита, отколкото използването на едно ниво на сигурност за цялата информация, тъй като позволява на организациите да приспособят своите мерки за сигурност към специфичната чувствителност на защитаваната информация.

На следващо място са разгледани системите за сигурност в среда за големи данни, тъй като това има отношение към разгледания проблем в текущия труд, в който се съчетават сигурност и среди за работа с големи данни.

Системи за сигурност в среда за големи данни

Значението на данните за получаване на конкурентно предимство е широко признато, което прави критично важно да се осигури съхранението и обработката на чувствителни данни. За да постигнат по-високо ниво на сигурност, организациите могат да внедрят механизми за удостоверяване и оторизация като LDAP, Active Directory и Kerberos. Технологии за криптиране като SSL/TLS протоколи, криптиране на данни в покой и управление на ключове могат да бъдат използвани за защита на чувствителни данни. От голяма значение е поддържането на регистрационни файлове за одит и осигуряването на съответствие с регулаторни изисквания. Мерките за мрежова сигурност, като конфигурация на защитна стена, мрежово сегментиране и VPN, могат да помогнат за предотвратяване на неоторизиран достъп до Hadoop възли и защитена комуникация между възли. В сложната среда за сигурност на системите за големи данни е необходим централизиран механизъм за удостоверяване, за да се гарантира, че само оторизирани потребители и услуги имат достъп до данните. Разгледани са Kerberos, Apache Knox и Apache Ranger като решения, които осигуряват ниво на сигурност при автентикирането и оторизирането на потребителите. Apache Knox служи като сигурна входна точка за REST и HTTP взаимодействия с екосистемата Hadoop [35], докато Apache Ranger предоставя платформа за управление на контрол на достъпа и политики за одит в екосистемата Hadoop [36]. Kerberos се използва за защита на процеса на удостоверяване и криптиране на данни, обменяни между клиенти и сървъри [37].

След разглеждането на основните компоненти и методи, следва обединение в единно решение, което да обхване всички аспекти на сигурност и да гарантира защитата на системата на различни нива. Едно подобно решение може да бъде оперативен център за информационна сигурност.

Оперативни центрове за информационна сигурност

Оперативните центрове за информационна сигурност играят ключова роля в защитата на тези среди и гарантирането на поверителността, целостта и наличността на чувствителни данни. Причината в използването и популяризирането на такива центрове се дължи главно на необходимостта от предотвратяване на големи кибер инциденти и произтичащото от това приемане на централизирани действия за сигурност в организации.

По своята същност Оперативният Център за Информационна Сигурност (ОЦИС, Information Security Operation Center - ОЦИС) може да бъде определен като централизирано съоръжение, което отговаря за наблюдението и управлението на сигурността на дадена организация. В ОЦИС обикновено има екип от анализатори и специалисти по сигурността, които имат за задача да откриват, анализират и реагират на инциденти по сигурността в реално време [38]. ОЦИС предоставя механизми за събиране, съхраняване и нормализиране на всякакъв вид данни, както и гарантира по-високо ниво на сигурност. Основната функция на ОЦИС е да осигури ситуационна осведоменост и функции за реагиране на инциденти на организацията. Това включва наблюдение на мрежовия трафик, откриване на инциденти и разследване на настъпили събития, свързани със сигурността, и координиране на екипа за реакция при инциденти [39].

Целта на ОЦИС е осигуряването на платформа за сътрудничество за разработване на мащабируем инструмент за анализ на сигурността, като в същото време центърът поддържа допълнителни функции за идентифициране на проблеми в сигурността [38].

ОЦИС се характеризира с използване на автоматизация и анализи за откриване и реагиране на заплахи за сигурността. Някои от автоматизираните му процеси са събиране и анализ на регистрационни файлове, сортиране и реагиране на инциденти и събиране на информация за заплахи (Фигура 8).

Фигура 8. Процеси на ОЦИС.

Този тип центрове интегрират множество технологии за сигурност и източници на данни, за да осигурят цялостен поглед върху състоянието на сигурността на организацията, включително интегриране на системи за информация за сигурността и управление на събития (SIEM), подходи за разузнаване на заплахи и други инструменти за сигурност в една платформа. Ключова характеристика на ОЦИС е непрекъснатото подобрение на центъра чрез редовни оценки и прегледи на процесите и процедурите за сигурност, използвайки показатели и анализи за измерване на ефективността на операциите за сигурност, идентифициране на области за подобрение и съответно коригиране на процеси и процедури. ОЦИС наблюдава спазването на политиките и разпоредбите за сигурност, като HIPAA, GDPR и др. [38].

ОЦИС за системи с големи данни са изправени пред нови предизвикателства поради големия обем и сложността на този вид данни. Има редица съществуващи решения за управление на информационната сигурност в системи с големи данни. Те включват анализ на поведението на потребителите и субектите (User and Entity Behavior Analytics - UEBA), предотвратяване на загуба на данни (Data Loss Prevention - DLP), управление на самоличността и достъпа (Identity and Access Management - IAM), разузнаване на заплахи и машинно обучение. UEBA открива аномалии в поведението на потребителите и

идентифицира потенциални заплахи за сигурността, докато DLP предотвратява изтичането на данни, а IAM управлява потребителския достъп до данни и системи. Разузнаването на заплахи (Threat Intelligence) предоставя информация в реално време за потенциални заплахи за сигурността и уязвимости, а машинно обучение идентифицира модели и аномалии в големи набори от данни, за да открива инциденти със сигурността по-бързо и точно.

Когато се проектира ОЦИС за системи с големи данни е добре да използва комбинация от гореспоменатите решения, за да осигури цялостен мониторинг и анализ на сигурността в реално време. Както и екипът на ОЦИС трябва да е в състояние да интерпретира данните и да реагира бързо на инциденти, свързани със сигурността.

Изводи

Нарастващата дигитализация на различни процеси доведе до генерирането на огромни количества данни онлайн, включително неструктурирани данни от социални медии и други платформи. За събирането и обработката на тези данни се използват големи данни и ИИ алгоритми, които също могат да откриват измами и да подобряват бизнес операциите. Запазването на тези данни и гарантирането на тяхната сигурност обаче е предизвикателство и различни мерки като криптиране на данни и многопластова сигурност трябва да се прилагат и наблюдават постоянно.

След реализираното обобщение на възможностите на средите за големи данни по избрани критерии, съобразени с нуждите на дисертационния труд, избрахме продукти, които ще използваме при проектирането на оперативен център за информационна сигурност. Такива продукти са Apache Hadoop и системата IDOL, които могат да се използват за съхраняване и анализиране на големи количества данни и за извличане на данни от социални медии. Тези продукти могат да бъдат интегрирани, за да осигурят допълнителна функционалност, използвайки Apache NiFi.

3. Проектиране на архитектура на оперативен център за информационна сигурност

Дигитализацията трансформира революционизира бизнес изискванията и използването на големи данни се превърна в ключов елемент за постигане на конкурентно предимство. В това отношение оперативните центрове по информационна сигурност (ОЦИС) стават незаменими за бизнеса, тъй като гарантират сигурността на данните. По-конкретно, системите за големи данни играят основна роля в киберсигурността и информационната сигурност. Тяхната роля се изразява в събирането и анализирането на обемни данни за откриване на потенциални киберзаплахи. Въпреки това, за ефективното внедряване на ОЦИС, правилният дизайн е наложителен и той трябва да се изпълнява от професионалисти с голям опит с ясни планове за реакция при инциденти.

За проектиране на ОЦИС трябва да се следват установени принципи и методи. Принципите предоставят насоки за вземане на решения, докато методите се отнасят до специфични техники и процеси, насочени към постигане на целите. От решаващо значение е да се отбележи, че за създаването на работещ проект за ОЦИС изисква цялостно разбиране на изискванията, рисковете и целите на организацията. Следователно са представени принципи и методи за проектиране на ОЦИС, които изграждат рамка за разбиране на изискванията, рисковете и целите на организацията.

Принципи за проектиране на оперативен център за информационна сигурност

При проектирането на ОЦИС за системи с големи данни трябва да се вземат предвид няколко принципа [40], [41].

Първо, мониторингът е от решаващо значение за функциониращ ОЦИС. Този принцип включва откриване на злонамерени атаки и наблюдение на злонамерени дейности от служители, подизпълнители, гости и външни лица. Ефективното наблюдение позволява бързо идентифициране на заплахи и улеснява ефективното сътрудничество между персонала по сигурността [41].

Второ, анализът на събраните данни е от решаващо значение за идентифицирането на заплахите, когато се появят. Анализът в реално време е от решаващо значение за реагиране на заплахи за сигурността [40].

Трето, реакцията при инциденти е жизненоважна за справяне с инциденти със сигурността, независимо дали са вътрешни или външни. Вътрешни инциденти могат да възникнат от служители, изпълнители или партньори с достъп до системите на организацията, докато външните инциденти се инициират от нападатели извън организацията [41].

Четвърто, откриването на пробиви и реакцията изискват цялостен план за реагиране при инцидент, който очертава стъпките, които трябва да бъдат предприети в случай на инцидент със сигурността. Планът трябва да включва процедури за ограничаване, разследване и възстановяване [41].

Пето, одитът на регистрационни файлове включва анализиране на регистрационни файлове на всички устройства и корелиране на събития в различни регистрационни файлове. ОЦИС играе важна роля в регистрирането и одита, като проверява съответствието и документира отговора на инциденти със сигурността [41].

Шесто, редовното тестване и оценка са критични за идентифициране на потенциални уязвимости и гарантиране, че съществуващите мерки за сигурност ефективно намаляват рисковете [41].

На седмо място, мащабируемостта е важна, тъй като големите данни могат бързо да нараснат по размер и скорост. ОЦИС трябва да бъде проектиран така, че да се мащабира хоризонтално чрез добавяне на повече сървъри или вертикално чрез увеличаване на процесорната мощност на съществуващите сървъри [40].

Осмо, поверителността на данните е от съществено значение, за да се гарантира, че данните не са компрометирани от вътрешни или външни нападатели. Трябва да се прилагат методи за криптиране и контрол на достъпа, за да се защитят данните при пренос и в покой [40].

И накрая, автоматизацията е от решаващо значение за бързото идентифициране на събития, свързани със сигурността, и навременната реакция на такива събития. Автоматизацията намалява риска от ръчни грешки и подобрява инструментите и процесите за наблюдение на сигурността [40].

Тези принципи служат като рамка за проектиране на ОЦИС за система за големи данни, която осигурява цялостна защита. Въпреки това, за да се разберат напълно стъпките на проектиране на ОЦИС, трябва да се разгледат различни методи.

Методи за проектиране на оперативен център за информационна сигурност

Проектирането на оперативен център за информационна сигурност за системи с големи данни изисква внимателно планиране и изпълнение. Ето някои методи, които важни и полезни при проектирането и внедряването на ОЦИС за системи за големи данни. Тяхната подредба е въз основа на значението им за гарантиране на сигурността и поверителността на системите за големи данни [42].

- **Оценка на риска** - извършването на цялостна оценка на риска е една от първите стъпки при проектирането на ОЦИС за системи за големи данни. Това помага да се идентифицират потенциалните уязвимости и рискове, свързани с данните, инфраструктурата и приложенията, които ще бъдат наблюдавани от ОЦИС [42].

- **Инструменти за управление на информация за сигурността и събития (SIEM)**: внедряването на инструменти за SIEM е от съществено значение за събиране, обобщаване и анализиране на данни за събития за сигурност от различни източници в реално време.

- **Контрол на достъпа** - прилага се строг контрол на достъпа, за да се ограничи броя на хората, които имат достъп до чувствителна информация, данни и системи [42].

- **Криптиране** - прилагат се механизми за криптиране на данни в покой и при пренос, което помага на защитата на чувствителни данни от неоторизиран достъп [43].

- **Сегментиране на мрежата** - мрежата се сегментира, за да се създадат отделни зони за различни видове данни и потребители. Това ще помогне за предотвратяване на неоторизиран достъп и ще ограничи потенциалните щети, причинени от пробив в сигурността [42].

- **Съответствие** - ОЦИС трябва да спазва всички съответни разпоредби и стандарти относно чувствителни или лични данни, като например Общия регламент за защита на данните (GDPR) и Стандарта за сигурност на данните на индустрията на платежните карти (Payment Card Industry Data Security Standard - PCI DSS) [81].

- **Обучение и повишаване на осведомеността** - използване на програми за обучение и повишаване на осведомеността, за информиране на служителите и потребителите за значението на информационната сигурност и за това как да идентифицират и докладват за потенциални заплахи за сигурността. Осигуряването на обучение и информираност за сигурността на целия персонал, участващ в операциите на системите за големи данни, помага за намаляване на риска от инциденти със сигурността, причинени от човешка грешка [81].

- **Непрекъснато подобрение** - прилагане на програма за непрекъснато усъвършенстване за редовен преглед и актуализиране на ОЦИС, за да се гарантира, че той остава ефективен и актуален [42].

За да се надграждат тези принципи и методи чрез добавяне на нови технологии, инструменти и процеси за подобряване на работата на ОЦИС, е необходимо да разгледаме съществуващите поколения архитектури на този тип центрове.

Поколения архитектури на оперативен център за сигурност

Поколенията на ОЦИС се отнася до еволюцията и съзряването на тези центрове с течение на времето. Съществуващите поколения се класифицират на четири етапа, всеки с различна степен на възможности, процеси и технологии [43]. Представени са на Фигура 9.

Фигура 9. Поколения ОЦИС.

Първото поколение се характеризира с липса на интеграция и автоматизация между инструментите за сигурност, докато второто поколение разширява възможностите отвъд мониторинга на предупрежденията за сигурност, за да включва проактивни операции за сигурност. Третото поколение включва усъвършенствани възможности за автоматизация и оркестрация, а четвъртото поколение включва интегрирането на усъвършенствани възможности за анализ и машинно обучение.

Тъй като всяко поколение на ОЦИС надгражда предишното, работата на екипа по сигурността се измества от събиране на регистрационни файлове (логове) и реакция след инцидент към разработване на нови възможности, процеси и технологии, които подобряват способността за откриване, предотвратяване и реагиране на заплахи за сигурността. Развитието на ОЦИС е непрекъснат процес, тъй като сигурността продължава да се развива, като с течение на времето се появяват нови заплахи. Разработването и интегрирането на нови технологии, инструменти и процеси ще продължи да бъде важно за подобряване на работата на ОЦИС. Тъй като тези центрове са изправени пред няколко предизвикателства, които трябва да бъдат адресирани за успешното проектиране и изграждане на следващото поколение от тези центрове.

Едно от основните предизвикателства е нарастващата сложност на системите, които трябва да се управляват централно и автоматично, което води до по-голям брой случващи

се събития и увеличаване на рисковете, свързани със сигурността на системите [44]. Въпреки това, липсата на контекст в данните, генерирани от системите за сигурност, може да затрудни разграничаването между истински заплахи и фалшиви положителни резултати. Освен това съществуващите информационни центрове за сигурност могат да бъдат претоварени от големия обем данни, което води до генериране на множество предупреждения и намалена способност за идентифициране на реални заплахи [44].

Друго предизвикателство е ограничената видимост в цялата организация. Съществуващите информационни системи за сигурност често се фокусират върху специфични области или функции на организация, което може да доведе до „слепи“ зони, които затрудняват откриването и реагирането на заплахи за сигурността, които възникват извън обсега на центъра [44]. Растежът на миграцията към облачни услуги и системи също налага нови изисквания към ОЦИС и скоростта, с която могат да се развиват съвременните заплахи, прави настоящите поколения уязвими към нови техники за атака [44].

За да се преодолеят тези предизвикателства, е започнат процес по проектиране на ново пето поколение ОЦИС. Това поколение използва технологии като изкуствен интелект, машинно обучение и автоматизация, за да позволи откриване на заплахи в реално време и реакция в цялата организация [45]. Едно от основните предимства на ОЦИС от пето поколение е способността му да анализира големи количества данни от множество източници в реално време и възприема по-проактивен и стратегически подход към сигурността, като се фокусира върху лов на заплахи и превенция, а не само реакция на инцидент [45]. Очаква се петото поколение да се съсредоточи върху сигурността в облака, справяйки се с уникалните предизвикателства пред сигурността, представени от преминаването към базирани на облак архитектури и нарастващото използване на контейнери и микроуслуги.

Някои доставчици предлагат усъвършенствани решения за оперативни центрове за сигурност, които включват ИИ за анализи и по-автоматизирани процеси, наричайки ги „пето поколение“ Такива компании са IBM, Atos, ProCyb и др. [46], [47], [48]. Понастоящем обаче няма общоприето определение за това какво точно представлява ОЦИС от пето поколение.

Тъй като заплахите и технологиите непрекъснато се развиват, от решаващо значение е непрекъснато да се оценяват и актуализират принципите на проектиране на центъра за операции по сигурността. Постоянно се откриват нови заплахи и уязвимости и нападателите винаги намират нови начини да ги използват. В същото време се разработват нови технологии, които могат да помогнат на организациите да откриват по-добре и да реагират навременно на инциденти със сигурността. Въз основа на анализа на съществуващите принципи и методи и представената информация за поколенията и различните предизвикателства, пред които са изправени, няколко актуални принципа могат да улеснят проектирането на ОЦИС и да доведат до създаването на метод, който оформя стъпките преди създаване на нов тип оперативен център за информационна сигурност.

Актуални принципи за проектиране на оперативен център за сигурност

Поради непрекъснато развиващите се заплахи и технологии, променящите се бизнес изисквания, увеличаващия се анализ на данни и популярността на облачните изчисления, е необходимо да се актуализират принципите за проектиране на ОЦИС.

Поради това се предлагат и актуални принципи за проектиране на ОЦИС, с което се отговоря на нуждите за сигурност и се извършва адаптация към променящите се технологии и бизнес. Тези принципи са:

1. **Максимално автоматизираната реакция и анализ на събития, включително използването на ИИ** – отнася се до автоматизирането на процесите за реагиране при инциденти с помощта на технологии за изкуствен интелект и машинно обучение. Това включва автоматичен анализ на събития и регистрационни файлове, генерирани от различни системи и устройства, което може да помогне за идентифициране на потенциални инциденти и аномалии в сигурността;

2. **Интеграция** – ОЦИС трябва да бъде интегриран с други системи за сигурност, като например защитни стени, системи за откриване на проникване и антивирусен софтуер, за да се осигури цялостна позиция за сигурност. Той трябва да бъде проектиран така, че да гарантира, че данните се споделят между системите, за да се подобрят възможностите за откриване на заплахи и реагиране;

3. **Използването на ИИ за анализ на структурирани и неструктурирани данни от събития -** включва използване на алгоритми за машинно обучение за анализиране както на структурирани, така и на неструктурирани данни от събития. Това може да помогне за идентифициране на модели и тенденции, които могат да бъдат пропуснати от традиционните методи и може да осигури по-точна и навременна представа за потенциални заплахи за сигурността;

4. **Анализът на съдържанието на неструктурирани данни** - обработка в почти реално време, както и на исторически данни от неструктурирани източници като имейли, социални медии и документи, за генериране на знания за анализ и за уведомяване и създаване на отчети;

5. **Приоритизиране** на предупрежденията за сигурност въз основа на тяхната тежест и въздействие. Това гарантира, че критичните сигнали получават незабавно внимание и че ресурсите се разпределят по съответния начин;

6. **Сътрудничество и споделяне на информация** включва възможности за комуникация между ОЦИС и други заинтересовани страни, като ИТ екипи, бизнес звена и външни партньори.

Прилагането на тези принципи за проектиране на ОЦИС би допринесло за това организациите да останат конкурентоспособни чрез подобряване на способността им да откриват и реагират бързо и ефективно на инциденти със сигурността.

Проектирането на ОЦИС за системи с големи данни може да бъде сложен и предизвикателен процес, но е от съществено значение да се гарантира, че чувствителните данни са защитени и инцидентите със сигурността се откриват и се реагира навреме. Поради тази причина е важно да се разгледат различните контролни на стандарта ISO 27001 [49], които имат отношение към създаването на метод, с който да се дефинират различните функционални нива на новото поколение на ОЦИС.

Използване на контроли за сигурност при проектирането на оперативен център за информационна сигурност

ISO 27001 може да се използва като референтна рамка за изграждане и управление на ОЦИС, тъй като предоставя изчерпателен набор от контроли, които организациите могат да прилагат, за да защитят своите информационни активи [50]. Този стандарт включва общо 114 контрола за сигурност в 14 области включително политики за информационна сигурност, организация на информационната сигурност, управление на активи, контрол на достъпа, криптография, физическа сигурност и сигурност на околната среда и управление на инциденти, които са предназначени да помогнат на организациите да управляват и смекчат рисковете за тяхната информационна сигурност. Стандартът е гъвкав и може да се адаптира, което означава, че не всички контроли могат да бъдат прилагани, а само тези, които отговарят на целите и оценените рискове [49], [50].

Следват някои от контролите на ISO 27001, които са подходящи, според нас, за внедряване в ОЦИС:

- **Политики за информационна сигурност** (A.5) - изисква организациите да установят и поддържат политики за управление на информационната сигурност. Екипите на ОЦИС трябва да се ръководят от политики, които очертават техните отговорности и процедури за справяне с инциденти, свързани със сигурността [49];

- **Физическа сигурност** (A.7) - изисква от организациите да прилагат мерки за физическа сигурност, за да защитят своите информационни активи, съоръжения за обработка на информация и други критични ресурси от физически заплахи като кражба, повреда или неоторизиран достъп [50];

- **Класификация на информацията** (A.8) - изисква от организациите да класифицират информационните активи въз основа на тяхното ниво на чувствителност и съответно да прилагат подходящи контроли. Екипите на ОЦИС могат да използват класификацията на информацията, за да съсредоточат усилията си върху най-критичните активи и да гарантират, че контролът за сигурност е съизмерим с нивото на риск [50];

- **Контрол на достъпа** (A.9) - изисква организациите да ограничат достъпа до информацията и съоръженията за обработка на информация до оторизирани потребители. Екипите на ОЦИС трябва да имат достъп до чувствителни данни и системи, но този достъп трябва да бъде внимателно контролиран, за да се предотврати неоторизиран достъп или пробиви на данни [51];

- **Управление на инциденти** (A.16) - изисква организациите да установят и поддържат процес за откриване, докладване и реагиране на инциденти, свързани със сигурността на информацията. Екипите на центровете за сигурност са отговорни за наблюдението на събитията по сигурността и реагирането на инциденти, така че този контрол е особено важен за операциите на ОЦИС [52].

Въз основа на съвременния характер на представените контроли от стандарта ISO 27001 от 2022 г. и тяхната значимост и уместност за областта на информационната сигурност, те ще бъдат приложими към проектирането на ОЦИС. В този контекст ние предлагаме интегрирането им в методологична рамка за проектиране на нов тип ОЦИС. Това би допринесло за гарантиране на поверителността, целостта и наличността на данните, обработвани и съхранявани в центъра по сигурност, както и за защитата на цялостната инфраструктура на ОЦИС срещу потенциални заплахи за сигурността.

Създаване на метод за проектиране на функционална архитектура на оперативен център за сигурност

След направен преглед и анализ на поколенията ОЦИС, предизвикателствата пред тях и текущото развитие на следващото поколение ОЦИС, считаме, че към момента липсва интегриран подход, спомагащ на всички компоненти на системата за сигурност за взаимосвързана работа, за осигуряване на цялостно решение за сигурност в реално време, използвайки данни от хетерогенни източници. Друга причина е бизнес необходимостта от подобрени мерки за сигурност в лицето на нарастващите заплахи за киберсигурността. В следствие на това в текущата дисертация предлагаме метод, който включва гореспоменатите контроли и изложените принципи за създаване на функционална архитектура, която ще позволи изграждането на актуално поколение ОЦИС (Фигура 10).

Фигура 10. Функционална архитектура на оперативен център за сигурност.

Основната цел на предложеното решение за ОЦИС е да отговори на развиващите се нужди на бизнеса, изправен пред повишени нива на кибер заплахи от различни категории недоброжелатели. Този център за сигурност включва адаптивна сигурност и предлага динамично коригиране на мерки за уязвимост, използвайки ИИ при обработка на настъпили събития, засягащи нивото на сигурност въз основа на съществуващи заплахи. Предложеният метод има за цел да осигури защита срещу кибератаки, като предлага още по-високо ниво на автоматизация, използвайки ИИ за събиране на различни видове данни от различни източници.

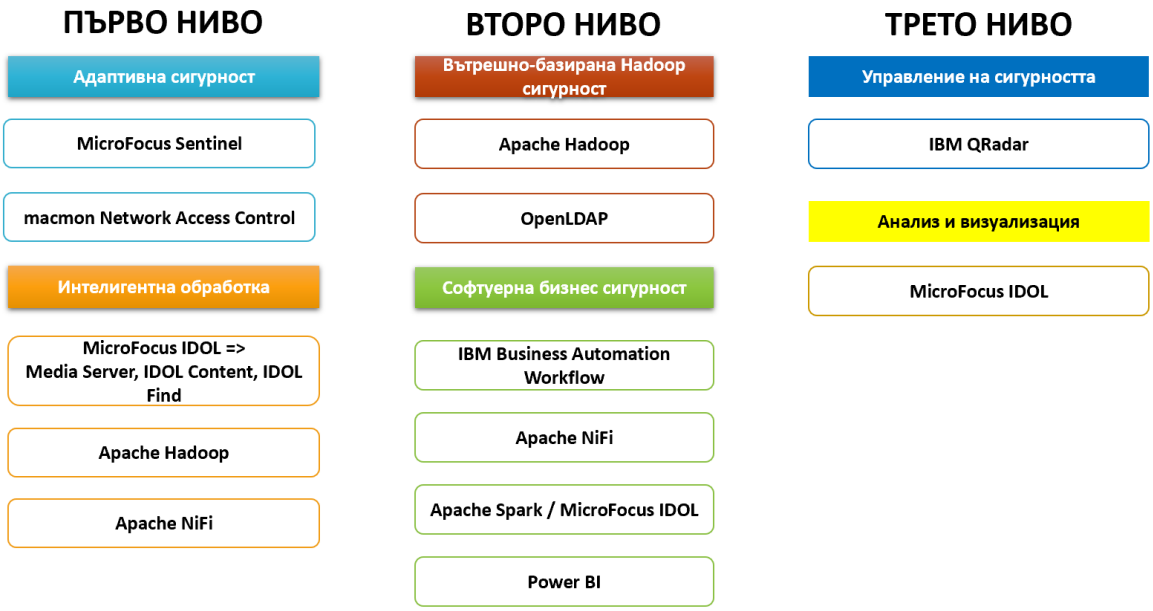
Предложената функционална архитектура се състои от три нива, като първото ниво е интелигентно управление на сигурността. Целта на това ниво е да открива и реагира на инциденти със сигурността в реално време, като използва ИИ и адаптивна сигурност. Системата събира данни от различни разнородни източници, включително видео потоци, данни от социални медии и други източници. ИИ, машинни и дълбоки методи за обучение се използват за когнитивно търсене, откриване на модели и знания и анализ на данни. Данните, събрани от социални медии и уебсайтове, могат да се използват за идентифициране на лица или групи, които могат да представляват заплаха за сигурността, както и за идентифициране на откраднати превозни средства. Освен това технологията за лицево разпознаване може да помогне за идентифицирането на лица, които могат да представляват заплаха за сигурността.

Второто ниво е базирана на Hadoop вътрешна сигурност и софтуерна бизнес сигурност, предназначени да защитават поверителността, целостта и достъпността на данните. Това ниво включва подходи, използващи централизирано удостоверяване на потребителите, създаване на потребителски права за достъп до клъстери на големи системи за данни и сегментиране на данни.

Най-високото трето ниво е анализ и визуализация на получените резултати, което има за цел да направи подходяща визуализация на събраните данни от предходните нива и да открие потенциални заплахи в реално време. Системата следи събития, случващи се в центъра, като например създаване и анализ на регистрационни файлове, за предотвратяване и откриване на потенциални заплахи в реално време.

Предложената функционална архитектура предоставя изчерпателен набор от контроли, които организациите могат да прилагат, за да защитят своите информационни активи. Архитектурата е гъвкава и може да се адаптира, за да отговори на специфичните нужди на всяка организация, независимо от нейния размер или индустрия.

На Фигура 11 са представени обобщено възможностите за използвани технологии на всяко от нивата, които реализират поставените цели при всяко от тях.



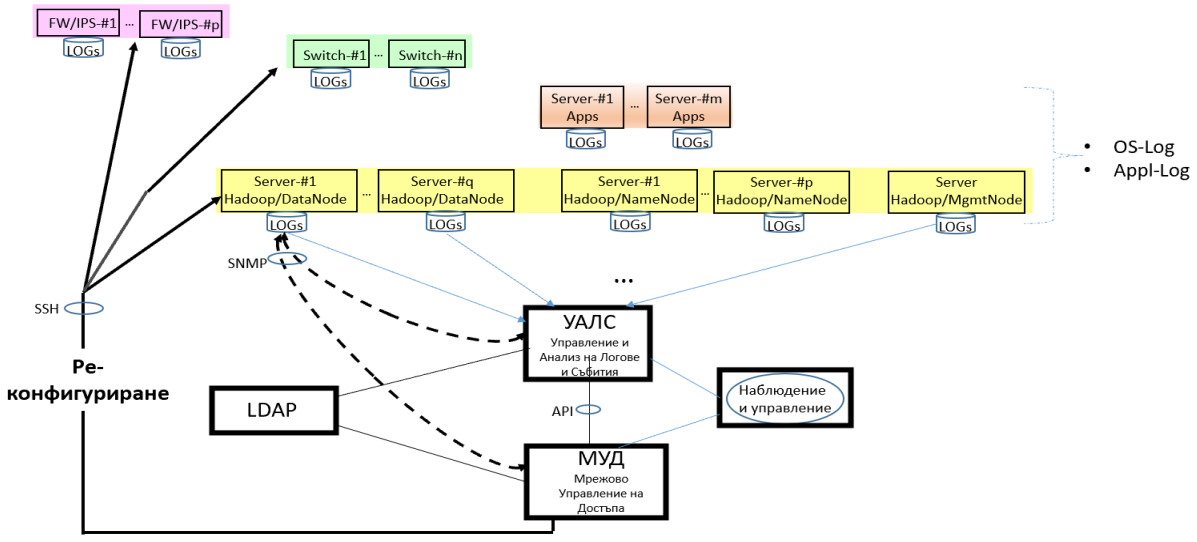
Фигура 11. Използвани технологии при различните нива на ОЦИС.

Следва подробно представяне на нивата в предложената архитектура и на технологични решения, обхващащи функционалности, предлагащи по-високо ниво на защита за всяко едно от тях.

Първо ниво на оперативен център за информационна сигурност – Адаптивна сигурност

Основно предизвикателство пред сигурността е постоянната заплаха от атака. Адаптивната сигурност (АС) е подход за киберсигурност, който постоянно анализира поведението и събитията в мрежата и има готовност да се адаптира към заплахи като ги проучва и анализира преди да се случат. Една организация може непрекъснато да оценява риска и да осигурява подходящо прилагане на защитни механизми и подходи, използвайки адаптивна сигурност [53].

Предлагаме компонентна архитектурата за адаптивна сигурност за система с големи данни, която трябва да бъде проектирана така, че да защитава поверителността, целостта и наличността на данните. На Фигура 12 е предложена архитектура за АС, която да бъде приложена за система с големи данни.



Фигура 12. Компоненти на адаптивна сигурност за системи с големи данни.

Предложената архитектура включва множество компоненти като защитни стени, превключватели, сървъри за приложения, сървъри, съдържащи компоненти на архитектурата на Hadoop Distributed File System (HDFS), и връзки на устройства. Средата

Hadoop има три основни компонента, а именно nameNode, dataNode и managementNode, които са отговорни съответно за управлението на пространството от имена на файловата система, съхраняването и управлението на блокове от данни и хостинг услугите за управление. Всички компоненти в системата имат активирана опция за създаване на лог файл, които се наблюдават от адаптивни системи за сигурност, за да осигурят мрежово състояние в реално време и да намалят потенциалните рискове за сигурността. Репликацията на данни се използва за осигуряване на наличност на данни в случай на компрометиране на възел. Системата също така включва система за управление и анализ на регистрационни файлове и събития (УАЛС) и система за управление на достъпа до мрежата (МУД) за управление и анализиране на регистрационни файлове и събития и осигуряване на достъпа до мрежата съответно. Като цяло тази архитектура осигурява цялостен и ефективен начин за защита на системи за големи данни [54]. Значението на защитения достъп и поверителността на данните в системите за големи данни е от първостепенно значение и може да се постигне чрез методи като многофакторно удостоверяване и криптиране на данни с помощта на техники като AES и 3DES. Редовното наблюдение и одит на системата помага за откриване на неразрешена дейност и гарантира спазването на разпоредбите. Използването на архитектура на микроуслуги и внедряването на нови функции за сигурност за контейнеризирани приложения също може да подобри сигурността. Мерките за мрежова сигурност като защитни стени и сегментиране на мрежата могат да помогнат за предотвратяване на атаки. УАЛС и МУД са компоненти на адаптивната сигурност със специфични задачи за изпълнение и технологични решения за постигане на целите за сигурност [55], [56].

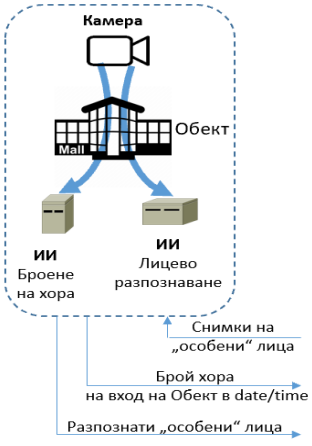
Предложеното адаптивно решение за сигурност за системата ОЦИС включва използването на системите Micro Focus Sentinel за УАЛС и за МУД - macmon Network Access Control. Sentinel събира и анализира свързани със сигурността данни от различни източници, докато macmon осигурява достъп до мрежата и контролира устройства и потребители. Интегрирането на тези две системи позволява по-голяма видимост в мрежата и бърза реакция при инциденти със сигурността. Споделянето на данни чрез API позволява корелация на събития и предупреждения за сигурност, улеснявайки идентифицирането на потенциални заплахи и незабавни действия за смекчаването им. Този подход позволява на

екипа на ОЦИС да бъде информиран за текущото състояние на системата и да предприеме навременни мерки за гарантиране на сигурността [57], [58].

Първо ниво на оперативен център за информационна сигурност – Интелигентна обработка на данни

Интелигентната обработка на данни се отнася до система или подход за управление на сигурността, който използва интелигентни или автоматизирани технологии, в които ИИ е с основна роля. Това ниво на сигурност е силно фокусирано върху събирането и анализирането на данни от различни източници по сигурен начин [59].

Автоматизирането на анализа на данни чрез алгоритми за машинно обучение и други ИИ техники улеснява откриването и реагирането на потенциални заплахи чрез идентифициране на модели в големи набори от данни. ИИ се използва за наблюдение на социални медии и други уебсайтове за подозрителна дейност и откриване на необичайно поведение от камери, разположени вътре или извън съоръжение. Тази технология може да идентифицира потенциални заплахи като разпространение на невярна информация или планиране на престъпни дейности, неоторизиран достъп или подозрителна дейност около зони с ограничен достъп (Фигура 13).

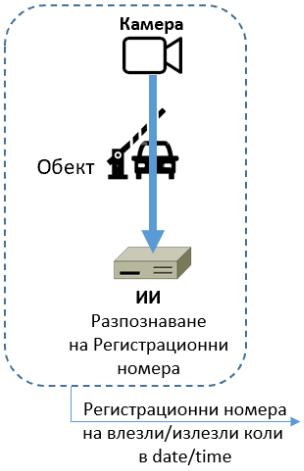


Фигура 13. Интелигентно първично ниво на сигурност с камери.

Фигура 14 илюстрира използването на видео потоци за откриване на необичайно поведение, като неоторизиран достъп или подозрителна дейност около зони с ограничен достъп. Чрез използване на усъвършенствани ИИ техники, алгоритмите могат да бъдат

обучени да броят хората, влизащи и напускащи дадено местоположение в реално време, предоставяйки на екипите по сигурността ценна информация за човешкия поток и нивата на заетост. Техниките за откриване и проследяване на обекти, включително дълбоки невронни мрежи, могат да се използват, за да се избегне двойното броене и изкуственото увеличаване на броя на обектите в рамката. ИИ може също да бъде обучен да разпознава конкретни лица въз основа на снимки на „посочени“ лица, което позволява бързо идентифициране и реакция на потенциални заплахи. Тази технология може да бъде особено полезна в среди с висока степен на сигурност като летища, железопътни гари и стадиони.

Разпознаването на регистрационни номера на автомобили от изображения или видео потоци включва използването на алгоритми, които използват оптично разпознаване на символи (OCR) за идентифициране и четене на номерата на регистрационните номера – Фигура 14.



Фигура 14. Регистриране на автомобилни номера на вход на обект.

OCR технологията има различни приложения като контрол на трафика, управление на паркиране и правоприлагане.

От друга страна, когнитивният анализ може да се използва за анализиране на неструктурирани данни като текст, изображения и видеоклипове от уебсайтове и блогове [25]. Това позволява на системата да разбере контекста на съдържанието на уебсайта, да открие негативни настроения или тон и да идентифицира визуални елементи, които допринасят за потребителското изживяване. Инструменти за уеб сканиране или обхождане

като BeautifulSoup, Scrapy и Selenium могат да се използват за извличане на съдържание на уебсайт, което след това се индексира в система за големи данни. Друг подход за изтегляне на уеб съдържание е чрез използване на система за управление на съдържанието (Content Management System - CMS), която отделя създаването и управлението на уеб съдържание от презентационния слой [106]. CMS може да предостави API за Hadoop за достъп до съдържание, съхранено в CMS, или данните могат да бъдат прехвърлени от CMS към HDFS с помощта на персонализиран скрипт или NiFi. Анализирането на данни за уебсайтове с AI позволява на бизнеса да получи ценна информация за поведението на потребителите, пазарните тенденции и други ключови бизнес показатели. Задвижваните от AI търсачки могат да разпознават и анализират голямо разнообразие от типове данни, включително структурирани, неструктурирани и полу-структурирани данни, осигурявайки по-цялостно разбиране на данните [60].

Второ ниво на оперативен център за информационна сигурност - Вътрешно-базирана сигурност за системи за големи данни

Вътрешно базираната сигурност за системи за големи данни е подход за сигурност на данните, който набляга на включването на мерки за сигурност директно в инфраструктурата за големи данни. Този подход има за цел да затрудни злонамерените участници да получат неоторизиран достъп или да компрометират системата. Традиционно сигурността на данните се основава на мерки за сигурност, базирани на периметъра, като защитни стени и системи за откриване на проникване. Въпреки това, вътрешната сигурност се фокусира върху интегрирането на функции за сигурност в рамките на инфраструктурата за големи данни.

Предложеният метод за вътрешна базирана сигурност за системи с големи данни използва системата за големи данни Hadoop, която включва няколко вградени функции за сигурност. Тези функции са предназначени да адресират ключови области на проблеми със сигурността, като удостоверяване и оторизация, криптиране и одит.

Централизираното удостоверяване на потребителя е основен подход за сигурност за Hadoop клъстери. Това може да бъде постигнато чрез механизми LDAP и/или Kerberos, за да се осигури правилно удостоверяване и контрол на разрешения за множество

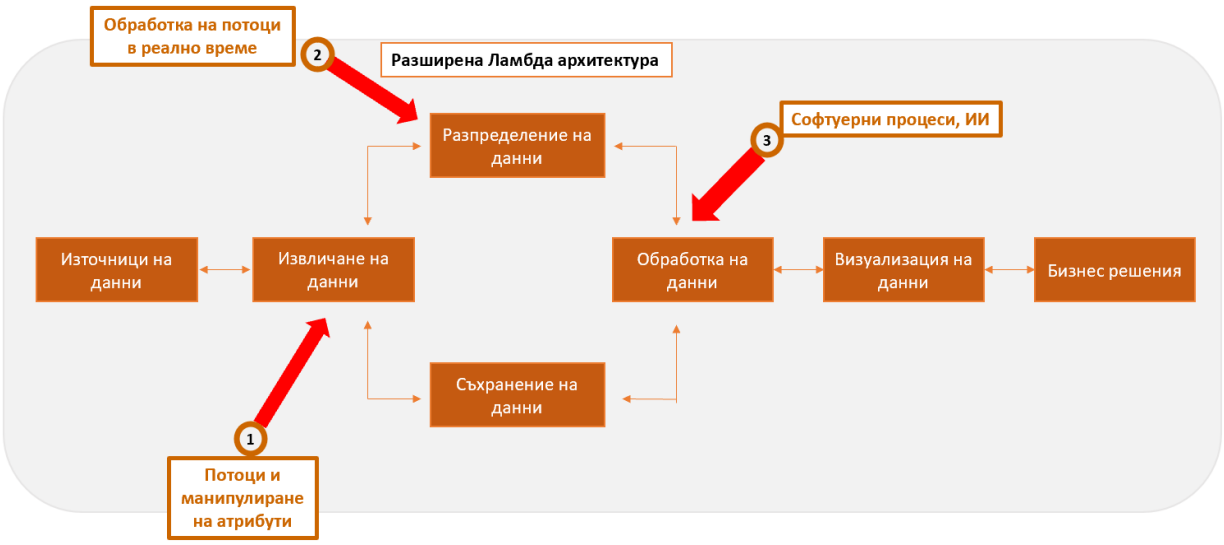
потребители, които имат достъп до клъстера [61]. Еднократните пароли предоставят допълнителен слой на сигурност, като изискват уникална парола, която може да се използва само веднъж и е валидна за ограничен период от време. MFA може да се приложи с LDAP, за да изисква два или повече фактора за удостоверяване за проверка на самоличността на потребителя. За контрол върху достъпа до сървърите на DataNode въз основа на потребителски акаунти и групи, може да се внедри System Security Services Daemon – SSSD. Последното е доставчик на идентичност и удостоверяване, който позволява интегрирането на различни механизми за удостоверяване и поддържа сегментиране на данни за контрол на достъпа до DataNode сървъри [62], [63]. Контролът на достъпа до HDFS директории и файлове може да се управлява чрез присвояване на разрешения на потребители и групи чрез модела на разрешения в стил Unix и командата "setfacl" от Hadoop CLI [64], [21].

Друг подход за осигуряване на сигурност е чрез създаване на централизиран лист за контрол на достъпа на Hadoop – Access Control List (ACL), който се управлява от интерфейса на командния ред на HDFS (Command-Line Interface - CLI) или инструмента за графичен потребителски интерфейс (Graphic User Interface - GUI) [65]. Този механизъм позволява на администратора да добавя или премахва потребители или групи от ACL и да променя правилата за достъп за конкретни файлове или директории. В допълнение, създаването на цялостен механизъм за одит е от решаващо значение за осигуряване на проследимост на достъпа до данни и активността на потребителите, което помага за предотвратяване на загуба на данни и поддържане на нормативно съответствие [66]. Проследяването на произхода на данните включва запис на метаданни за произхода, историята и трансформацията на данни в инфраструктурата, докато проследяването на потребителската активност включва запис на метаданни за действията, предприети от потребителите в рамките на инфраструктурата. И накрая, защитата и криптирането на данни в покой и в движение са критични компоненти на защитен и добре управляван Hadoop клъстер, което може да бъде постигнато чрез различни методи за криптиране и TLS протокол [67]. Тези подходи помагат да се поддържа сигурността на клъстера Hadoop и защитават ценните данни, на които разчитат фирмите [68].

Второ ниво на оперативен център за информационна сигурност – Софтуерна бизнес сигурност

Сигурността на бизнес софтуера (СБС) е от съществено значение за защита на чувствителните данни, събрани от бизнеса, от киберзаплахи като злонамерен софтуер, хакерство и фишинг. СБС включва контрол на достъпа, криптиране на данни, сигурни програмни практики, оценки на уязвимостите, тестове за проникване в системата, мониторинг на сигурността и реакция при инциденти. Статистическите методи могат да се комбинират с други мерки за сигурност за анализиране на данни и откриване на аномалии, тенденции и модели, които могат да показват заплахи за сигурността. Корелационният анализ се използва за идентифициране на потенциални заплахи за сигурността и предприемане на подходящи мерки за смекчаването им [69]. Освен това управлението на бизнес процеси (Business Process Management - BPM) се използва за идентифициране на потенциални рискове за сигурността и уязвимости в бизнес процесите и за проектиране и прилагане на мерки за сигурност за смекчаване на тези рискове [70]. Използвайки тези методи, организациите могат да разработят цялостна стратегия за сигурност, която адресира широк спектър от потенциални заплахи и уязвимости.

Интегрирането на СБС с Ламбда архитектурата гарантира, че данните, които се обработват, са защитени. Ламбда архитектурата е подход за обработка на данни, предназначен за обработка на огромни количества данни чрез комбиниране на групова обработка и методи за обработка на данни в реално време. Въпреки това, използването на Ламбда архитектурата може да доведе до несъответствия в резултатите, генерирани от пакетния слой и слоя за обработка в реално време. За да се преодолеят тези ограничения, се използва разширена Ламбда архитектура, която въвежда допълнителни слоеве като слой за събиране на данни, слой за съхранение на данни и слой за обработка на данни [71], [72] - (Фигура 15). Тези слоеве помагат за увеличаване на производителността, скалируемостта и последователността на системата.



Фигура 15. Разширена Ламбда архитектура.

Разширената Ламбда архитектура интегрира сигурността на софтуерния бизнес в своя работен процес, което включва дефиниране на източници на данни и извличане на данни от тях с помощта на слой за събиране на данни. След това данните се съхраняват директно в разпределена система за съхранение или се разпространяват за обработка. След подготовката на данните се извършва пакетна, поточна или обработка с ИИ алгоритъм, последвана от анализ на данните с помощта на инструменти за анализ на големи данни, за да се предоставят информирани знания на бизнеса. Визуализацията на резултатите се извършва с помощта на платформи за визуализация на големи данни или софтуер за системи за планиране на ресурсите на предприятието (Enterprise Resource Planning - ERP). Инструментите за управление на бизнес процеси (BPM), като IBM Business Automation Workflow (BAW), се използват за управление и оптимизиране на работните процеси за обработка на данни. Различни приложения като NiFi, IDOL и Spark се използват за поддръжка на обработка на данни, анализ и визуализация на различни етапи. NiFi се използва за поглъщане на данни от различни източници, Spark се използва за обработка на потоци от данни в почти реално време, а MLlib се използва за обучение на модели за машинно обучение и извършване на прогнози и класификация в реално време. IDOL се използва за разширени функции за извличане на информация от неструктурирани данни.

Визуализирането на данни е важен аспект от анализа на данни, позволяващ на вземащите решения да идентифицират тенденции, модели и аномалии и да предскажат бъдещи тенденции. Инструменти като PowerBI, Tableau и Plotly предлагат адаптивни и интерактивни визуализации, които могат да бъдат адаптирани към конкретни бизнес нужди. Освен това IDOL има вградени функции за обработка на данни и извличане на знания, включително възможност за създаване на графики и диаграми. Използвайки разширената Ламбда архитектура, компаниите могат да се възползват от цялостен подход към обработката на данни и сигурността, който позволява наблюдение в реално време, откриване и отговор на заплахи за сигурността, като същевременно гарантира последователност и точност на данните. Архитектурата може също да интегрира статистически методи и изкуствен интелект за подобряване на възможностите за анализ на данни и поддържане на сигурността на данните.

Трето ниво на оперативен център за информационна сигурност

Третото ниво на ОЦИС включва управление на сигурността и визуализация на данни от различни източници като регистрационни файлове, видео потоци и социални медии. Ефективният анализ и визуализация на тези данни може да помогне за прилагането на цялостни мерки за сигурност и стратегии за предотвратяване на уязвимости. SIEM системи като QRadar и MicroFocus IDOL могат да бъдат интегрирани в ОЦИС, за да осигурят мониторинг в реално време, анализи и възможности за реакция при инциденти. QRadar използва корелационен механизъм за откриване на модели и аномалии в данните за сигурност, докато IDOL се използва за анализиране и визуализиране на неструктурирани данни от различни източници. Чрез комбиниране на QRadar с IDOL, фирмите могат да получат представа за състоянието на сигурността си и да идентифицират потенциални заплахи, които са били пропуснати от други инструменти за сигурност. Тази интеграция предоставя разширени възможности за визуализация и позволява на бизнеса да изследва данните за сигурността по нови начини [73].

Изводи

В тази глава се подчерта значението на използването на центрове за бизнес сигурност, като се представиха четирите поколения ОЦИС и предизвикателствата, пред които са изправени

преди появата на следващото поколение. Изведохме актуални принципи за проектиране на ОЦИС и се фокусирахме върху контролите за управление на сигурността, предоставени от международния стандарт ISO 27001, които са подходящи за този тип центрове за сигурност. Въз основа на реализирания анализ предложихме метод за проектиране на функционална архитектура за нов тип оперативен център за сигурност на системата за големи данни. Предложената архитектура включва три нива на управление на сигурността, включително адаптивна сигурност, използваща ИИ, интелигентна обработка на данни, вътрешна сигурност и цялостен системен анализ. На всяко ниво се предлагат технологични решения, които да бъдат внедрени в работещ модел на предложената ОЦИС. Функционалната архитектура предлага многопластов подход към управлението на сигурността, осигурявайки цялостна защита за системи с големи данни чрез интегриране на технологии за постигане на поставените цели.

4. Прилагане на метода за оперативен център за сигурност с архитектурни решения

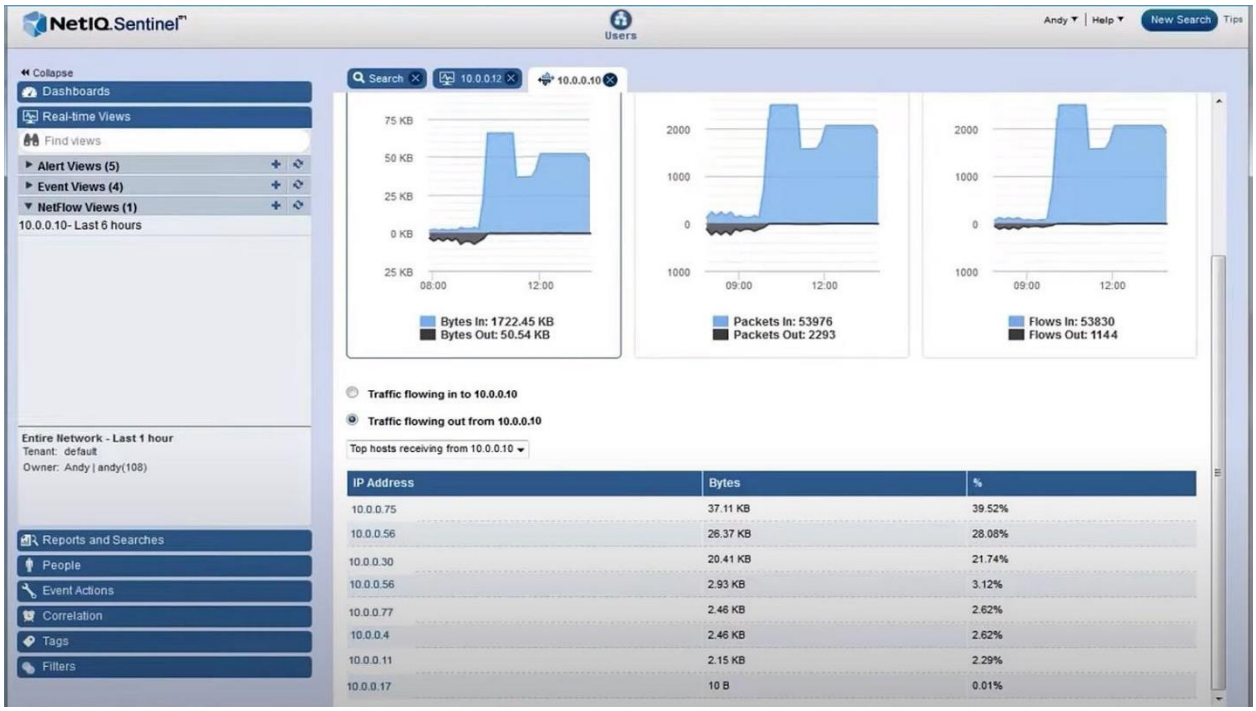
Предложената функционална архитектура предлага многопластов подход към управлението на сигурността, осигурявайки цялостна защита за системи с големи данни, интегрирайки технологии за постигане поставяне на цели. Чрез прилагане на предложения метод за проектиране ОЦИС с представените решения, организациите могат да подобрят своята информационна сигурност. Някои от технологичните решения, представени в глава трета, вече са внедрени от различни компании и поради това не се обсъждат в тази глава. На част от тях ще се обърне по-голямо внимание поради специфики в прилагането им, което демонстрира, че същите са реализуеми за предложения оперативен център за информационна сигурност.

Прилагане на адаптивна сигурност

Адаптивният подход за сигурност е от съществено значение при управлението на най-ниското ниво на предложения център за операции по сигурността, което е нивото на мрежата. Две допълнителни функции за сигурност, включени в предложената архитектура, са управление и анализ на логове и събития (УАЛС) и управление на достъпа до мрежата

(МУД), които осигуряват детайлен контрол върху достъпа до мрежата и постоянно наблюдение на логове от свързаните устройства. УАЛС работи като събира регистрационни файлове от всички компоненти на ниско ниво, а МУД управлява достъпа до тях. Micro Focus Sentinel и macmon Network Access Control са две решения, които могат да се използват за наблюдение на мрежовия трафик и предоставят на администраторите навременна информация за състоянието на системата. Активирането на тези решения за сигурност може да подобри сигурността на организацията чрез изграждане на цялостен и многопластов подход към сигурността.

На Фигура 16 е представено таблото за постоянен мониторинг от Sentinel, в който се виждат мрежовата натовареност на наблюдаваните устройства в предложения ОЦИС.



Фигура 16. Табло за постоянен мониторинг на мрежовия трафик в ОЦИС.

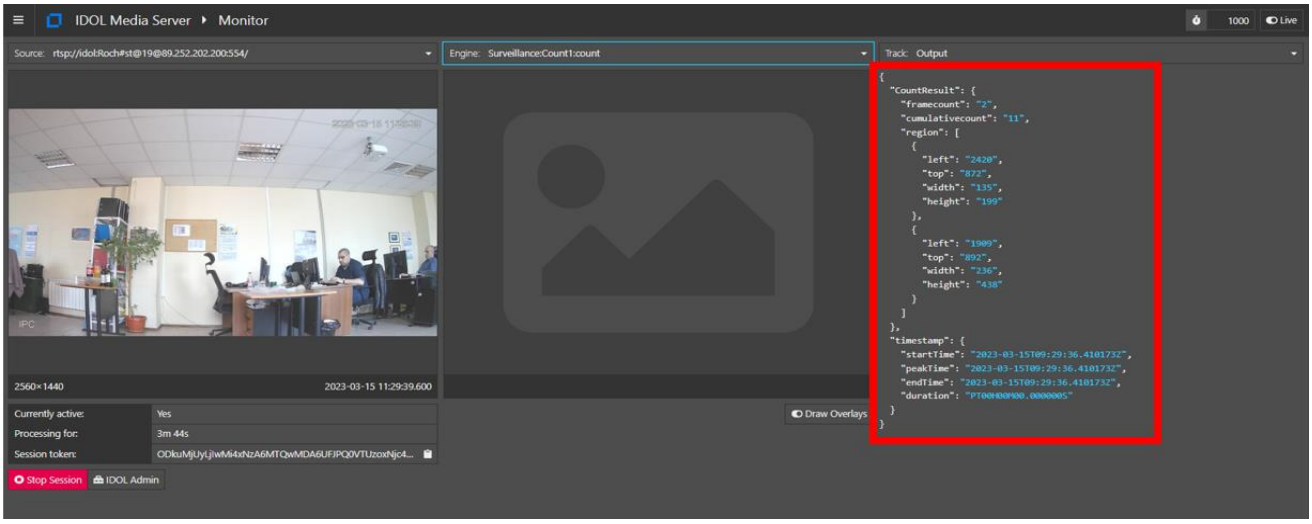
Използвайки двете системи, администраторите на ОЦИС могат да взимат навременни мерки и да бъдат информирани за текущото състояние на системата.

Прилагане на решения за интелигентна обработка на данни

След като разгледахме предложение за защита на мрежата и компонентите в средата за големи данни, следва начина на защита при извличането и обработката им. За да се реализира тази функционалност, се използва системата на Micro Focus – IDOL (описана подробно във Втора глава).

Поради необходимостта от извличане на данни от хетерогенни източници, са разгледани методи на извличане, съхранение и обработка на данни от видео потоци, включващи действия като преброяване на хора, разпознаване на лица и регистрационни номера на автомобили, както и извличане на данни от уеб-съдържание и социални медии.

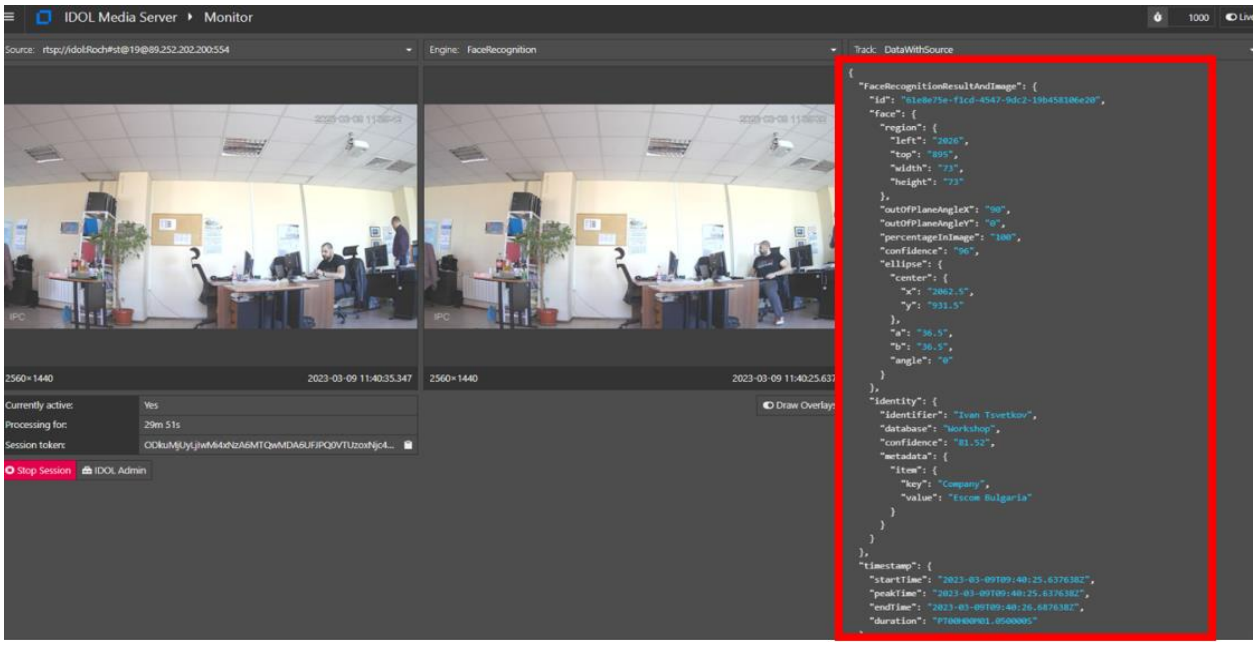
Предложената архитектура за извличане на данни от видео поток включва запис или заснемане на видеото, последвано от обработка на данни с помощта на ИИ алгоритми на IDOL Media Server. Последното анализира видео потока, за да открива лица и обекти, и има възможности за преобразуване на реч в текст за проследяване на аудио емисии. Алгоритмите на ИИ в Media Server могат също да броят хора от видео поток, постигайки една от целите на предложената архитектура. Изходът за броя на хората съдържа данни като брой хора на кадър и кумулативен брой, а също така определя продължителността, началния час и крайния час на видеопотока. Фигура 17 показва визуализация в реално време на видео потока и изхода, генериран от Media Server.



Фигура 17. Преброяване на хора в кадър от видео поток.

Media Server разделя видеокадъра на различни области, по-добро откриване на обекти и ефективност на проследяване. Използването на области в кадър допринася за съсредоточаване върху отделни ;части от кадъра, където има вероятност да се появят важни събития или дейности, и да игнорира други области, където се очаква малко или никаква активност. В случая на Фигура 17, има две области на разделение – там, където има струпване на хора и там, където няма.

IDOL Media Server е технология, която може да анализира видео потоци за откриване на лица и обекти с помощта на ИИ алгоритми. Една от целите на предложената архитектура е да преброи хората както вътре, така и извън съоръжението. Media Server е обучен да брои хора от видео поток с помощта на ИИ алгоритми. Функцията за лицево разпознаване включва създаване на математическо представяне на лицето на човек, което улавя ключови черти на лицето. Media Server сравнява извлечените характеристики от видеопотока с база данни с характеристики на известни или посочени лица, като използва алгоритъм за съпоставяне, за да присвои оценка за сходство – Фигура 18.

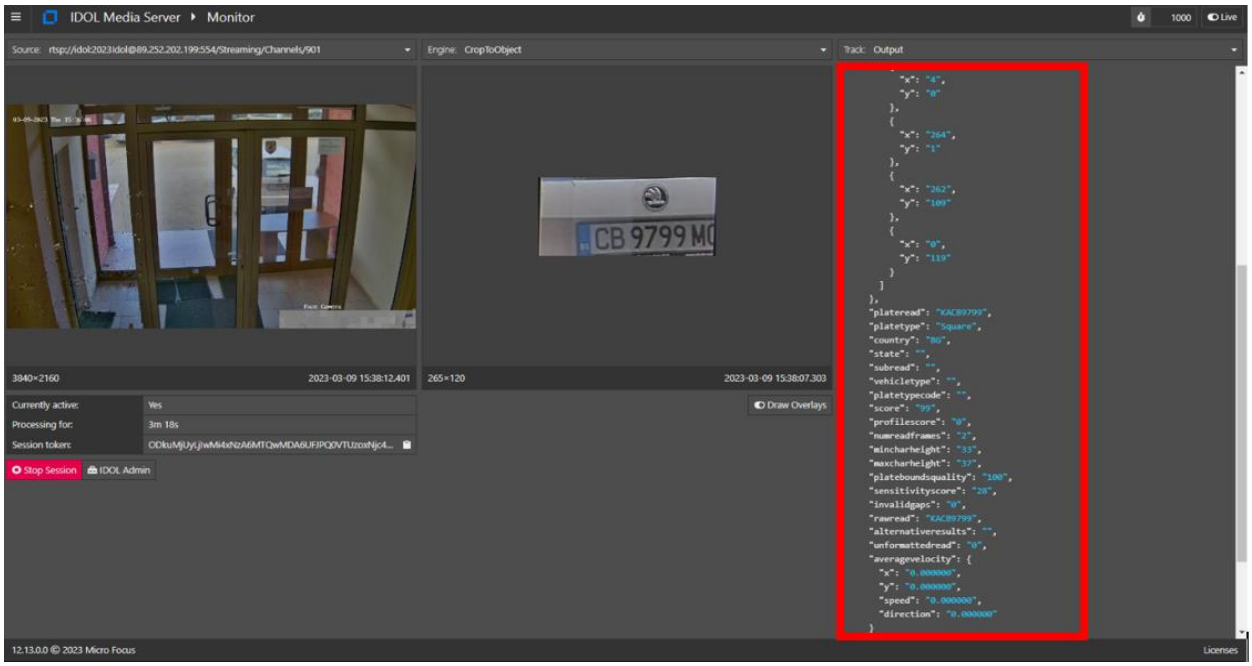


Фигура 18. Разпознаване на лица от Media Server.

Резултатът, генериран от софтуера, предоставя ценна информация като местоположението на откритите лица във видеокадрите, името на разпознатия индивид, базата, от която лицето е открито, и процент на увереност, че откритото лице е това, на

което съответства . Тази информация се използва, за да се определи дали резултатите от откриването са достатъчно надеждни, за да се предприемат по-нататъшни действия или да се предупредят служителите по сигурността.

IDOL Media Server анализира видеопоток, за да открие и локализира регистрационните номера и след това използва OCR алгоритми, за да разпознае знаците – Фигура 19.



Фигура 19. Разпознаване на автомобилни номера с Media Server.

Софтуерът сравнява регистрационния номер с база данни с известни номера, като използва алгоритъм за съвпадение. Генерираният изход включва информация за разпознатия регистрационен номер, неговите атрибути и нивата на достоверност на резултатите от разпознаването. Този изход включва също допълнителна информация за разпознатата регистрационна табела, като държава или държава на регистрация, марка и модел на превозното средство и дата и час на разпознаване. След като IDOL Media Server разпознае лице или номер на регистрационния номер, той може да анализира метаданните, свързани с този обект, като например времето и местоположението, когато е бил открит.

Тази информация може да се използва за сигурност, наблюдение, контрол на трафика или анализ на клиента.

В последните години данните, генерирани в социалните медии, се превръщат в ценен ресурс за бизнеса. В тях се идентифицира клиентско поведение и се извършва анализ, който да подобри качеството на бизнеса.

На това ниво използваме продуктите Micro Focus IDOL, Apache Hadoop и Apache NiFi, а избраните социални медии са Фейсбук (Facebook) и Туитър (Twitter).

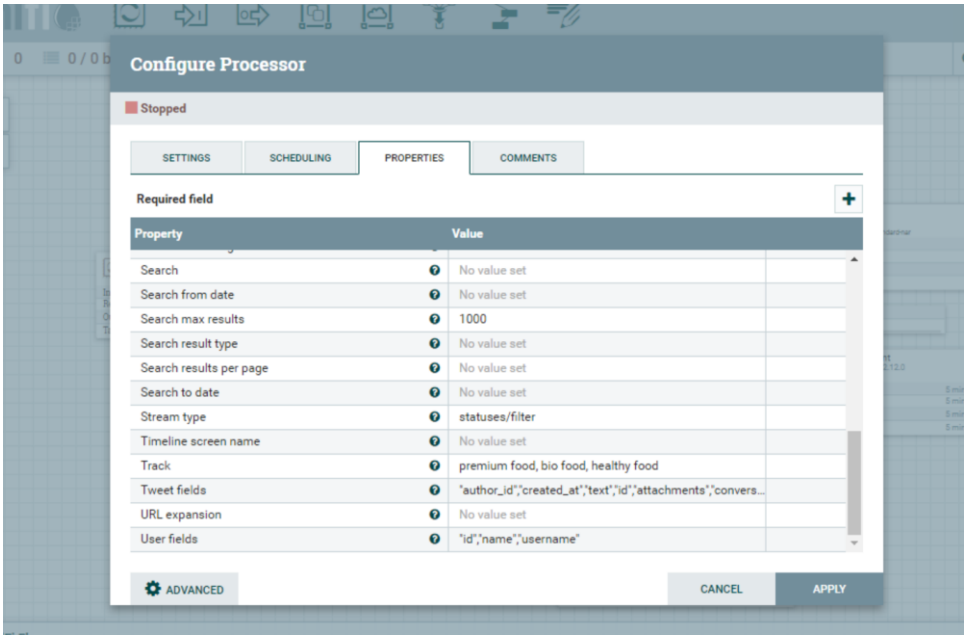
Процесът на извличане, съхранение и обработка на данни е показан на Фигура 20.

Фигура 20. Процес на обработка на неструктурирани данни от социални медии.

Процесът на обработка на неструктурирани данни от социални данни включва инсталиране и конфигуриране на IDOL сървъра и неговите компоненти, създаване на база данни за съхраняване на извлечените данни, дефиниране на роли за достъп и използване на вградени ИИ функционалности. След това Apache NiFi се инсталира и конфигурира за извличане и съхраняване на данни. Следващата инсталация е Hadoop, който е конфигуриране за съхранение на данни. Конекторите са инсталирани и конфигурирани за извличане на данни от социални медийни платформи като Twitter и Facebook чрез създаване на приложения в средата за разработчици на приложения и използване на клиентски ключове и токени за свързване с API. След това системата NiFi се използва за дефиниране на конектори за социални медии за извличане на желано съдържание. Фигура 21 демонстрира този процес.

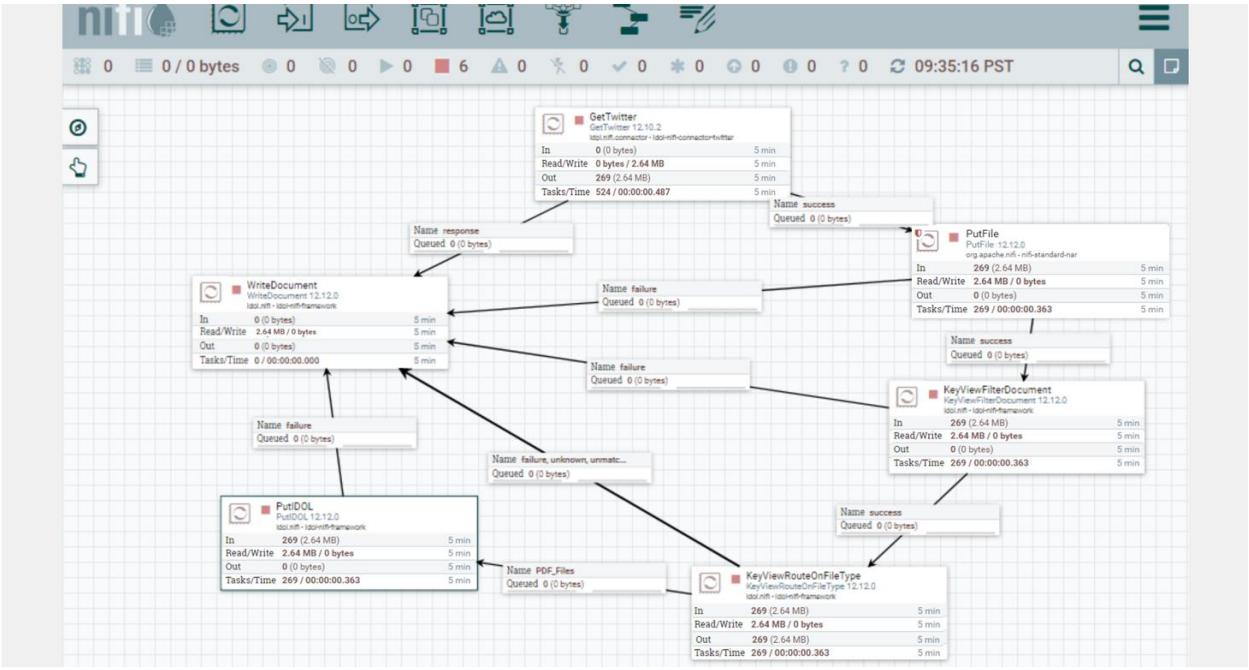
Фигура 21. Конфигурация на конектор с предоставени ключове в създадено приложение на социална медия.

Конекторите предлагат възможност за филтриране на данните по ключови думи или фрази, които се съдържат в социалните медии, както е показано на Фигура 22.



Фигура 22. Конфигуриране на полета за извличане на данни от социални медии.

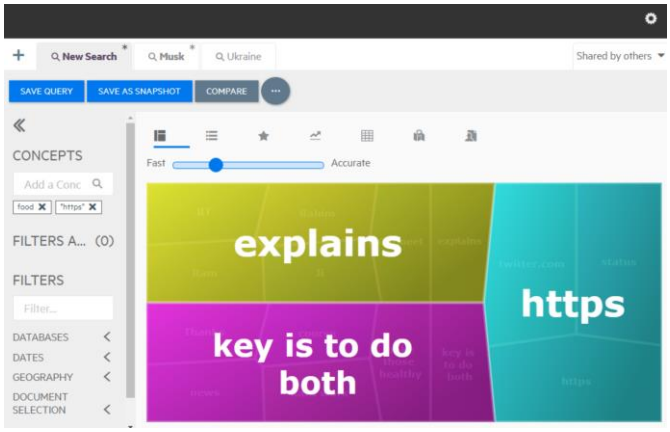
При изграждането на процеса за обработка на неструктурирани данни от социални медии в среда NiFi (Фигура 23), е необходимо създаването на процесори.



Фигура 23. Процес по извличане и съхранение на данни от Туитър.

Това става след като изберем опцията за избор на тип компонент на IDOL, към който ще бъде причислен. След това конфигурираме конектора и потвърждаваме, че връзката между компонентите функционира и данните могат да бъдат изтеглени.

След като процесът е завършен и данните се намират в IDOL, следват обработки и анализ на данните по допълнителни заявки за търсене, филтриране или сортиране от IDOL Find, който използва ИИ при тази обработка (Фигура 24).



Фигура 24. Зареждане на данни в IDOL Find.

IDOL Find използва техники за предварителна обработка, като токенизация и нормализиране за трансформиране на необработените данни във формат, който може да бъде анализиран. Това включва процеси по разделяне на данните на по-малки единици, като думи или фрази, и премахване на шум и неуместна информация.

Данните, които имат визуално съдържание, можем да визуализираме (Фигура 25) преди да бъдат обработени или съхранени в хранилището на Hadoop, откъдето могат по всяко време да бъдат извлечени и използвани за последващи анализи.

Фигура 25. Визуализация на данни в IDOL Find.

Много фирми разчитат на данни, за да вземат информирани решения. Изследователи в различни области събират данни от уебсайтове, за да провеждат проучвания, да анализират тенденции и да получат представа за конкретни теми.

Използваме сървъра IDOL и платформата IDOL Content, за да индексираме конкретно уеб съдържание. Процесът включва извличане на съдържанието с помощта на Scrapy и генериране на XML файл, който се проверява за съвместимост с IDOL. Файлът съдържа полета като DRREFERENCE, DRETITLE, DRECONTENT, DREDBNAME, DREINDEX, DREFILENAME и DREFIELD. След индексиране данните са готови за обработка и анализ в IDOL Find, където могат да се търсят. Чрез анализиране на данни за уебсайтове с AI, фирмите могат да получат ценна информация за поведението на потребителите и пазарните тенденции, които могат да информират при вземането на

решения. Разширената Ламбда архитектура, използваща NiFi и Spark, предлага сигурност от второ ниво в допълнение към вътрешната сигурност на Hadoop.

Прилагане на сигурност на второ ниво от функционалната архитектура

За да се гарантира сигурността на система с големи данни, е необходимо да се осигури функционалност за сигурност от самото начало – входът на системата. LDAP и Kerberos са две решения, които могат да постигнат това. В този контекст LDAP се разглежда като специфично приложение за Hadoop. Първата стъпка е инсталиране и конфигуриране на OpenLDAP LDAP сървъра. Това включва създаване на LDAP директория, дефиниране на потребители и групи и конфигуриране на правила за удостоверяване и оторизация. Системата OpenLDAP използва LDAP файл за обмен на данни (LDIF), който се създава с текстов редактор. Потребители и групи могат да се добавят към базата данни на OpenLDAP с помощта на командата ldapadd.

```
Dn: cn=ivelkova,ou=users,dc=example,dc=com
objectClass: top
objectClass: person
objectClass: organizationalPerson
objectClass: inetOrgPerson
cn: ivelkova
sn: Velkova
givenName: Ivona Velkova
userPassword: {SHA}nU4fzW93lj8q3I3TQDwBsjJREO
mail: ivelkova @example.com
```

Политиките за удостоверяване и оторизация могат да бъдат конфигурирани с помощта на OpenLDAP Access Control Language (ACL).

```
access to * by dn="cn=admin,dc=example,dc=com" write by * read
```

LDAP удостоверяването за Hadoop може да бъде активирано чрез редактиране на системния core-site.xml конфигурационен файл на Hadoop.

След подсигуряване на входа в системата, следва да се обърне внимание на софтуерната бизнес сигурност на това ниво. Софтуерната бизнес сигурност се отнася за защитата на бизнес активностите, инфраструктурата и данните. Инструментите за софтуерната бизнес сигурност са подходящи за цялостна обработка на данни от различни източници. Това е важно, тъй като такъв се смята за по-висока информираност на бизнеса относно потенциални заплахи в онлайн пространството като обработката им във време близо до реалното, би помогнало за съвременна реакция срещу тях. Тук използваме инструментите NiFi и IDOL, като последният се използва и за визуализация на резултата. Процесите, които са свързани с обработката на данни за извличане от източници, съхранение в хранилище, обработка за извличане на информация и анализ на информацията, които да помогнат за вземане на управлявано решение. Използвайки тези инструменти в разширена Ламбда архитектура и сигурен вход в системата, организациите могат да изградят стабилни и мащабируеми канали за обработка на данни, които могат да обработват големи обеми данни и да поддържат обработка и анализ на данни в реално време.

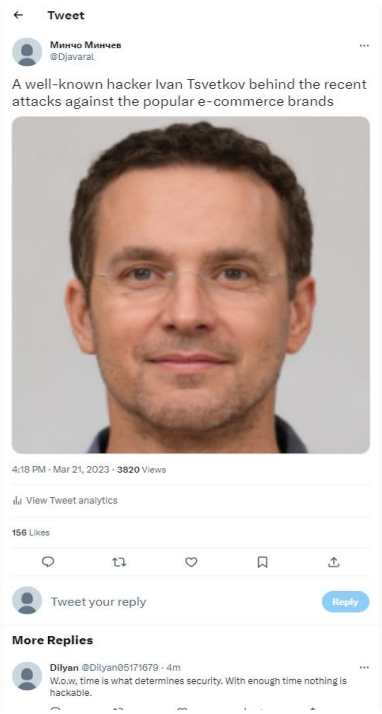
Управление на сигурността на трето ниво от архитектурата на оперативен център за информационна сигурност

За управление на сигурността от трето ниво се използват системи за управление на информация за сигурността и събития (SIEM) за агрегиране и консолидиране на данни от множество източници, включително мрежи, сигурност, сървъри и бази данни. IBM Qradar е една такава система, която използва усъвършенствани алгоритми за анализ и машинно обучение за откриване и приоритизиране на заплахи за сигурността в реално време. Micro Focus Sentinel събира данни от различни източници и ги нормализира до общ формат, което прави анализа и корелацията по-лесни. Когато Qradar идентифицира потенциална заплаха, той генерира предупреждение, което може да бъде персонализирано въз основа на сериозността и типа. За анализиране и визуализиране на данни, събрани от лог файлове, се използва Micro Focus IDOL. IDOL използва текстови анализи и техники за обработка на естествен език, за да извлече подходяща информация от регистрационни файлове, а

неговият набор от инструменти включва табла за управление, диаграми и графики за изследване и анализиране на регистрационни данни. Когато бъдат открити потенциални проблеми или аномалии, Qradar извежда известие за потенциалната заплаха. Заедно Qradar и IDOL позволяват на бизнеса да анализира и визуализира данните за сигурността в реално време, за да идентифицира потенциални заплахи за сигурността.

Верифициране на дефинирания метод с използване на архитектурни решения за известяване на потенциална заплаха

Предложеният оперативен център за информационна сигурност използва различни софтуерни инструменти за наблюдение и защита на мрежи и данни. Представен е примерен случай, който демонстрира ефективността на системата. Първият процес включва извличане на публикация от платформа за социални медии, анализиране на данните и съхраняване на резултатите в хранилището на Hadoop – Фигура 26.



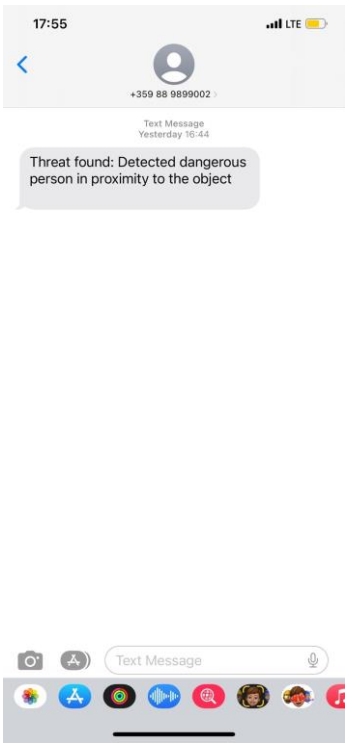
Фигура 26. Публикация в социална медия относно известен хакер.

Вторият процес включва извличане на данни от видео поток от камера, извършване на анализ от Media Server и разпознаване на лица – Фигура 27, Фигура 28.

Фигура 27. Визуализиране на извлечената публикация от социална медия.

Фигура 28. Изходен код от Media Server с данни на съответното разпознато лице.

След това мултимедийният сървър съпоставя идентифицираното лице с данните, съхранени в Hadoop, и ако има съвпадение, QRadar генерира SMS аларма към екипа на ОЦИС. Работейки заедно, тези софтуерни инструменти предоставят на организациите по-ефикасен и ефективен начин за наблюдение и защита на техните мрежи и данни – Фигура 29.



Фигура 29. SMS аларма за потенциална опасност за близост на обект.

Съвместната работа на тези софтуерни инструменти може да доведе до по-голям обхват и ефективно решение на проблеми, свързани с ИТ и киберсигурността, което от своя страна ще позволи на организациите да наблюдават, анализират и защитават по-добре своите мрежи и данни.

Изводи

Извършена е проверка на предложената архитектура за ОЦИС в система за големи данни, демонстрираща предимствата от съвместната работа на софтуерни решения. Micro Focus Sentinel и macmon Network Access Control се използват за управление на първо ниво, събиране на защитени данни от устройства и системи и наблюдение за заплахи за сигурността. Micro Focus IDOL и NiFi се използват за сигурно извличане и съхраняване на данни в среда с големи данни, докато LDAP и многофакторно удостоверяване с OpenLDAP защитен достъп до системата. BPM и разширената Lambda архитектура се използват като един процес с NiFi, IDOL и PowerBI за генериране на графики и диаграми за бизнес заключения. IDOL се използва за визуализация на анализ на лог файл, а QRadar се използва за известия в реално време и анализ на лог файл.

5. Заключение

Осигуряването на информационна сигурност в системите за големи данни е сложна задача, изискваща ефективни оперативни центрове за управление на информационната сигурност, които могат да откриват и реагират на заплахи своевременно. В тази връзка предлагаме метод за проектиране на оперативен център за информационна сигурност, състоящ се от три нива на управление, използващи принципи и методи от света на информационната и киберсигурността, както и контроли от ISO 27001. Първото ниво включва адаптивна сигурност и интелигентна обработка, второто осигурява вътрешна сигурност в средата Hadoop и се фокусира върху сигурността на софтуерния бизнес, докато третото ниво използва SIEM система за наблюдение и анализ на събития за сигурност в реално време. За валидиране на предложения метод е извършена проверка с помощта на няколко предложени инструмента, което доведе до по-високо ниво на сигурност.

6. Списък на публикациите по темата на дисертационния труд

1. Ivona Velkova, Security challenges for big data platforms, 10TH INTERNATIONAL CONFERENCE ON APPLICATION OF INFORMATION AND COMMUNICATION TECHNOLOGY AND STATISTICS IN ECONOMY AND EDUCATION ICAICTSEE – 2020, November 27 – 28th, 2020, University of National and World Economy, Sofia, Bulgaria, ISSN 2367-7635 (PRINT),ISSN 2367-7643 (ONLINE), достъпна на: https://icaictsee.unwe.bg/past-conferences/ICAICTSEE-2020.pdf, стр. 482-488

2. Ivona Velkova, Mariana Kovacheva, Digitalization in Bulgarian Higher Education – Present and Future Opportunities, 4th International Academic Conference on Education, 2021, 10th December 2021, Barcelona, Spain, ISBN: 978-609-485-239-8), достъпна на: https://www.dpublication.com/proceeding/4th-iaceducation/#Table-of-Contents, стр. 1-13

3. Ivona Velkova, Unstructured Data Processing and Analysis Using Artificial Intelligence, Обработка и анализ на неструктурирани данни с помощта на изкуствен интелект, сп. Автоматика и информатика, ISSN 0861-7562 (Print), ISSN 2683-1279 (Online), Year LV No. 4/2022, достъпна на: https://sai-bg.com/wp-content/uploads/2023/01/AI-4-2022.pdf, стр. 27-30

4. Ivona Velkova, Unstructured social media data processing with artificial intelligence, VIII INTERNATIONAL SCIENTIFIC CONFERENCE HIGH TECHNOLOGIES. BUSINESS.SOCIETY 2023, 06-09.03.2023, BOROVETS, BULGARIA, ISSN 2535-0005(PRINT), ISSN 2535-0013 (ONLINE), достъпна на: http://hightechsociety.eu/sbornik/2023.pdf, стр. 45-48

Очаква се публикуване: Ivona Velkova, Approaches to higher security level for Hadoop environment

Литература

[1] “Total data volume worldwide 2010-2025,” *Statista*. https://www.statista.com/statistics/871513/worldwide-data-created/ (accessed Oct. 09, 2022).

[2] K. Stefanova and D. Kabakchieva, “Challenges and Perspectives of Digital Transformation,” *Conferences of the department Informatics*, no. 1, pp. 13–23, 2019.

[3] G. Sriram, “SECURITY CHALLENGES OF BIG DATA COMPUTING,” p. 8, Jan. 2022.

[4] InfoSec (www.infosec.gov.hk), “InfoSec: Core Security Principles,” *InfoSec*. https://www.infosec.gov.hk/en/knowledge-centre/core-security-principles (accessed Mar. 11, 2023).

[5] “7 SecOps roles and responsibilities for the modern enterprise | TechTarget,” *Security*. https://www.techtarget.com/searchsecurity/feature/7-SecOps-roles-and-responsibilities-for-the-modern-enterprise (accessed Jan. 29, 2022).

[6] D. Hain, R. Jurowetzki, S. Lee, and Y. Zhou, “Machine learning and artificial intelligence for science, technology, innovation mapping and forecasting: Review, synthesis, and applications,” *Scientometrics*, Jan. 2023, doi: 10.1007/s11192-022-04628-8.

[7] “The Global State of Digital in July 2022 — DataReportal – Global Digital Insights.” https://datareportal.com/reports/digital-2022-july-global-statshot (accessed Feb. 15, 2023).

[8] “Number of internet users worldwide 2022 | Statista.” https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/ (accessed Mar. 29, 2023).

[9] B. Scalzo, “Securing Structured Data,” *Stealthbits Technologies*, Apr. 24, 2019. https://stealthbits.com/blog/securing-structured-data/ (accessed Mar. 11, 2023).

[10] P. Lo Giudice, L. Musarella, G. Sofo, and D. Ursino, “An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake,” *Information Sciences*, vol. 478, pp. 606–626, Apr. 2019, doi: 10.1016/j.ins.2018.11.052.

[11] MicroFocus, “Saba Cloud Pro: IDOL01WBT - IDOL Essentials 12.6 Digital Learning for Administrators with Specialist Exam.”

[12] “What Is Unstructured Data?,” *MongoDB*. https://www.mongodb.com/unstructured-data (accessed Mar. 29, 2023).

[13] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, “Uncertainty in big data analytics: survey, opportunities, and challenges,” *Journal of Big Data*, vol. 6, no. 1, p. 44, Jun. 2019, doi: 10.1186/s40537-019-0206-3.

[14] “Internet and social media users in the world 2023,” *Statista*. https://www.statista.com/statistics/617136/digital-population-worldwide/ (accessed Feb. 27, 2023).

[15] “The Latest Facebook Statistics: Everything You Need to Know — DataReportal – Global Digital Insights.” https://datareportal.com/essential-facebook-stats (accessed Mar. 01, 2023).

[16] “10 Google Search Statistics You Need to Know in 2023 | Oberlo,” Jan. 13, 2023. https://www.oberlo.com/blog/google-search-statistics (accessed Mar. 02, 2023).

[17] “The 42 V’s of Big Data and Data Science,” *Elder Research*. https://www.elderresearch.com/blog/the-42-vs-of-big-data-and-data-science/ (accessed Mar. 09, 2023).

[18] К. Стефанова and С. Йорданова, “ПРЕДИЗВИКАТЕЛСТВАТА НА ГОЛЕМИТЕ ДАННИ СЪЩНОСТ, ХАРАКТЕРИСТИКИ И ТЕХНОЛОГИИ.” https://docplayer.bg/166167087-предизвикателствата-на-големите-данни-същност-характеристики-и-технологии.html (accessed Nov. 20, 2022).
[19] Techreviewer, “The Most Popular Big Data Frameworks in 2022 | Techreviewer Blog.” https://techreviewer.co/blog/the-most-popular-big-data-frameworks-in-2022 (accessed Oct. 09, 2022).
[20] W. Inoubli, S. Aridhi, H. Mezni, M. Maddouri, and E. Mephu Nguifo, “An experimental survey on big data frameworks,” *Future Generation Computer Systems*, vol. 86, pp. 546–564, Sep. 2018, doi: 10.1016/j.future.2018.04.032.
[21] Y. Filaly, N. Berros, H. Badri, F. E. mendil, and Y. E. B. EL Idrissi, “Security of Hadoop Framework in Big Data,” in *Artificial Intelligence and Smart Environment*, Y. Farhaoui, A. Rocha, Z. Brahmia, and B. Bhushab, Eds., in Lecture Notes in Networks and Systems. Cham: Springer International Publishing, 2023, pp. 709–715. doi: 10.1007/978-3-031-26254-8_103.
[22] “Apache Hadoop.” https://hadoop.apache.org/ (accessed Mar. 09, 2023).
[23] IBM Cloud Education, “What is Apache Spark?,” Sep. 14, 2022. https://www.ibm.com/cloud/learn/apache-spark (accessed Oct. 09, 2022).
[24] “What is Apache Hive? | AWS.” https://aws.amazon.com/big-data/what-is-hive/ (accessed Mar. 11, 2023).
[25] MicroFocus, “Powered By Idol,” *Micro Focus*. https://content.microfocus.com/idol-analytics-21/powered-by-idol-sdks (accessed Oct. 09, 2022).
[26] N. Millman, “8 considerations when selecting big data technology,” *Computerworld*, Feb. 27, 2014. https://www.computerworld.com/article/2475840/8-considerations-when-selecting-big-data-technology.html (accessed Feb. 24, 2023).
[27] “7 Criteria for Choosing Big Data Developers - N-iX,” *Software Development Company - N-iX*. https://www.n-ix.com/7-criteria-choosing-big-data-developers/ (accessed Sep. 28, 2021).
[28] “Apache NiFi Overview.” https://nifi.apache.org/docs/nifi-docs/html/overview.html (accessed Mar. 11, 2023).
[29] “Supervised vs. Unsupervised Learning: What’s the Difference?,” Nov. 15, 2022. https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning (accessed Mar. 05, 2023).
[30] “What is Artificial Intelligence (AI)? | Definition from TechTarget,” *Enterprise AI*. https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence (accessed Mar. 11, 2023).
[31] “What is Natural Language Processing? | IBM.” https://www.ibm.com/topics/natural-language-processing (accessed Mar. 11, 2023).
[32] N. James, “160 Top Cybersecurity Statistics 2023: Figures, Facts & Trends,” Dec. 19, 2022. https://www.getastra.com/blog/security-audit/cyber-security-statistics/ (accessed Mar. 06, 2023).
[33] “Critical cybersecurity areas worldwide 2023,” *Statista*. https://www.statista.com/statistics/1292944/critical-cybersecurity-area-worldwide/ (accessed Mar. 06, 2023).
[34] “What is Information Security | Policy, Principles & Threats | Imperva,” *Learning Center*. https://www.imperva.com/learn/data-security/information-security-infosec/ (accessed Mar. 29, 2023).

[35] Y. Soo, “Discussion and Comparison of Several Hadoop Security Tools | by Yinyi Soo | Medium.” https://ysoo23.medium.com/discussion-and-comparison-of-several-hadoop-security-tools-b4532a8c67f9 (accessed Feb. 10, 2022).
[36] Cloudera, “Apache Ranger,” *Cloudera*. https://www.cloudera.com/products/open-source/apache-hadoop/apache-ranger.html (accessed Feb. 11, 2022).
[37] A. Javed, “3 Basic A’s of Identity and Access Management -Authentication, Authorization, and Accounting.” https://www.xorlogics.com/2019/04/15/3-basic-as-of-identity-and-access-management-authentication-authorization-and-accounting/ (accessed Feb. 10, 2022).
[38] EC-Council, “Understanding the Role of a Security Operations Center,” *Cybersecurity Exchange*, Apr. 28, 2022. https://www.eccouncil.org/cybersecurity-exchange/security-operation-center/responsibilities-security-operations-center-soc-team/ (accessed Mar. 29, 2023).
[39] M. Vielberth, F. Bohm, I. Fichtinger, and G. Pernul, “Security Operations Center: A Systematic Study and Open Challenges,” *IEEE Access*, vol. 8, pp. 227756–227779, 2020, doi: 10.1109/ACCESS.2020.3045514.
[40] S. MS, “Useful KPIs for a Security Operation Center (SOC),” Dec. 19, 2019. https://www.cloudcybersafe.com/useful-kpi-elements-for-a-security-operation-center-soc/ (accessed Mar. 29, 2023).
[41] LogRhythm, “7 Steps to Building A Security Operations Center (SOC),” *LogRhythm*, Jun. 16, 2020. https://logrhythm.com/blog/7-steps-to-build-your-security-operations-center/ (accessed Jan. 18, 2023).
[42] “How to Build a Security Operations Center (SOC): Peoples, Processes, and Technologies,” *Digital Guardian*. https://www.digitalguardian.com/blog/how-build-security-operations-center-soc-peoples-processes-and-technologies (accessed Mar. 29, 2023).
[43] “The Evolution of Security Operations and Strategies for Building an Effective SOC,” *ISACA*. https://www.isaca.org/resources/isaca-journal/issues/2021/volume-5/the-evolution-of-security-operations-and-strategies-for-building-an-effective-soc (accessed Mar. 29, 2023).
[44] A. T. A. J. Kienzle, “What Is a SOC? Top Security Operations Center Challenges,” *IIoT World*, Jan. 28, 2022. https://www.iiot-world.com/ics-security/cybersecurity/top-challenges-soc-are-facing/ (accessed Mar. 29, 2023).
[45] “5G/SOC: SOC Generations -HP ESP Security Intelligence and Operations Consulting Services - Business white paper”.
[46] E. R. https://www.emergenresearch.com, “SOC as a Service Market Size, Share | Industry Forecast by 2030.” https://www.emergenresearch.com/industry-report/security-operations-center-as-a-service-market (accessed Apr. 17, 2023).
[47] “Atos opens new global next-gen Security Operations Center in Bulgaria and strengthens its sovereign security offering in Europe,” *Atos*, Mar. 22, 2022. https://atos.net/en/2022/press-release_2022_03_22/new-global-security-operations-center-in-bulgaria (accessed Apr. 17, 2022).
[48] “Next Gen CR - PROCYB srl,” Oct. 06, 2021. https://procyb.io/en/next-generation-soc-en/ (accessed Apr. 17, 2023).
[49] “ISO 27001 Annex A.5 - Information Security Policies,” *https://www.isms.online/*. https://www.isms.online/iso-27001/annex-a-5-information-security-policies/ (accessed Apr. 23, 2023).

[50] D. Kosutic, “What are the 11 new security controls in ISO 27001:2022?” https://advisera.com/27001academy/explanation-of-11-new-iso-27001-2022-controls/ (accessed Apr. 23, 2023).
[51] “ISO 27001 Annex A.9 Access Control - Your Step-by-Step Guide,” *https://www.isms.online/*. https://www.isms.online/iso-27001/annex-a-9-access-control/ (accessed Apr. 24, 2023).
[52] “ISO 27001 Annex A.16 - Information Security Incident Management,” *https://www.isms.online/*. https://www.isms.online/iso-27001/annex-a-16-information-security-incident-management/ (accessed Apr. 24, 2023).
[53] “Adaptive architecture: Key to True Cybersecurity | Kaspersky official blog.” https://www.kaspersky.com/blog/asa-key-to-true-cybersecurity/6678/ (accessed Dec. 10, 2022).
[54] “Designing an Adaptive Security Architecture for Protection From Advanced Attacks,” *Gartner*. https://www.gartner.com/en/documents/2665515 (accessed Mar. 28, 2023).
[55] “Set Up Containerize and Test a Single Hadoop Cluster using Docker and Docker compose,” *Engineering Education (EngEd) Program | Section*. https://www.section.io/engineering-education/set-up-containerize-and-test-a-single-hadoop-cluster-using-docker-and-docker-compose/ (accessed Mar. 29, 2023).
[56] K. Miao, J. Li, W. Hong, and M. Chen, “A Microservice-Based Big Data Analysis Platform for Online Educational Applications,” *Scientific Programming*, vol. 2020, p. e6929750, Jun. 2020, doi: 10.1155/2020/6929750.
[57] “Enterprise Business Intelligence | Sentinel.” https://www.microfocus.com/en-us/cyberres/secops/sentinel (accessed Mar. 29, 2023).
[58] “Zero Trust Network Access.” https://www.macmon.eu/en/ (accessed Mar. 29, 2022).
[59] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, “Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda,” *International Journal of Information Management*, vol. 48, pp. 63–71, Oct. 2019, doi: 10.1016/j.ijinfomgt.2019.01.021.
[60] “Big data Hadoop and MapReduce solutions for CMS content management problems,” *TheServerSide.com*. https://www.theserverside.com/tutorial/How-big-data-solved-the-content-management-CMS-problem (accessed Mar. 02, 2023).
[61] KARDEN, “Authentication in Hadoop cluster: MIT Kerberos and Active Directory – DekarLab,” May 23, 2020. https://dekarlab.de/wp/?p=883 (accessed Mar. 29, 2022).
[62] “Set up Okta Verify (MFA) | eSolutions.” https://www.monash.edu/esolutions/phones/change-device-multi-factor-authentication (accessed Mar. 29, 2023).
[63] “Best Practices Guide for Systems Security Services Daemon Configuration and Installation - Part 1 - Cloudera Blog.” https://blog.cloudera.com/best-practices-guide-for-systems-security-services-daemon-configuration-and-installation-part-1/ (accessed Mar. 29, 2022).
[64] A. Luntovskyy and D. Gütter, “From Big Data to Smart Data: Best Practices for Data Analytics,” in *Highly-Distributed Systems: IoT, Robotics, Mobile Apps, Energy Efficiency , Security*, A. Luntovskyy and D. Gütter, Eds., Cham: Springer International Publishing, 2022, pp. 79–96. doi: 10.1007/978-3-030-92829-2_4.
[65] P. P. Sharma, “Securing Big Data Hadoop : A Review of Security Issues , Threats and Solution,” 2014. https://www.semanticscholar.org/paper/Securing-Big-Data-Hadoop-%3A-

A-Review-of-Security-%2C-Sharma/fedad0e2e8a2eb2a81a572679c233ee94b1a2bf8 (accessed Feb. 10, 2022).
[66] "Apache Hadoop Amazon Web Services support – Object Store Auditing." https://hadoop.apache.org/docs/current/hadoop-aws/tools/hadoop-aws/auditing (accessed Feb. 10, 2023).
[67] W. Rowe, "Introduction to Hadoop Security," *BMC Blogs*. https://www.bmc.com/blogs/hadoop-security/ (accessed Feb. 10, 2022).
[68] "Hadoop KMS – Hadoop Key Management Server (KMS) - Documentation Sets." https://hadoop.apache.org/docs/stable/hadoop-kms/index.html (accessed Mar. 29, 2023).
[69] R. Ross, M. McEvilley, and J. Carrier Oren, "Systems Security Engineering: Considerations for a Multidisciplinary Approach in the Engineering of Trustworthy Secure Systems," National Institute of Standards and Technology, NIST SP 800-160, Nov. 2016. doi: 10.6028/NIST.SP.800-160.
[70] "Use BPM Software to Improve Enterprise Security," *abas*. https://abas-erp.com/en/resources/erp-blog/improve-enterprise-security-bpm-software (accessed Mar. 29, 2023).
[71] "Artificial Intelligence: Transforming Business Security Strategies." https://www.securityinformed.com/insights/artificial-intelligence-transforming-business-security-strategies-co-3425-ga.1555480394.html (accessed Mar. 29, 2023).
[72] "What Is Lambda Architecture?" https://www.databricks.com/glossary/lambda-architecture (accessed Mar. 29, 2023).
[73] "Definition of SIEM - IT Glossary | Gartner." https://www.gartner.com/en/information-technology/glossary/security-information-and-event-management-siem (accessed Mar. 29, 2023).

Списък фигури

Списък таблици

UNIVERSITY OF NATIONAL AND WORLD ECONOMY

Faculty of Applied Informatics and Statistics

Department of Information Technologies and Communications

„Principles and methods for designing an operational center for managing information security for big data systems“

Abstract of the dissertation

for acquiring an educational

and scientific PhD degree

Doctoral student:

Ivona Velkova

Scientific Manager:

Prof. Dr. Lyuben Boyanov

Sofia, 2023

Table of Contents

General characteristics of the dissertation

1. The relevance of the problem

The growing digitalization of processes, objects and activities in our everyday life is a modern trend caused by technological progress through the last few decades. One of the results of this digitalization is the significant increase in the volume of data generated from different sources [1]. With the entry of digital transformation into almost every area of human activity, data has become one of the main assets for obtaining a competitive advantage for any business. This is a ubiquitous process and affects practice in all areas of human activity - from health, finance, logistics, transport to ecology and sports. All these activities constantly collected huge amounts of data and information. The data may be of a different nature - for markets, customers, competitors, various processes, quantities of goods, state of environment, video surveillance, social media data, etc. Thanks to this data, people receive information that can be analyzed and benefit business. The process of extracting knowledge from data collection and their analysis is able to optimize and bring greater efficiency to any business or parts of him. With the knowledge gained, companies can identify what is important to them and what additional or alternative actions they may or should take to update or optimize their ultimate success targets [2].

According to Statista, the total amount of data generated and consumed worldwide scale reached 97 zetabytes (ZB) in 2022, a significant increase compared to previous years [1]. For comparison, in 2020 the number of data generated was 64 ZB [1]. As the volume of data generated increases, so do the risks, related to their protection - the larger the data, the more complex they are the mechanisms and approaches by which these data are collected and analyzed [3].

Information security is essential for the protection of information in an organization. This applies to customer data, financial records, intellectual property, etc., which may be subject to cyber threats such as malware, phishing, attacks through social engineering, etc. Information security is a process that aims to prevent unauthorized access, counteract various

types of threats, provide confidentiality and reduce the risk of destruction or modification of the stored information [4].

In order to be effectively managed, information security is necessary to identify, assess and manage the risks of information assets of the organization. This requires an in-depth and integrated approach to security, which includes the implementation of policies, standards, procedures and technical control for protection against potential threats. Because of the above reasons, the existence of an operational information management center security that can monitor and respond to security threats is from crucial for protecting digital information and ensuring appropriate business security level [5].

Data protection and information during storage, processing and analysis of big data in a number of cases is carried out by cloud and software solutions, which are characterized by a high level of security. Increasingly often these systems include Artificial Intelligence (III), which increases the efficiency of processing large amounts of data, including unstructured data in the form of text, video, audio and web content [6].

The present work proposes an approach that ensures security of different levels of data, processes and communications in large systems data. The data is extracted from heterogeneous sources, some of which are video streams from cameras, data from websites, social media, etc. The latter are actively used in business, as they offer a set of opportunities - from increasing awareness of a trademark, model or service, to informing potential customers and sales stimulation. As such, they are one of the important ones modern digital tools for even greater business success. According to a report by Datareporta as of July 2022, 59% of people who use the Internet worldwide use social media platforms [7]. This shows the great importance of these media as a source of business data.

It is not known now that a method is proposed that unites different software solutions for big data systems to extract and process data with AI from different sources to which cognitive and adaptive have been applied security, and which at the same time be confirmed, presented and separated into a single operations center. Each big data system has technical specifics, depending on the specifics of its application, the number of active users, place of activity and technologies of implementation. The presented study can help professionals

working in the field of big data to increase the level of security through the application of the proposed principles and methods, building a single management center to provide different levels of security. The presented solution can be further developed and extended with additional functionalities so as to meet the needs of different types of business and users.

2. Object and subject of the study

The object of the study is principles and methods for the design of an operational center for security management in big data systems.

The subject of the study is the output of a method for building a new type of operational security management center based on treatment with AI of unstructured data, big data systems and Security Information and Event Management (Security information and event management - SIEM), integrated in general architecture.

The research problem of the dissertation is the combination of technologies and providing different levels of security in the construction of an operational center in big data systems.

3. Purpose and tasks of the dissertation

The purpose of the dissertation is to offer principles and methods for design and the construction of an operational center to manage information security, for systems operating with big data.

Tasks that are performed in the process of realization of the set goal of the dissertation is:

- To study the types of data and the advantages and disadvantages of existing means for storage and processing of big data;
- After analyzing existing solutions and good practices worldwide in this area, consider principles and methods of management different levels of information security;
- Review models and approaches when using artificial intelligence to ensure a higher level of security;

- Review operational security centers (TACIS) as consider the different generations of the CCIS;
- Based on the analysis, to define an approach for designing OSIC in systems for large security management data;
- Design a functional architecture with different levels of security to propose a solution to the defined research problem in the dissertation work and the proposed architecture to be able to use the set principles and methods for designing and building an operational center for management of security;
- To extract data from heterogeneous sources and to store in a single system for big data and to establish a level of security in their extraction and subsequent processing;
- Analyze the level of security in the storage and processing system big data and to define a level of adaptive security; to be considered the internal-organized security of the big data system; to be separated software business security;
- To propose approaches for building an operational center that will manage security in different processes in big data systems.

The design approach and software systems „ bottom-up “. With this approach, the process begins with the design of the main components of the system, which then should be combined to create complex operational modules center.

4. Working hypotheses of the dissertation

The main working hypotheses for the current dissertation are:

Hypothesis I:

It is possible to create a model that describes the design of an operational information security management center at different levels for large systems data and data extraction from heterogeneous sources.

Hypothesis II:

It is possible to achieve effective integration of various technologies and systems for working with big data.

Hypothesis III:

It is possible to create a multilayer architecture with different approaches to providing security.

5. Scientific and scientific-applied contributions

1. The nature and component structure of „ operational center for security “ and an up-to-date definition and necessary elements have been defined, in relation to the modern conditions for the functioning of such a center in the environment of big data.
2. Criteria for comparing technologies for processing large ones are proposed data, in relation to the objectives of the development in terms of collection, the organization and storage of unstructured data, with opportunities for application of artificial intelligence tools.
3. Basic principles and methods for managing an operational center for information security covering the process, functionalities, and levels of security.
4. Current principles have been proposed for the design and construction of the OSIC, necessary to cover the specifics of security management systems operating in a big data environment.
5. A method for designing and creating a functional architecture has been defined of the OSIC. The proposed architecture has three levels of security management, covering the network level of security, the process of extraction and processing data, verification of internal-based security in a big data environment and analysis of the obtained results.
6. A prototype built with the technologies of the proposed architecture has been realized purpose covering all necessary functionalities. The prototype connects NiFi, Micro Focus IDOL and Apache Hadoop technologies and has been tested with data from social media, video stream, data from websites and log files.

6. Volume and structure of the dissertation

The dissertation is in a volume of 170 pages, of which 153 contain the actual study without applications. The structure of the dissertation consists of:

- Introduction;
- Three chapters in which an analysis of the subject area is realized, the subject area is removed basic definitions related to the research problem are presented the prerequisites for the realization of labor and a method is defined with which to achieve this. The proposed method is verified using different technologies;
- Conclusion, which includes a list of scientific and scientific applications contributions, List of publications made on the topic of the dissertation labor, Future work, Literature, Lists of figures and tables and List of terms and abbreviations.

Chapter One (Introduction) focuses on the timeliness of the view problem by considering the object, subject, research problem and objectives of dissertation. Three hypotheses have been defined that need to be proven.

Chapter Two is the research part of the research. In it place the theoretical foundations of the meaning and types of data, the means are considered and approaches for processing big data and comparison is made according to selected criteria, to help choose the appropriate tools for the dissertation. Chapter Two discusses artificial intelligence (AI) applications in the field security, principles and methods are derived that are used to ensure high level of security in organizations and its definition is presented the component structure of the operational security information center (OSIC).

Chapter Three focuses on the design of the OSIC. They are taken out in it principles and methods to be applied for the design of a center of this kind, the existing generations of OSIC have been considered and the challenges have been identified in front of them. As a result of these things, current principles are proposed and controls of the international standard ISO 27001 for the design of a new generation OSIC. This chapter proposes a method that describes the new type of OSIC in functional architecture at three levels for security management. First

level covers adaptive security and intelligent data processing by heterogeneous sources, the second level focuses on the internal-based security of Hadoop system for big data and software business security of processes in organization while the third level focuses on visualization of the analyzed information from the previous two levels and in the presence of potential security vulnerability - notifying the OSIC team to take specific actions.

Chapter Four focuses on proving the functionality of the proposed one architecture using different technologies at each of the proposed three levels. A working prototype with an experimental goal for verification of a worker is presented process using the proposed OSIC.

Chapter 5 (Conclusion) summarizes the work on the dissertation, the scientific and applied contributions are derived, as well as guidance is given for the future work of scientific work.

Brief presentation of the dissertation

1. Introduction

The introduction provides an overview of the timeliness of the chosen topic, the object, the object, defines the tasks to be performed, as well as formulates three hypotheses. The purpose of the dissertation is to offer principles and methods for design and construction of an operational center to manage information security, for systems operating with big data.

2. Security in a big data environment and an Operational Security Information Center

Data has been used for centuries, but nowadays digital technologies are ubiquitous to over 5 billion users in a connected Internet environment [8]. Each digital action generates data that a business can use to improve efficiency and to discover new opportunities. The data are classified in structured, semistructured and unstructured types and can provide valuable information for making strategic decisions and competitive market advantage.

- **Structured data** is organized quantifiable data, often in tabular form as SQL databases. Columns have data types - text or numeric and rows have specific values. This data are secure thanks to controlled access and established technology. They are used in sensitive applications such as finance or healthcare [9].

- **Semi-structured data** do not have a fixed scheme and allow changes without violating their structure. They are good for dynamic data sets and use metadata as tags or attributes to describe the shape, structure, and their meaning. It is usually extracted in XML, HTML and JSON documents [10].

- **Unstructured data** includes data generated by humans and machines, such as social media posts, images and data from sensors [11]. It constitutes 80% of global data, challenging organizations that try to derive a value of [12].

Unlike structured and semi-structured data, where there is specific processing methods, in the case of unstructured data, the extraction of knowledge and analysis is a problem, as the diversity of formats require specialized tools for their processing [13]. Part of these tools are Apache Hadoop, Micro Focus IDOL, IBM Watson, AWS, and others.

The analysis of information extracted from unstructured data is important because c such an approach organization can get an idea of the behavior of customers, market trends and other business-relevant information that would it was difficult or impossible to obtain otherwise.

Big data

Big data is a large, complex and fast-growing collection of structured, semi-structured and unstructured data generated from various sources such as social media, web search engines and IoT devices. Statistics Statistics Statistics for social media show that in 2022 the number of consumers was 4.26 billion worldwide – the number is projected to increase over the next few years [14]. Only for the period October 2022 to January 2023 there is an increase in the number of active users of the Facebook platform worldwide with 5 million [15]. The bigger one part of the global growth of social media is due to the growing use of mobile devices with which users write posts, express opinions, upload photos or respond to posts of other users. In addition to social media, web search engines for various inquiries are also widely used. A study shows that by the end of 2022. Google receives 8.5 billion requests a day, equivalent to almost 100 thousands of requests per second [16]. This data indicates that the information generated is constantly growing.

Big data can be best defined with the six V's: volume, speed, variety, reliability, value and variability. Initially, big data were often presented as the 3 V – volume, speed and variety, but up to 42 V [17].

- The volume refers to the huge amount of data generated and collected by individuals, organizations, and machines.
- Speed refers to the speed at which the data is generated, processed, and analyzed in real time.

- Diversity refers to the variety of data types, sources, and formats, including structured, semi-structured and unstructured data.
- Reliability refers to the reliability, accuracy, and consistency of data.
- The value refers to the extraction of knowledge and benefits that can be obtained from the analysis and use of big data.
- Variability refers to the unpredictability of the data generated and collected from various sources, including social media, IoT devices and sensors. For example, social media data can vary greatly due to differences in languages, writing styles and idioms used by different users. Data variability management requires constant advanced data processing and analysis techniques that can process different data formats and quality levels.

The data volumes collected by the business also increase exponentially with the development of technologies are distributed in local, cloud and hybrid systems [18]. This increases both the complexity of management and access and data processing. Therefore, big data systems are developing and upgrading, and this process is expected to continue in the future.

Systems for big data and their processing

Specialized systems have been developed for storage and processing of big data. These environments use distributed file systems and parallel processing techniques for storing and processing large volumes of data. Examples of big data systems include Apache Hadoop, Amazon Redshift, Snowflake, and NoSQL databases such as MongoDB and Apache Cassandra [19], [20].

The work discusses Apache Hadoop, Apache Spark, Apache Hive, Micro Focus IDOL, and Apache NiFi.

- ***Apache Hadoop***

Hadoop is an open-source system for storing and processing various types of data in a distributed environment. It consists of HDFS for data storage and MapReduce for parallel processing between nodes [21]. To process the data, the Hadoop system first needs to store it in its directories, which is done by HDFS (Figure 1).

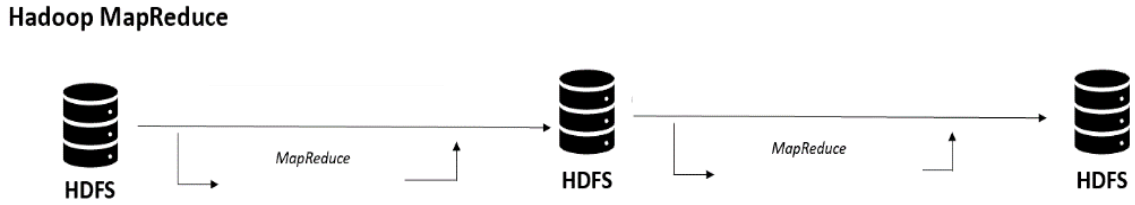


Figure 1. Processing with Hadoop.

Hadoop clusters can be scaled horizontally to process large amounts of data. MapReduce divides the data into smaller chunks and processes them in parallel, reading from HDFS for each step. The results are stored in HDFS or sent for further analysis. Hadoop can be integrated with other big data tools such as Spark for advanced analysis [22].

- ***Apache Spark***

Spark is an open-source software solution for big data processing with high performance, flexibility, and AI capabilities. It provides APIs for popular programming languages and consists of Spark Core, Spark SQL, Spark Streaming, and MLlib [23]. Spark Core provides distributed processing functionality and uses Resilient Distributed Datasets (RDD) as the primary data structure (Figure 2).

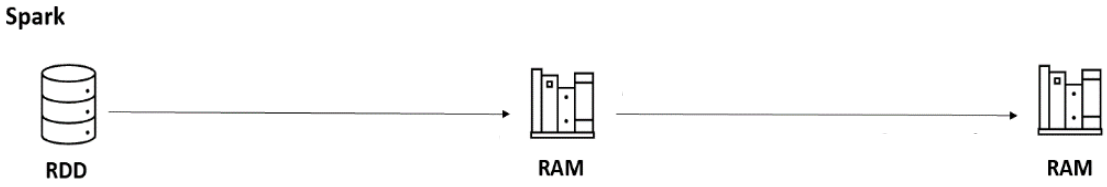


Figure 2. Processing with Spark.

RDDs provide fault tolerance and parallelism and can be created from various data sources. Spark SQL allows SQL-based queries, Spark Streaming enables real-time processing, and MLlib provides machine learning capabilities. Spark reads data from RDDs and stores them in RAM for faster processing and parallelizes the process across multiple nodes in a cluster.

- ***Apache Hive***

Hive is a data storage and query tool for big data sets stored in Hadoop's HDFS or other compatible file systems. It supports various data formats and uses a query language called HiveQL, based on SQL. Hive consists of components such as a query processor (Figure 3), Metastore, server, and execution module, which work together to process big data [24].

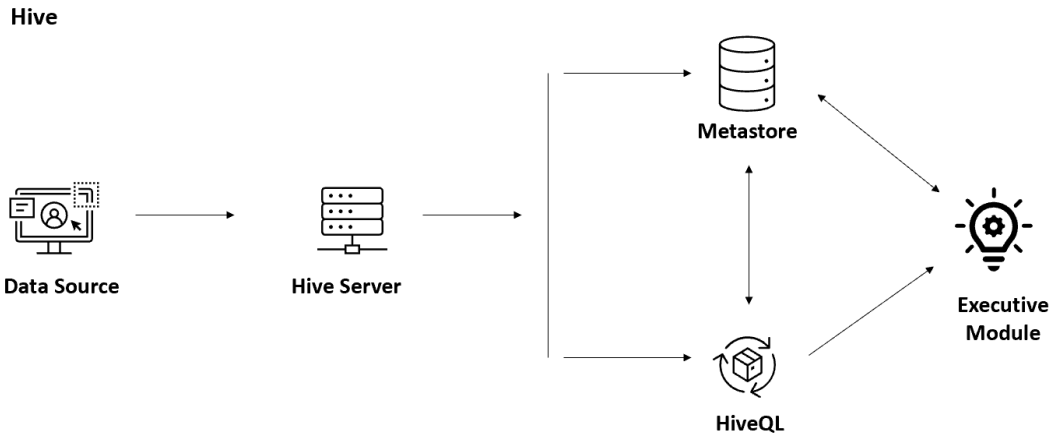


Figure 3. Processing with Hive.

Hive is not suitable for processing unstructured data such as text, images, or video and is optimized for batch processing. It can be used with other tools for advanced analysis.

- ***Micro Focus IDOL***

IDOL (Intelligent Data Operating Layer) uses techniques such as natural language processing, AI, and semantic analysis to extract knowledge from various types of data. The system has several components such as connectors, indexers, a distributed index manipulator, content analyzer, and query server (Figure 5) [25].

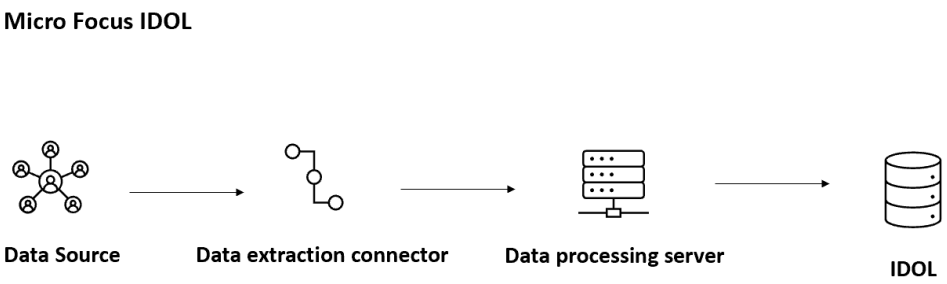


Figure 4. Processing with IDOL.

IDOL uses machine learning and semantic analysis to make predictions, classify data, and identify relationships between different data. The processed data and results are stored in IDOL's big data repository.

Comparison of selected criteria for big data

Based on the presented technological solutions, a summary of their capabilities is given in Table 1 based on selected criteria. These criteria are important because they ensure that data can be effectively managed and used to drive business success [26], [27].

The criteria presented in Table 1 were selected based on the specific goals and objectives of the current dissertation work. The presented technological products are compared based on the location for storing and processing big data, and attention is focused on the ability to process unstructured data, as they play a key role in business success. To make a decision on which tools to use, a comparison of the type of data processing is presented. This is an important condition because some of the products do not support real-time processing, which would be critical in critical moments for the organization. Next, attention is paid to the location of the data storage during processing, as this can affect the efficiency and speed of data processing, as well as the overall data storage costs. To ensure data security in case of potential vulnerability or data encryption on any of the servers, a criterion for error tolerance is presented. The use of AI is suitable for analyzing data from different types with proper understanding of the meaning of the data. It would contribute to improved speed in data analysis and extraction of dependencies, trends, and models.

Table 1. A comparison of big data technology environment.

Criteria	**Hadoop**	**Spark**	**Hive**	**IDOL**
Big data storage and processing	HDFS file system for storage and MapReduce for batch data processing	HDFS file system for storage and resilient distributed data set RDD for data processing	HDFS file system for storing and using HiveQL for Hadoop queries	Connectors for data extraction and IDOL server for processing and analysis
Processing of unstructured data	Yes	Yes	No	Yes
Processing type	Volumetric batch processing of historical data	Real-time and batch data processing	Batch processing of structured data historical data	Real-time and batch data processing
Data storage during processing	On the cluster disk in the HDFS system	In RAM	On the cluster disk in the HDFS system	Stores in RAM or cluster disk
Fault tolerance	Yes, by replicating the data with HDFS on the cluster nodes	Yes, by RDD abstraction with replication on cluster nodes	Yes, by replicating the data with HDFS on the cluster nodes	Yes, by configuring with redundant nodes for availability and automatic
Use of AI	No	Yes	No	Yes

Table 1 compares the strengths and weaknesses of four big data processing tools: Hadoop, Spark, Hive, and IDOL. Hadoop is suitable for batch processing but can be slow for real-time processing. Spark is faster but more complex and requires more memory than Hadoop. Hive has an easy-to-use interface but is not suitable for real-time processing or unstructured data. IDOL processes unstructured data with AI and is tolerant to errors. All four tools replicate data for fast recovery in case of system breaches.

Based on the obtained derived characteristics, a decision is made on which technological solutions will be used and which of them could be combined to achieve the result of the current work.

The NiFi system is used for integrating data processing and analysis tools.

- ***Apache NiFi***

NiFi is an open-source platform for automating the flow of data between different systems. It has various components such as processors, flow controllers, file storage, and connectors that work together to provide data integration capabilities (Figure 6) [28].

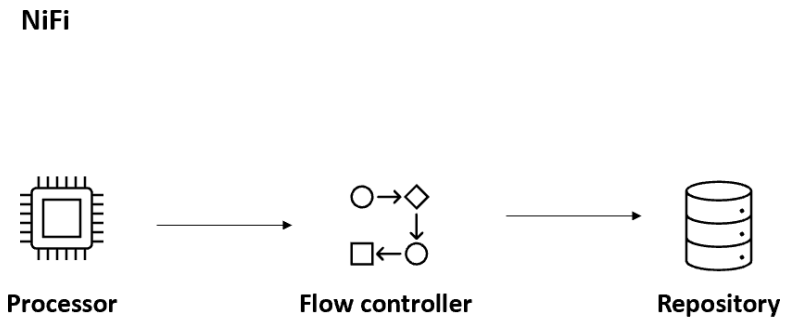


Figure 5. NiFi workflow process.

NiFi can direct and distribute data based on various criteria, transform data in different ways, and validate data quality. It is important to integrate different systems for maximum benefit. The three main models of artificial intelligence (natural language processing, machine learning, and deep learning) are also being explored for vulnerabilities and breaches in the system.

All of the aforementioned systems have different applications. To maximize their benefit, it is appropriate to integrate them into a unified system that meets business needs. To make a decision on which of the already discussed systems to use to achieve results, the three main models of artificial intelligence - natural language processing, machine learning, and deep learning - and their application in the field of security are examined to gain an understanding of the possibilities provided for detecting vulnerabilities and breaches in the system.

An artificial intelligence application for big data security

The development of artificial intelligence (AI) is one of the most significant technological achievements in recent times. AI is being applied in many places and its use in various fields of application has increased significantly in recent years. Examples of its use in our daily lives include smartphones, where facial or fingerprint recognition is used to unlock them. Another example is the detection of credit card fraud, where AI algorithms analyze patterns and anomalies in transactional data. This means that if a transaction is significantly larger than usual or if a customer starts making purchases that are not typical of their behavior, AI can mark it for further review and notify the system administrator of a potential threat [29].

AI is based on the idea of simulating human intelligence and applying it to digital devices that are programmed to perform tasks that usually require human intelligence, which is associated with processes such as learning, problem-solving, decision-making, and visual perception of generated information [30]. A disadvantage of using AI is that if it is trained on data that contains biases, it can lead to inaccurate results, which can be a serious challenge to identify and correct errors [31]. To overcome this drawback, the application of pre-processing techniques such as data cleaning and data normalization is necessary to reduce data deviation.

AI is used in big data security to protect against cyber-attacks by analyzing user behavior patterns, detecting anomalies in data, and developing predictive threat models. These AI applications can mark potential threats and help prevent future attacks. AI models are also used in information security to improve the accuracy and effectiveness of security systems.

Three models of AI usage in security have been discussed: natural language processing, machine learning, and deep learning, which provide different methods for managing security.

Principles and approaches for managing information security

Information security is a critically important issue for businesses as more and more data is stored and processed online or in systems connected to the internet. However, security isn't just about protecting data and processes online, but also in physical environments.

According to research and data from Astra, about 2,200 hacker attacks are carried out every day, which translates to about one online attack every 39 seconds [32]. A hacker attack is an

attempt by an unauthorized person to gain access to or damage a computer system, network, or data. Protection against these attacks is achieved by implementing strong security measures such as protective walls, intrusion detection and prevention systems, access controls, and training employees to recognize and avoid attacks such as social engineering [32].

According to a Statista survey from the end of 2022, the top five critical areas in cybersecurity for 2023 are data security, followed by personal data confidentiality, cybersecurity, risk analysis, and compliance [33]. The latter ensures that the organization's security measures are sufficient to protect sensitive information and prevent unauthorized access. Data security is the highest-rated critical area in cyberspace, as shown in Figure 7. This type of security should be viewed as a continuous process, part of an organization's culture, rather than a one-time event.

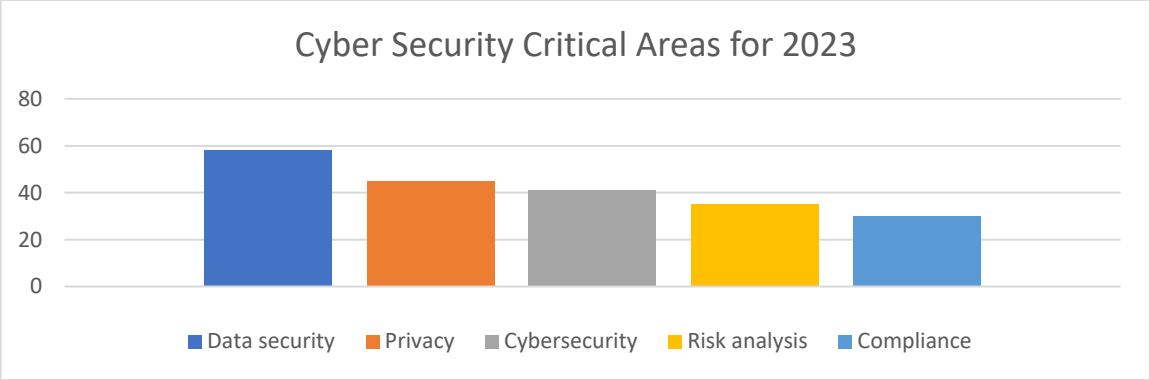


Figure 6. Cybersecurity zones.

The importance of data security management cannot be overstated, as a breach can have severe consequences for all affected parties. Data breaches are one of the most important and widespread problems in data security, as companies keep not only corporate data but also sensitive data about their customers.

To effectively manage security, several principles must be followed that help organizations design and implement effective measures to ensure the confidentiality, integrity, and availability of their information [34]:

- *Confidentiality*: This principle ensures that sensitive information is kept secret and only accessible to authorized persons. This can be achieved through methods such as data encryption, access control, and regular security checks;

- *Integrity*: This principle ensures that data remains accurate and unchanged both during transmission and at rest. This can be achieved through methods such as data archiving, hashing, version control, and secure coding practices;
- *Availability*: This principle ensures that data and resources are accessible to authorized persons when they need them. This can be achieved through methods such as backup systems, disaster recovery planning, and service-level agreements*;*
- *Authentication*: This principle ensures that users are who they claim to be before granting them access to resources. This is achieved through methods such as multi-factor or multi-step authentication such as sending a message via email or mobile device, entering a code from an image, entering a second password, etc.;
- *Authorization*: The principle of authorization ensures that users have the appropriate level of access to resources based on their role or level of authority. By providing authorization, organizations can limit potential damage that may result from a security breach or attack;
- *Non-repudiation*: This principle ensures that the authenticity of data or transactions cannot be denied by participating parties. This is achieved through methods such as digital signatures and protocols for secure communication.

While security principles provide a framework for effective information security, prevention, detection, and response are specific actions that organizations take to implement these principles and protect their sensitive information.

After outlining principles for an effective information security program, important methods used to achieve information security goals are described. There are many different security methods used to protect systems, data, and networks from unauthorized access, theft, or damage [3].

• Process Security: Implementation of controls and procedures to ensure that business processes are protected and not vulnerable to attacks or breaches, including policies for confidential information, training on best security practices, and security audits.

• Zero Trust: Extending security measures to all devices, applications, and users within an organization, requiring authentication and authorization before access is granted, and

focusing on protecting individual devices and assets with data through multifactor authentication, access control, and continuous monitoring.

• Multifactor Authentication: Requiring users to provide two or more forms of authentication to verify their identity before access to systems or data is granted, reducing the risk of unauthorized access and password disclosure attacks.

• Software Security: Using security-related software practices for building, testing, and correcting the system, including designing and coding software to prevent common vulnerabilities such as buffer overflow and injection attacks, vulnerability testing, and applying updates and fixes to address vulnerabilities.

• Multilevel Security: Implementing multilevel security (MLS) is a way to protect information in computer systems that contain data with different levels of sensitivity or classifications. Multilevel security offers a higher level of protection than using one level of security for all information, as it allows organizations to adapt their security measures to the specific sensitivity of the protected information.

Next, security systems in a big data environment are discussed, as this is relevant to the problem addressed in this paper, which combines security and working with big data environments.

Security systems in a big data environment

The importance of data for gaining a competitive advantage is widely recognized, making it critically important to ensure the storage and processing of sensitive data. To achieve a higher level of security, organizations can implement authentication and authorization mechanisms such as LDAP, Active Directory, and Kerberos. Encryption technologies such as SSL/TLS protocols, data-at-rest encryption, and key management can be used to protect sensitive data. It is of great importance to maintain audit logs and ensure compliance with regulatory requirements. Network security measures such as firewall configuration, network segmentation, and VPN can help prevent unauthorized access to Hadoop nodes and secure communication between nodes. In the complex security environment of big data systems, a centralized authentication mechanism is necessary to ensure that only authorized users and services have access to the data. Kerberos, Apache Knox,

and Apache Ranger are solutions that provide a level of security in authenticating and authorizing users. Apache Knox serves as a secure entry point for REST and HTTP interactions with the Hadoop ecosystem [35], while Apache Ranger provides a platform for access control management and audit policies in the Hadoop ecosystem[36]. Kerberos is used to protect the authentication process and encrypt data exchanged between clients and servers [37].

After discussing the basic components and methods, the next step is to consolidate them into a unified solution that covers all aspects of security and ensures the protection of the system at different levels. Such a solution can be an operational security information center.

Operational Security Information Center

Operational Security Information Center play a key role in protecting these environments and ensuring the confidentiality, integrity, and availability of sensitive data. The reason for using and popularizing such centers is mainly due to the need to prevent major cyber incidents and the resulting adoption of centralized security actions in organizations.

By its nature, the Information Security Operations Center (Operational Security Information Center - OSIC) can be defined as a centralized facility responsible for monitoring and managing the security of a given organization. OSIC typically has a team of security analysts and specialists tasked with detecting, analyzing, and responding to security incidents in real-time [38]. OSIC provides mechanisms for collecting, storing, and normalizing any type of data, as well as ensuring a higher level of security. The main function of OSIC is to provide situational awareness and incident response functions for the organization. This includes monitoring network traffic, detecting security incidents, investigating security-related events, and coordinating the incident response team [39].

The goal of OSIC is to provide a collaborative platform for developing a scalable security analysis tool while maintaining additional features for identifying security issues [38]. OSIC is characterized by the use of automation and analysis to detect and respond to security threats. Some of its automated processes include collecting and analyzing log files, sorting and responding to incidents, and collecting threat intelligence (Figure 8) [38].

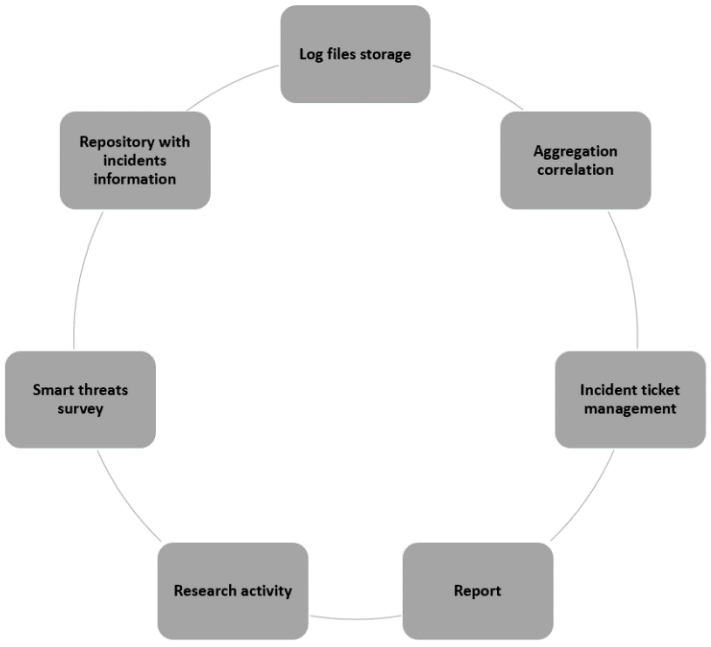


Figure 7. Processes of OSIC.

These types of centers integrate various security technologies and data sources to provide a holistic view of an organization's security status, including integrating security information and event management (SIEM) systems, threat intelligence approaches, and other security tools into one platform. A key feature of these centers is continuous improvement through regular assessments and reviews of security processes and procedures, using metrics and analyses to measure the effectiveness of security operations, identifying areas for improvement, and accordingly adjusting processes and procedures. OSIC monitor compliance with security policies and regulations such as HIPAA, GDPR, and others [38].

OSIC for big data are facing new challenges due to the large volume and complexity of this type of data. There are several existing solutions for managing information security in big data systems, including User and Entity Behavior Analytics (UEBA), Data Loss Prevention (DLP), Identity and Access Management (IAM), Threat Intelligence, and machine learning. UEBA detects anomalies in user behavior and identifies potential security threats, while DLP prevents data leakage, and IAM manages user access to data and systems. Threat Intelligence provides real-time

information on potential security threats and vulnerabilities, and machine learning identifies patterns and anomalies in big datasets to detect security incidents more quickly and accurately.

When designing OSIC for big data, it is best to use a combination of these solutions to ensure comprehensive real-time monitoring and analysis of security. The OSIC team should also be able to interpret the data and react quickly to security-related incidents.

Conclusions

In conclusion, the increasing digitization of various processes has led to the generation of massive amounts of online data, including unstructured data from social media and other platforms. Big data and AI algorithms are used to collect and process this data, which can also detect fraud and improve business operations. However, ensuring the security of this data is a challenge, and various measures such as data encryption and multilayer security must be constantly applied and monitored.

After summarizing the capabilities of big data environments according to selected criteria that are relevant to the dissertation work, we selected products that we will use in designing an operational security information center. Such products include Apache Hadoop and the IDOL system, which can be used for storing and analyzing large amounts of data and for extracting data from social media. These products can be integrated to provide additional functionality, using Apache NiFi.

3. Designing the architecture of an Operational Security Information Center

Digitalization is transforming and revolutionizing business requirements, and the use of big data has become a key element in achieving competitive advantage. In this regard, Operational Security Information Center (OSIC) have become indispensable for businesses as they ensure data security. Specifically, big data systems play a crucial role in cybersecurity and information security by collecting and analyzing large volumes of data to detect potential cyber threats. However, for the effective implementation of OSIC, proper design is essential, and it should be carried out by experienced professionals with clear plans for incident response.

Designing an OSIC requires following established principles and methods. Principles provide guidelines for decision-making, while methods refer to specific techniques and processes aimed at achieving the objectives. It is crucial to note that creating a functional OSIC project requires a comprehensive understanding of the organization's requirements, risks, and goals. Therefore, principles and methods for designing OSIC are presented to establish a framework for understanding the requirements, risks, and goals of the organization.

Principles for designing an Operational Security Information Center

When designing an operational center for big data systems, several principles should be taken into account [40], [41].

First, monitoring is crucial for the functioning of an operational security information center. This principle involves detecting malicious attacks and monitoring malicious activities by employees, subcontractors, guests, and external individuals. Effective monitoring allows for quick identification of threats and facilitates efficient collaboration among security personnel [41].

Second, the analysis of the collected data is crucial for identifying threats as they arise. Real-time analysis is essential for timely response to security threats [40].

Third, incident response is crucial for handling security incidents, whether they are internal or external. Internal incidents may arise from employees, contractors, or partners with access to

the organization's systems, while external incidents may be initiated by attackers outside the organization [41].

Fourth, detecting breaches and responding to them requires a comprehensive incident response plan that outlines the steps to be taken in case of a security incident. The plan should include procedures for containment, investigation, and recovery [41].

Fifth, auditing log files involves analyzing log files from all devices and correlating events across different log files. The OSIC plays a crucial role in logging and auditing by verifying compliance and documenting the response to security incidents [41].

Sixth, regular testing and assessment are critical for identifying potential vulnerabilities and ensuring that existing security measures effectively mitigate risks [41].

Seventh, scalability is important as big data can rapidly grow in size and speed. The OSIC should be designed to scale horizontally by adding more servers or vertically by increasing the processing power of existing servers [40].

Eighth, data confidentiality is crucial to ensure that data is not compromised by internal or external attackers. Methods such as encryption and access control should be implemented to protect data during transmission and at rest [40].

Lastly, automation is crucial for timely identification of security events and prompt response to such events. Automation reduces the risk of manual errors and enhances the tools and processes for security monitoring [40].

These principles serve as a framework for designing a SOC for a big data system that provides comprehensive security. However, in order to fully understand the steps of designing a SOC, various methods need to be examined.

Methods for designing an Operational Security Information Center

Designing an OSIC for big data systems requires careful planning and execution. Here are some methods that are important and useful in designing and implementing an OSIC for big data systems. Their order is based on their significance in ensuring the security and confidentiality of big data systems [42].

- **Risk assessment** - conducting a comprehensive risk assessment is one of the first steps in designing an OSIC for big data systems. This helps identify potential vulnerabilities and risks associated with the data, infrastructure, and applications that will be monitored by the OSIC [42].

- **Security Information and Event Management (SIEM)** - implementing SIEM tools is crucial for collecting, aggregating, and analyzing security event data from various sources in real-time.

- **Access Control** - strict access control is applied to limit the number of people who have access to sensitive information, data, and systems [42].

- **Encryption** - data encryption mechanisms are applied at rest and in transit, which helps protect sensitive data from unauthorized access [43].

- **Network segmentation** - the network is segmented to create separate zones for different types of data and users. This will help prevent unauthorized access and limit potential damages caused by security breaches [42].

- **Compliance** - The OSIC must adhere to all relevant regulations and standards regarding sensitive or personal data, such as the General Data Protection Regulation (GDPR) and the Payment Card Industry Data Security Standard (PCI DSS)[81].

- **Training and awareness raising** - using training programs and awareness campaigns to inform employees and users about the importance of information security and how to identify and report potential security threats. Providing training and awareness about security to all personnel involved in big data system operations helps reduce the risk of security incidents caused by human error[81].

- **Continuous improvement** - implementing a program for continuous improvement to regularly review and update the ISMS to ensure it remains effective and up-to-date [42].

To further enhance these principles and methods by incorporating new technologies, tools, and processes to improve the operation of the ISMS, it is necessary to examine the existing generations of architectures for such centers.

Generations of Operational Security Information Center (OSIC) architectures

The generations of SOC refer to the evolution and maturation of these centers over time. The existing generations are classified into four stages, each with different capabilities, processes, and technologies[43]. These are presented in Figure 8.

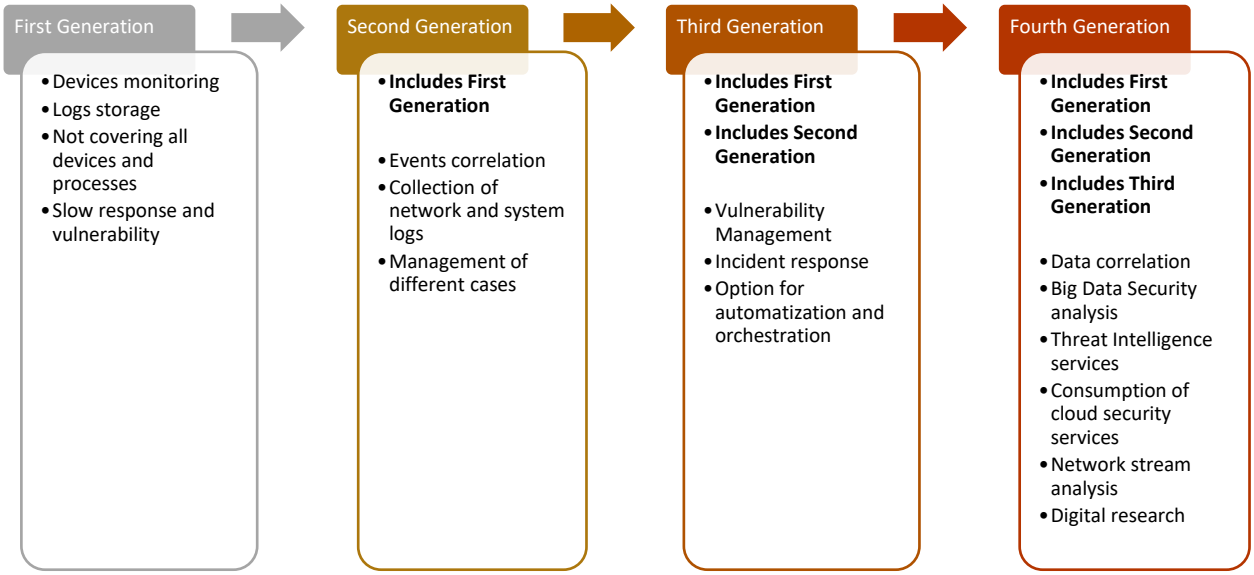


Figure 8. Generations OCIS.

The first generation is characterized by a lack of integration and automation among security tools, while the second generation expands capabilities beyond security alert monitoring to include proactive security operations. The third generation includes advanced automation and orchestration capabilities, and the fourth generation includes integration of advanced analytics and machine learning capabilities.

As each generation of OSIC builds upon the previous one, the work of the security team shifts from collecting log files and incident response to developing new capabilities, processes, and technologies that improve the ability to detect, prevent, and respond to security threats. The evolution of OSICs is a continuous process, as security threats continue to evolve and new challenges arise over time. The development and integration of new technologies, tools, and

processes will continue to be important for enhancing the work of OSIC. These centers face several challenges that need to be addressed for the successful design and construction of the next generation of OSIC.

One of the main challenges is the increasing complexity of systems that need to be managed centrally and automatically, leading to a higher number of events and increased security risks associated with these systems [44]. However, the lack of context in the data generated by security systems can make it difficult to distinguish between genuine threats and false positives. Additionally, existing security information centers can be overwhelmed by the large volume of data, resulting in the generation of numerous alerts and reduced ability to identify real threats [44].

Another challenge is the limited visibility across the entire organization. Existing security information systems often focus on specific areas or functions of the organization, which can result in "blind spots" that hinder the detection and response to security threats that may occur outside the scope of the center [44]. The growth of migration to cloud services and systems also imposes new requirements on security information centers, and the speed at which modern threats can evolve makes current generations vulnerable to new attack techniques [44].

To overcome these challenges, a process has been initiated to design a new fifth generation OSIC. This generation utilizes technologies such as artificial intelligence, machine learning, and automation to enable real-time threat detection and response across the entire organization [45]. One of the main advantages of fifth generation OSIC is their ability to analyze large amounts of data from multiple sources in real-time and adopt a more proactive and strategic approach to security, focusing on threat hunting and prevention rather than just incident response [45]. The fifth generation is expected to focus on cloud security, addressing the unique challenges posed by the transition to cloud-based architectures and the increasing use of containers and microservices.

Some vendors offer advanced solutions for OSIC that incorporate AI for analytics and more automated processes, calling them "fifth generation" OSIC. Such companies include IBM, Atos, ProCyb, and others [46], [47], [48]. However, currently there is no universally accepted definition of what exactly constitutes a fifth generation OSIC.

As threats and technologies constantly evolve, it is of critical importance to continuously assess and update the design principles of an OSIC. New threats and vulnerabilities are constantly

being discovered, and attackers always find new ways to exploit them. At the same time, new technologies are being developed that can help organizations better detect and respond to security incidents in a timely manner. Based on the analysis of existing principles and methods, as well as the information presented about generations and the different challenges they face, several current principles can facilitate the design of a OSIC and lead to the creation of a methodology that shapes the steps before creating a new type of security operations center.

Current principles for designing a security operations center (SOC)

Due to constantly evolving threats and technologies, changing business requirements, increased data analysis, and the growing popularity of cloud computing, it is necessary to update the principles for designing an OSIC.

As a result, updated principles for designing an OSIC are proposed to address security needs and adapt to changing technologies and business requirements. These principles include:

7. **Fully automated incident response and event analysis, including the use of AI**– refers to the automation of incident response processes using artificial intelligence and machine learning technologies. This includes automatic analysis of events and log files generated by various systems and devices, which can help identify potential security incidents and anomalies;

8. **Integration** – The SOC must be integrated with other security systems, such as firewalls, intrusion detection systems, and antivirus software, to ensure a holistic security posture. It should be designed in a way that facilitates data sharing between systems to enhance capabilities for threat detection and response;

9. **Use of AI for analysis of structured and unstructured event data-** involves the use of machine learning algorithms to analyze both structured and unstructured event data. This can help identify patterns and trends that may be missed by traditional methods and can provide a more accurate and timely understanding of potential security threats;

10. **Analysis of unstructured data content -** processing of near real-time and historical data from unstructured sources such as emails, social media, and documents to generate insights for analysis, notification, and report creation;

11. **Prioritization of security alerts based on their severity and impact**. This ensures that critical alerts receive immediate attention and resources are allocated accordingly.

12. **Collaboration and information sharing** include communication capabilities between the SOC and other stakeholders, such as IT teams, business units, and external partners.

Applying these design principles to a OSIC would contribute to organizations remaining competitive by improving their ability to quickly and effectively detect and respond to security incidents.

Designing a SOC for big data systems can be a complex and challenging process, but it is essential to ensure that sensitive data is protected, and security incidents are detected and responded to in a timely manner. For this reason, it is important to consider the various controls in the ISO 27001 standard [49] that are relevant to defining the different functional levels of the new generation SOC.

Using security controls in the design of an OSIC

ISO 27001 can be used as a reference framework for building and managing an OSIC as it provides a comprehensive set of controls that organizations can implement to protect their information assets [50]. This standard includes a total of 114 security controls in 14 areas, including information security policies, organization of information security, asset management, access control, cryptography, physical security and environmental security, and incident management, designed to help organizations manage and mitigate risks to their information security. The standard is flexible and can be adapted, meaning that not all controls need to be implemented, but only those that align with the organization's objectives and assessed risks [49], [50].

Here are some of the ISO 27001 controls that are suitable for implementation in an OSIC:

- **Information security policies (A.5)**- requires organizations to establish and maintain policies for managing information security. OSIC teams should be guided by policies that outline their responsibilities and procedures for handling security-related incidents [49];

- **Physical security (A.7)** - requires organizations to implement measures for physical security to protect their information assets, information processing facilities, and other critical resources from physical threats such as theft, damage, or unauthorized access [50];

- **Information classification (A.8)** - requires organizations to classify information assets based on their level of sensitivity and apply appropriate controls accordingly. The teams of the OSIC can use information classification to focus their efforts on the most critical assets and ensure that security controls are commensurate with the level of risk [50];

- **Access control (A.9)** - requires organizations to restrict access to information and information processing facilities to authorized users. The teams of the OSIC should have access to sensitive data and systems, but this access should be carefully controlled to prevent unauthorized access or data breaches [51];

- **Incident management (A.16)** - requires organizations to establish and maintain a process for detecting, reporting, and responding to information security incidents. The teams of security operations centers are responsible for monitoring security events and incident response, making this control particularly important for the operations of the OSIC [52].

Based on the modern nature of the controls presented in the ISO 27001:2022 standard and their relevance and appropriateness for the field of information security, they will be applicable to the design of OSIC projects. In this context, we propose integrating them into a methodological framework for designing a new type of OSIC. This would contribute to ensuring the confidentiality, integrity, and availability of data processed and stored in the security center, as well as protecting the overall infrastructure of the OSIC against potential security threats.

Creating a method for designing the functional architecture of an OSIC

After a review and analysis of the generations of OSIC, their challenges, and the current development of the next generation OSIC, we believe that there is currently a lack of an integrated approach that facilitates interconnected operation of all security system components, to provide a comprehensive real-time security solution using data from heterogeneous sources. Another reason is the business need for improved security measures in the face of increasing cybersecurity threats. As a result, in this dissertation, we propose a method that incorporates the controls and principles for creating a functional architecture that will enable the construction of a modern generation OSIC (Figure 9).

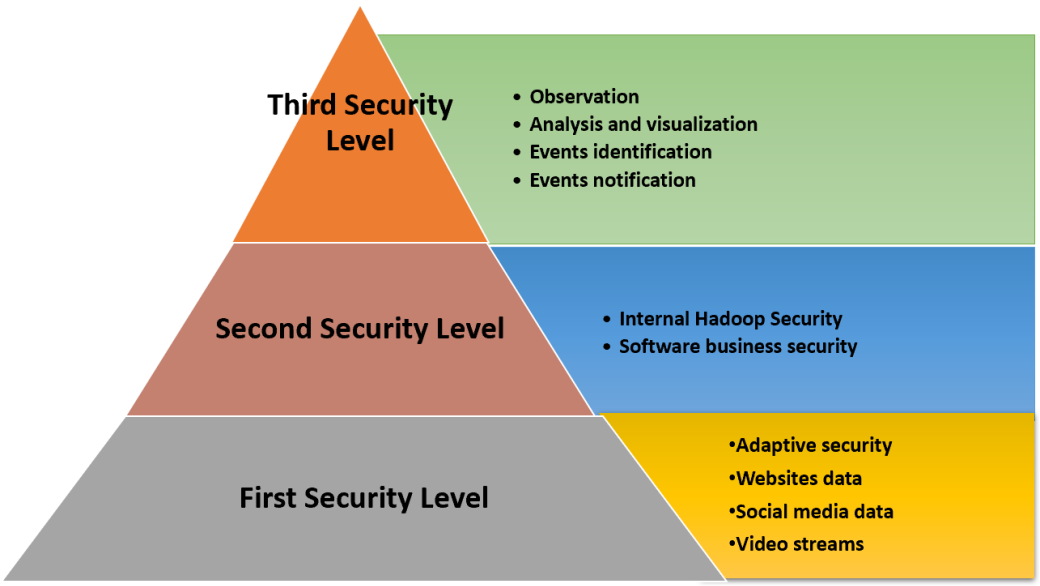


Figure 9. Security Operations Center Functional Architecture.

The main objective of the proposed solution for an OSIC is to address the evolving needs of businesses facing increased levels of cyber threats from various malicious actors. This security center incorporates adaptive security and offers dynamic adjustment of vulnerability measures using artificial intelligence (AI) for event processing that impacts the security level based on existing threats. The proposed method aims to provide protection against cyber-attacks by offering an even higher level of automation, utilizing AI for collecting various types of data from different sources.

The proposed functional architecture consists of three levels, with the first level being intelligent security management. The goal of this level is to detect and respond to security incidents in real-time, utilizing AI and adaptive security. The system collects data from various heterogeneous sources, including video streams, social media data, and other sources. AI, machine, and deep learning methods are used for cognitive search, pattern and knowledge discovery, and data analysis. Data collected from social media and websites can be used for identifying individuals or groups that may pose a security threat, as well as identifying stolen vehicles. Additionally, facial recognition technology can aid in identifying individuals who may pose a security threat.

The second level is based on Hadoop internal security and software business security, designed to protect the confidentiality, integrity, and availability of data. This level includes approaches that utilize centralized user authentication, creation of user access rights to clusters of large data systems, and data segmentation.

The highest third level is the analysis and visualization of the obtained results, aimed at providing appropriate visualization of the collected data from the previous levels and detecting potential threats in real-time. The system monitors events occurring in the center, such as creation and analysis of log files, to prevent and detect potential threats in real-time.

The proposed functional architecture provides a comprehensive set of controls that organizations can implement to protect their information assets. The architecture is flexible and can be adapted to meet the specific needs of any organization, regardless of its size or industry.

On Figure 10, the capabilities of the technologies used at each level are summarized, which help achieve the objectives set for each level.

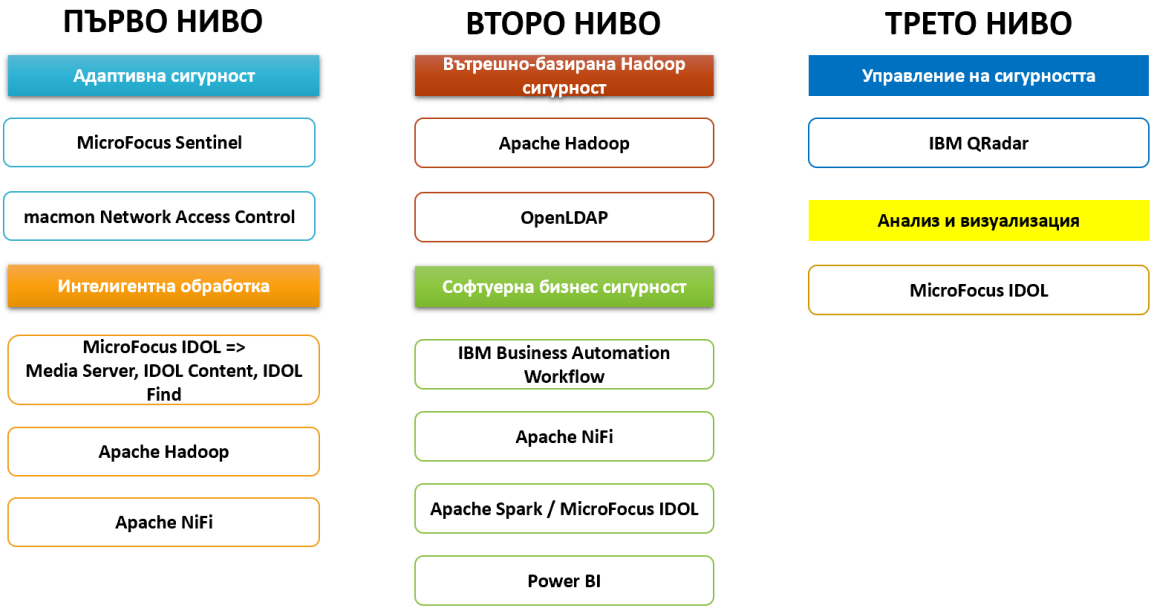


Figure 10. Technologies used at the different levels of OSIC.

Below is a detailed presentation of the levels in the proposed architecture and the technological solutions covering the functionalities that offer a higher level of protection for each of them.

First Level - Adaptive Security at the Operational Security Center

The main challenge in security is the constant threat of attacks. Adaptive Security (AS) is an approach to cybersecurity that continuously analyzes network behavior and events, and is ready to adapt to threats by studying and analyzing them before they occur. An organization can continuously assess risk and ensure appropriate implementation of defensive mechanisms and approaches using adaptive security [53].

We offer a component-based architecture for adaptive security for big data systems, designed to protect the confidentiality, integrity, and availability of data. Figure 11 presents a proposed architecture for Adaptive Security (AS) that can be applied to a big data system.

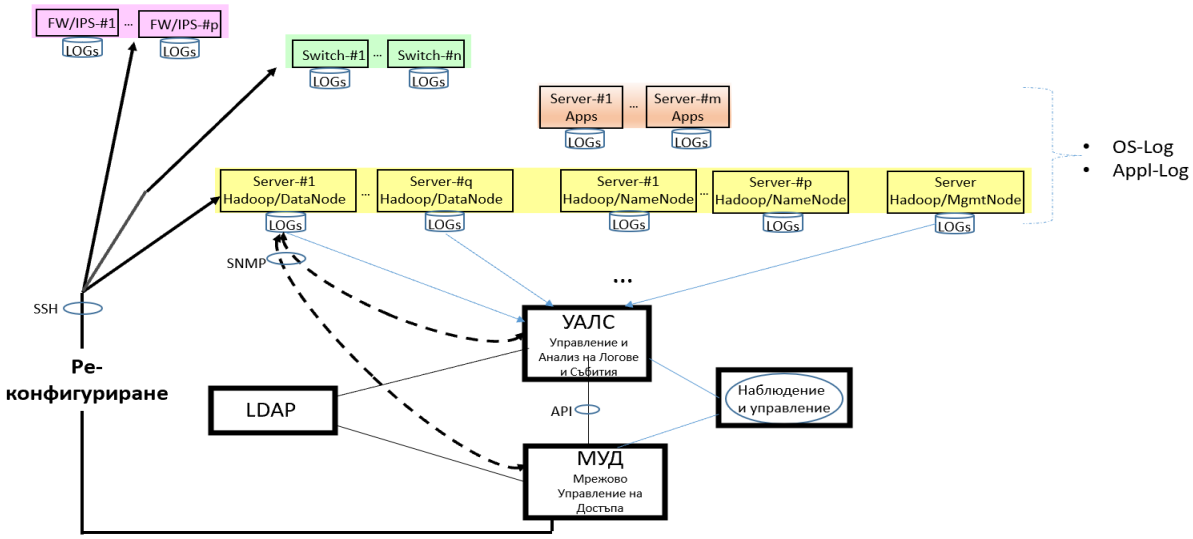


Figure 11. Adaptive Security Components for Big Data Systems.

The proposed architecture includes multiple components such as firewalls, switches, application servers, servers containing components of the Hadoop Distributed File System (HDFS), and device connections. The Hadoop environment has three main components, namely the nameNode, dataNode, and managementNode, which are responsible for managing the namespace of the file system, storing and managing data blocks, and hosting management services, respectively. All components in the system have an option for creating log files, which are monitored by adaptive security systems to ensure real-time network state monitoring and reduce potential security risks. Data replication is used to ensure data availability in case of node compromise. The system also includes a log and event management system (LEMS) and a network access management system (NAMS) for managing and analyzing log files and events and providing network access control, respectively. Overall, this architecture provides a comprehensive and efficient way to protect big data systems[54]. Ensuring secured access and data confidentiality in big data systems is of paramount importance and can be achieved through methods such as multi-factor authentication and data encryption using techniques like AES and 3DES. Regular monitoring and auditing of the system help detect unauthorized activities and ensure compliance with regulations. The use of microservices architecture and implementation of new security features for containerized applications can also enhance security. Network security measures such as firewalls and network segmentation can help prevent attacks. LEMS and NAMS

are components of adaptive security with specific tasks for execution and technological solutions to achieve security goals [55], [56].

The proposed adaptive security solution for the OSIC system includes the use of Micro Focus Sentinel for UALs and macmon Network Access Control for MUD. Sentinel collects and analyzes security-related data from various sources, while macmon provides network access and controls devices and users. Integrating these two systems allows for greater visibility into the network and quick response to security incidents. Data sharing through APIs allows for correlation of events and security alerts, facilitating the identification of potential threats and immediate actions to mitigate them. This approach enables the OSIC team to stay informed about the current state of the system and take timely measures to ensure security [57], [58].

The first level of the operational center for information security - Intelligent data processing

Intelligent data processing refers to a system or approach for security management that utilizes intelligent or automated technologies, where AI plays a key role. This level of security is heavily focused on the collection and analysis of data from various sources in a secure manner [59].

Automating data analysis using machine learning algorithms and other AI techniques facilitates the detection and response to potential threats by identifying patterns in large datasets. AI is used for monitoring social media and other websites for suspicious activities, as well as detecting unusual behaviors from cameras positioned inside or outside facilities. This technology can identify potential threats such as the spread of false information or planning of criminal activities, unauthorized access, or suspicious activities around restricted access areas (Figure 12).

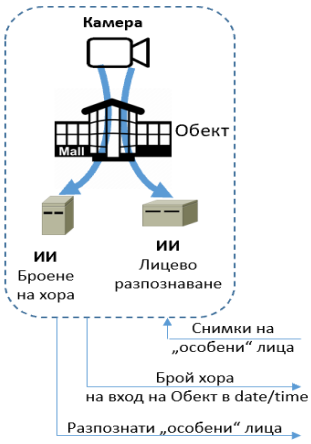


Figure 12. Intelligent primary security with cameras.

Figure 12 illustrates the use of video streams for detecting unusual behaviors, such as unauthorized access or suspicious activities around restricted access areas. By employing advanced AI techniques, algorithms can be trained to count the number of people entering and exiting a specific location in real-time, providing valuable information to security teams about human flow and occupancy levels. Object detection and tracking techniques, including deep neural networks, can be used to avoid double counting and artificial inflation of object numbers within the frame. AI can also be trained to recognize specific individuals based on "enrolled" face images, enabling rapid identification and response to potential threats. This technology can be particularly useful in high-security environments such as airports, railway stations, and stadiums.

The recognition of vehicle license plates from images or video streams involves the use of algorithms that utilize optical character recognition (OCR) to identify and read the license plate numbers - Figure 13.

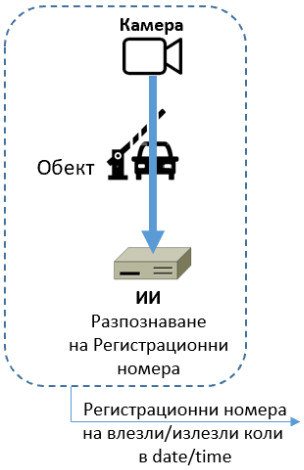


Figure 13. Registration of car numbers at the entrance of a site.

OCR technology has various applications such as traffic control, parking management, and law enforcement.

On the other hand, cognitive analytics can be used to analyze unstructured data such as text, images, and video clips from websites and blogs [25]. This allows the system to understand the context of the website's content, detect negative sentiments or tones, and identify visual elements that contribute to the user experience. Web scraping or crawling tools such as BeautifulSoup, Scrapy, and Selenium can be used to extract website content, which is then indexed in a big data system. Another approach to retrieving web content is through the use of a Content Management System (CMS), which separates content creation and management from the presentation layer [106]. CMS can provide an API for Hadoop access to content stored in CMS, or data can be transferred from CMS to HDFS using a custom script or NiFi. Analyzing website data with AI allows businesses to gain valuable insights into user behavior, market trends, and other key business indicators. AI-driven search engines can recognize and analyze a wide variety of data types, including structured, unstructured, and semi-structured data, providing a more comprehensive understanding of the data [60].

Second level of operational center for information security - Internal-based security for big data system

Internal-based security for big data systems is an approach to data security that emphasizes the inclusion of security measures directly within the big data infrastructure. This approach aims

to make it difficult for malicious actors to gain unauthorized access or compromise the system. Traditional data security has relied on perimeter-based security measures such as firewalls and intrusion detection systems. However, internal-based security focuses on integrating security functions within the framework of the big data infrastructure.

The proposed method for internal-based security for big data systems utilizes the Hadoop big data framework, which includes several built-in security features. These features are designed to address key areas of security concerns, such as authentication and authorization, encryption, and auditing.

Centralized user authentication is a key security approach for Hadoop clusters. This can be achieved through LDAP and/or Kerberos mechanisms to ensure proper authentication and permission control for multiple users accessing the cluster [61]. One-time passwords provide an additional layer of security by requiring a unique password that can only be used once and is valid for a limited period of time. MFA (Multi-Factor Authentication) can be implemented with LDAP to require two or more authentication factors to verify the user's identity. For access control to DataNode servers based on user accounts and groups, System Security Services Daemon (SSSD) can be implemented. SSSD is an identity and authentication provider that allows integration of different authentication mechanisms and maintains data segmentation for access control to DataNode servers[62], [63]. Access control to HDFS directories and files can be managed through assigning permissions to users and groups using the Unix-style permission model and the "setfacl" command from Hadoop CLI [64], [21].

Another approach to ensuring security is by creating a centralized access control list (ACL) for Hadoop, which is managed through the HDFS Command-Line Interface (CLI) or the Graphic User Interface (GUI) [65]. This mechanism allows the administrator to add or remove users or groups from the ACL and modify access rules for specific files or directories. In addition, implementing a comprehensive auditing mechanism is crucial for ensuring traceability of data access and user activity, which helps prevent data loss and maintain regulatory compliance [66]. Data lineage tracking involves recording metadata about the origin, history, and transformation of data in the infrastructure, while user activity tracking involves recording metadata about actions taken by users within the infrastructure. Lastly, data-at-rest and data-in-motion protection through encryption are critical components of a secure and well-managed Hadoop cluster, which can be

achieved through various encryption methods and TLS protocol[67]. These approaches help maintain the security of the Hadoop cluster and protect the valuable data that organizations rely on [68].

Second level of operational center for information security - Software business security

Business software security (BSS) is crucial for protecting sensitive data collected by businesses from cyber threats such as malware, hacking, and phishing. BSS includes access control, data encryption, secure coding practices, vulnerability assessments, penetration testing, security monitoring, and incident response. Statistical methods can be combined with other security measures for data analysis and anomaly detection, trends, and patterns that may indicate security threats. Correlation analysis is used to identify potential security threats and take appropriate mitigation measures[69]. Additionally, Business Process Management (BPM) is used for identifying potential security risks and vulnerabilities in business processes and designing and implementing security measures to mitigate those risks[70]. By employing these methods, organizations can develop a comprehensive security strategy that addresses a wide range of potential threats and vulnerabilities.

Integration of BSS with the Lambda architecture ensures that the data being processed is protected. The Lambda architecture is an approach to data processing designed for handling large volumes of data by combining batch processing and real-time data processing methods. However, the use of the Lambda architecture can result in inconsistencies in the results generated by the batch layer and the real-time processing layer. To overcome these limitations, an extended Lambda architecture is used, which introduces additional layers such as the data ingestion layer, data storage layer, and data processing layer [71], [72] - (Figure 14). These layers help improve the performance, scalability, and consistency of the system.

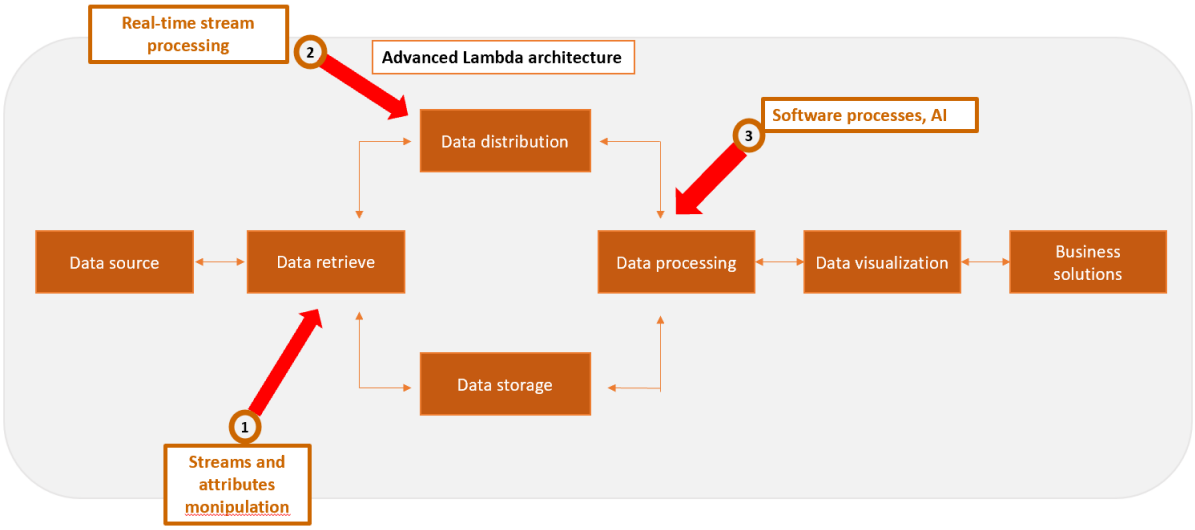


Figure 14. Extended Lambda architecture.

The extended Lambda architecture securely integrates the security of the software business into its workflow, including defining data sources and extracting data from them using a data ingestion layer. Then, the data is stored directly in a distributed storage system or distributed for processing. After data preparation, batch processing, stream processing, or AI algorithm processing is performed, followed by data analysis using big data analytics tools to provide informed business insights. Visualization of results is done using big data visualization platforms or enterprise resource planning (ERP) software. Business process management (BPM) tools such as IBM Business Automation Workflow (BAW) are used for managing and optimizing data processing workflows. Various applications such as NiFi, IDOL, and Spark are used for data processing, analysis, and visualization at different stages. NiFi is used for data ingestion from various sources, Spark is used for near real-time stream data processing, and MLlib is used for machine learning model training and real-time prediction and classification. IDOL is used for advanced information extraction from unstructured data.

The visualization of data is an important aspect of data analysis, allowing decision-makers to identify trends, patterns, and anomalies, and predict future trends. Tools like PowerBI, Tableau, and Plotly offer adaptive and interactive visualizations that can be tailored to specific business needs. Additionally, IDOL has built-in data processing and knowledge extraction capabilities,

including the ability to create charts and diagrams. By leveraging the extended Lambda architecture, companies can benefit from a holistic approach to data processing and security, enabling real-time monitoring, detection, and response to security threats, while ensuring data consistency and accuracy. The architecture can also integrate statistical methods and artificial intelligence to enhance data analysis capabilities and maintain data security.

Third level of OSIC for Information Security

The third level of OSIC includes security management and data visualization from various sources such as log files, video streams, and social media. Effective analysis and visualization of this data can help implement comprehensive security measures and vulnerability prevention strategies. SIEM systems such as QRadar and MicroFocus IDOL can be integrated into SOC to provide real-time monitoring, analysis, and incident response capabilities. QRadar uses correlation mechanisms to detect patterns and anomalies in security data, while IDOL is used for analyzing and visualizing unstructured data from various sources. By combining QRadar with IDOL, organizations can gain insights into their security posture and identify potential threats that may have been missed by other security tools. This integration provides advanced visualization capabilities and enables businesses to explore security data in new ways [73].

Conclusions

In this chapter, the importance of using business security centers was emphasized, presenting the four generations of Security Information and Event Management (SIEM) systems and the challenges they face before the emergence of the next generation. We derived current design principles for SIEM systems and focused on the security management controls provided by the international standard ISO 27001, which are suitable for such security centers. Based on the analysis conducted, we proposed a method for designing a functional architecture for a new type of big data security operations center. The proposed architecture includes three levels of security management, including adaptive AI-driven security, intelligent data processing, internal security, and holistic system analysis. Technological solutions are offered at each level to be implemented in a working model of the proposed security operations center. The functional architecture offers a multi-layered approach to security management, providing comprehensive protection for big data systems by integrating technologies to achieve the set goals.

4. Application of the operational security center method with architectural solutions

The proposed functional architecture offers a multilayered approach to security management, providing comprehensive protection for large data systems by integrating technologies to achieve goals. By applying the proposed method for designing an OSIC (Operational Security Information Center) with the presented solutions, organizations can improve their information security. Some of the technological solutions presented in chapter three have already been implemented by various companies and are not discussed in this chapter. Some of them will be given greater attention due to specificities in their application, demonstrating that they are feasible for the proposed Operational Security Information Center (OSIC).

Application of adaptive security

The adaptive security approach is crucial for managing the lowest level of the proposed security operations center, which is the network level. Two additional security functions included in the proposed architecture are log and event management (LEM) and network access control (NAC), which provide detailed control over network access and constant monitoring of logs from connected devices. LEM works by collecting log files from all low-level components, and NAC manages access to them. Micro Focus Sentinel and macmon Network Access Control are two solutions that can be used to monitor network traffic and provide administrators with timely information about system status. Enabling these security solutions can improve the organization's security by building a comprehensive and multilayered approach to security.

Figure 15 shows the monitoring dashboard from Sentinel, which displays the network load of the monitored devices in the proposed OSIC.

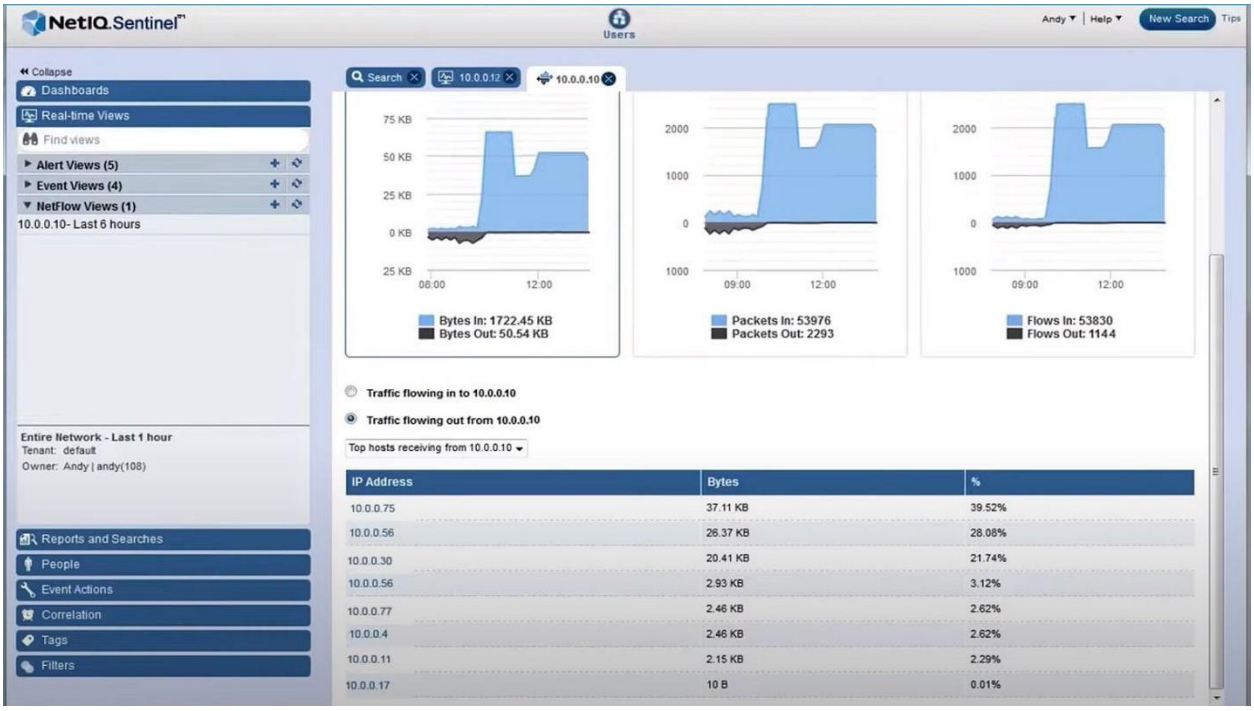


Figure 15. Dashboard for constant monitoring of network traffic in OSIC.

Using both systems, OSIC administrators can take timely action and be informed about the status of the system.

Applying solutions for intelligent data processing

After considering a proposal for network and component protection in the big data environment, the next step is to address the protection during data extraction and processing. To achieve this functionality, the Micro Focus system - IDOL (described in detail in Chapter 2) is used.

Due to the need for data extraction from heterogeneous sources, methods of data extraction, storage, and processing from video streams have been examined, including actions such as people counting, facial recognition, license plate recognition, as well as data extraction from web content and social media.

The proposed architecture for data extraction from a video stream includes recording or capturing the video, followed by data processing using AI algorithms on the IDOL Media Server. The latter analyzes the video stream to detect faces and objects and can convert speech to text for audio tracking. The AI algorithms in the Media Server can also count people in the video stream, achieving one of the goals of the proposed architecture. The output for people counting contains data such as the number of people per frame and cumulative count, as well as determining the duration, start time, and end time of the video stream. Figure 16 shows real-time visualization of the video stream and output generated by the Media Server.

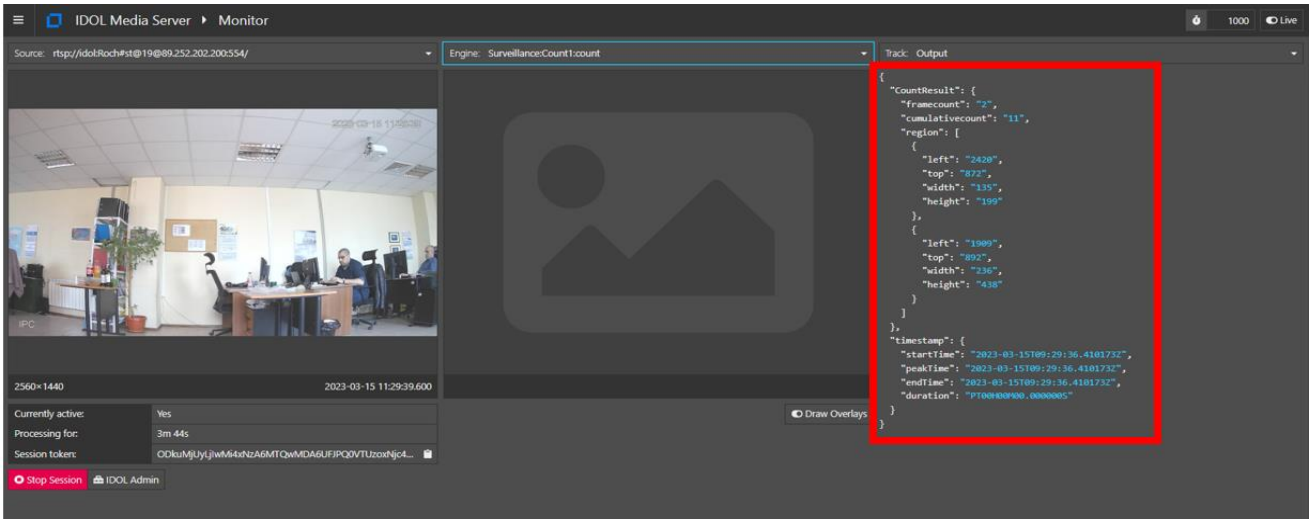


Figure 16. People counting in a frame from a video stream.

Media Server divides the video frame into different areas for better object detection and tracking efficiency. The use of areas in the frame contributes to focusing on individual parts of the frame where important events or activities are expected to occur and ignoring other areas where little or no activity is expected. In the case of Figure 17, there are two areas of division - where there is a gathering of people and where there is not.

IDOL Media Server is a technology that can analyze video streams to detect faces and objects using AI algorithms. One of the goals of the proposed architecture is to count people both inside and outside the facility. Media Server is trained to count people from a video stream using AI algorithms. The facial recognition function involves creating a mathematical representation of a person's face that captures key features of the face. The Media Server compares the extracted features from the video stream with a database of characteristics of known or specified faces, using a matching algorithm to assign a similarity score - Figure 17.

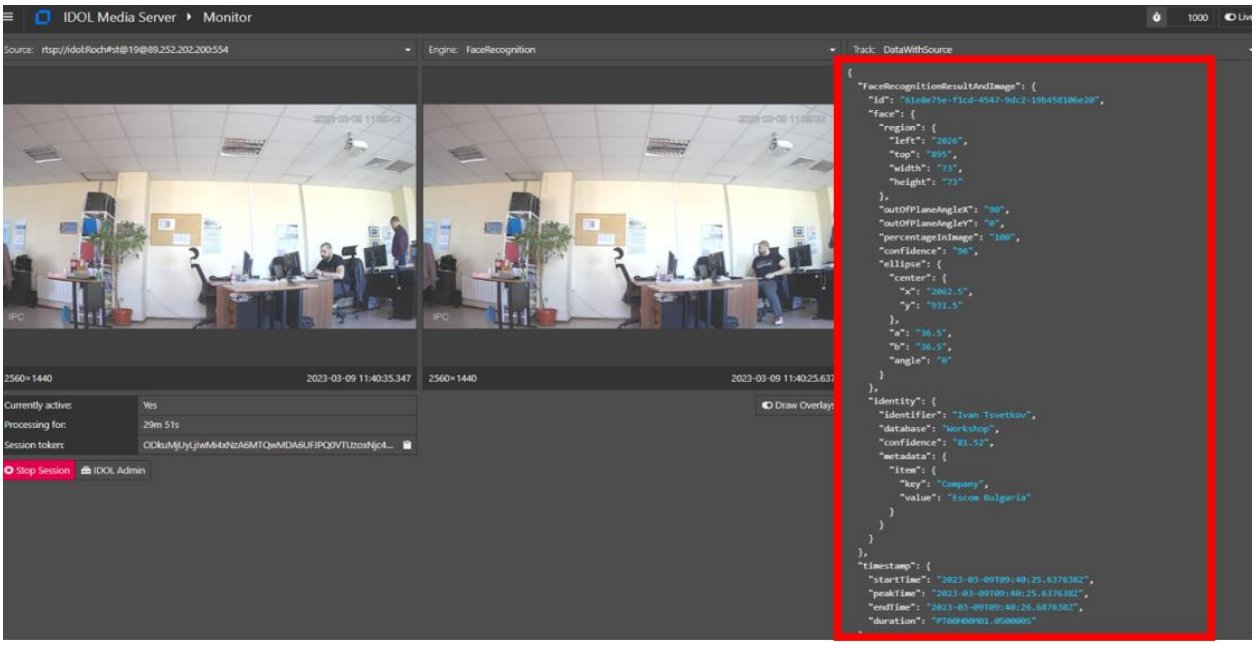


Figure 17. Media Server facial recognition.

The output generated by the software provides valuable information such as the location of the detected faces in the video frames, the name of the recognized individual, the base from which the face was detected, and the percentage of confidence that the detected face is the one it matches. This information is used to determine whether the detection results are reliable enough to take further action or alert security officials.

IDOL Media Server analyzes a video stream to detect and locate license plates and then uses OCR algorithms to recognize the characters - Figure 18.

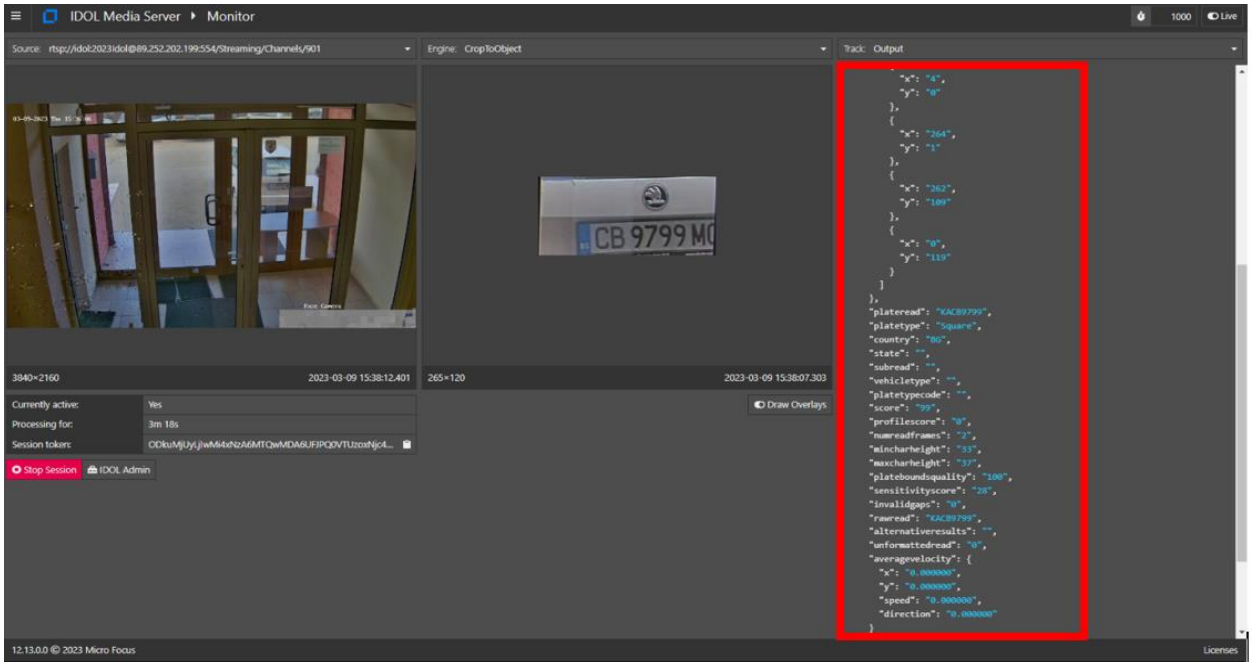


Figure 18. License plate recognition with Media Server.

The software compares the license plate number with a database of known numbers using a matching algorithm. The generated output includes information about the recognized license plate number, its attributes, and the level of accuracy of the recognition results. This output also includes additional information about the recognized license plate, such as the country or state of registration, make and model of the vehicle, and the date and time of recognition. Once IDOL Media Server recognizes a face or license plate number, it can analyze metadata associated with that object, such as the time and location when it was detected. This information can be used for security, surveillance, traffic control, or customer analysis purposes.

In recent years, data generated on social media has become a valuable resource for businesses. They identify customer behavior and conduct analysis to improve business quality.

At this level, we use Micro Focus IDOL, Apache Hadoop, and Apache NiFi products, and the chosen social media platforms are Facebook and Twitter.

The process of data extraction, storage, and processing is shown in Figure 19.

Figure 19. Processing of unstructured social media data.

The process of processing unstructured social data involves installing and configuring the IDOL server and its components, creating a database to store the extracted data, defining access roles, and using built-in AI functionalities. Then, Apache NiFi is installed and configured to extract and store data. The next installation is Hadoop, which is configured for data storage. Connectors are installed and configured to extract data from social media platforms such as Twitter and Facebook by creating applications in the application developer environment and using client keys and tokens to connect to the API. Then, the NiFi system is used to define social media connectors to extract desired content. Figure 20 demonstrates this process.

Figure 20. Connector configuration with provided keys in a built social media application.

Connectors offer the ability to filter data by keywords or phrases that are contained in social media, as shown in Figure 21.

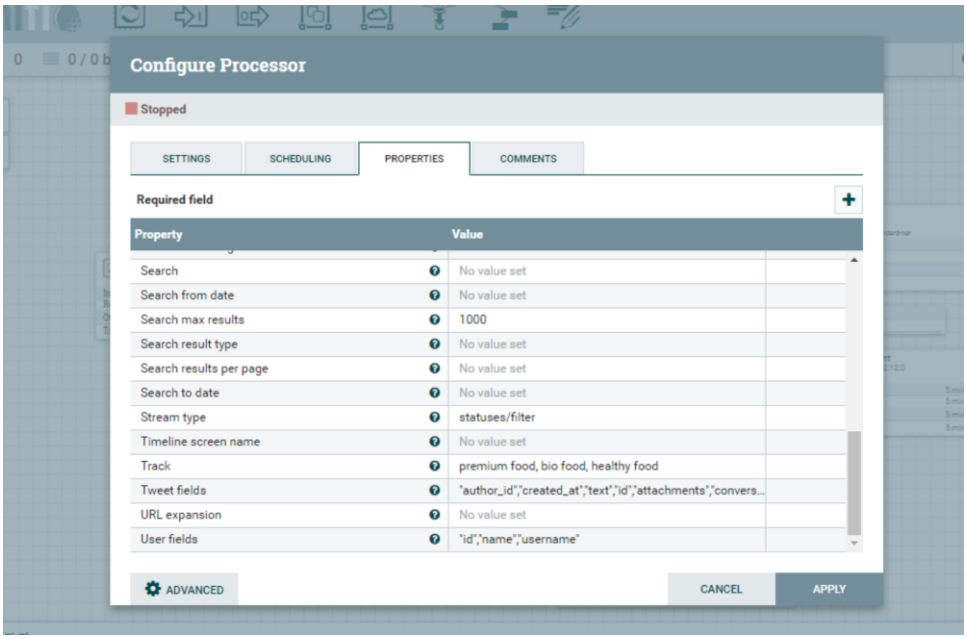


Figure 21. Configure fields to extract data from social media.

In building the process for processing unstructured social media data in a NiFi environment (Figure 22), it is necessary to create processors.

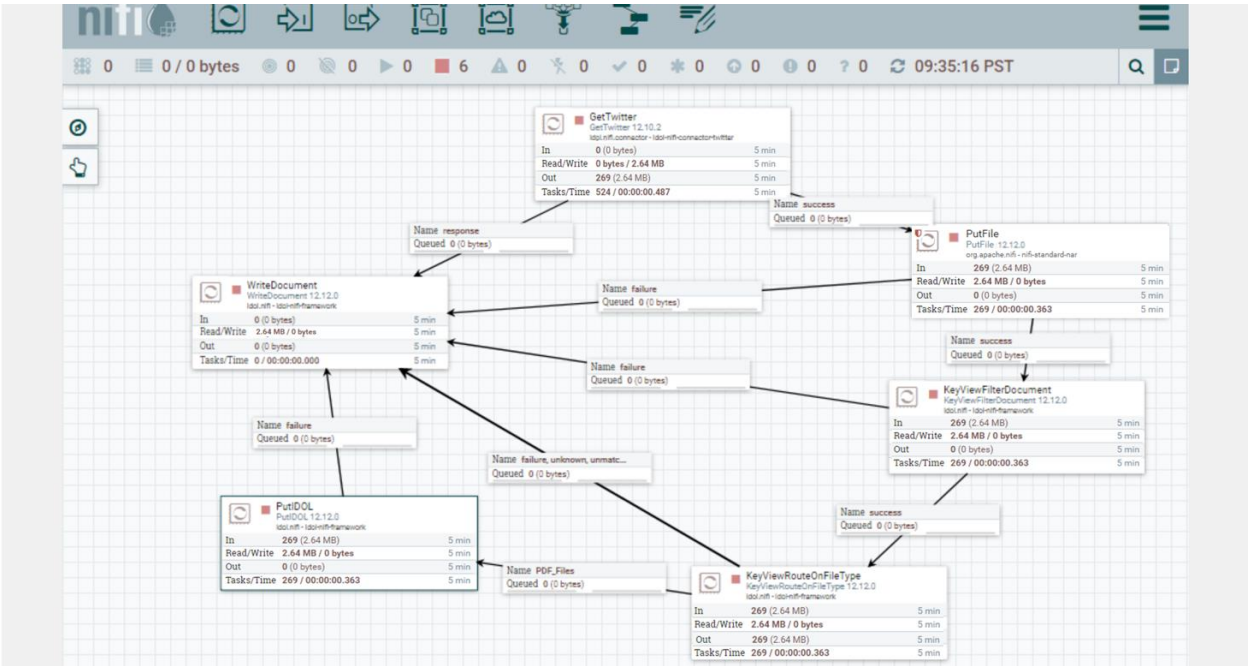


Figure 22. The process of retrieving and storing data from Twitter.

This is done after we select the option to select the type of IDOL component to which it will be assigned. Next, we configure the connector and confirm that the connection between the components is working, and data can be downloaded.

Once the process is complete and the data is in IDOL, further processing and analysis of the data follows for further search, filtering or sorting requests by IDOL Find, which uses AI in this processing (Figure 23).

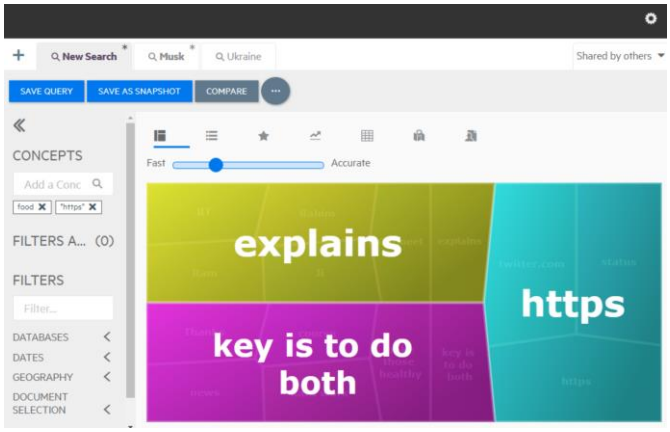


Figure 23. Loading data into IDOL Find.

IDOL Find uses preprocessing techniques such as tokenization and normalization to transform raw data into a format that can be analyzed. This includes processes such as splitting the data into smaller units, such as words or phrases, and removing noise and irrelevant information. For data that has visual content, we can visualize it (Figure 24) before it is processed or stored in the Hadoop repository, where it can be extracted and used for subsequent analysis at any time.

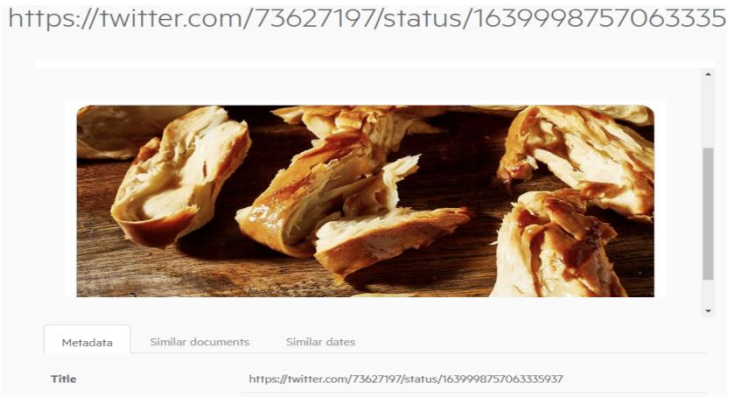


Figure 24. Data visualization in IDOL Find.

Many companies rely on data to make informed decisions. Researchers in various fields collect data from websites to conduct studies, analyze trends, and gain insights into specific topics.

We use the IDOL server and IDOL Content platform to index specific web content. The process involves extracting content using Scrapy and generating an XML file, which is checked for compatibility with IDOL. The file contains fields such as DRREFERENCE, DRETITLE, DRECONTENT, DREDBNAME, DREINDEX, DREFILENAME, and DREFIELD. After indexing, the data is ready for processing and analysis in IDOL Find, where it can be searched. By analyzing web data with AI, companies can gain valuable information about user behavior and market trends, which can inform decision-making. The Extended Lambda architecture, using NiFi and Spark, provides second-level security in addition to Hadoop's internal security.

Applying second-level security to the functional architecture

To ensure the security of a big data system, it is necessary to provide security functionality from the very beginning - at the entry point of the system. LDAP and Kerberos are two solutions that can achieve this. In this context, LDAP is considered as a specific application for Hadoop. The first step is to install and configure the OpenLDAP LDAP server. This includes creating an LDAP directory, defining users and groups, and configuring authentication and authorization rules. The OpenLDAP system uses an LDAP data interchange format (LDIF) file, which is created with a text editor. Users and groups can be added to the OpenLDAP database using the ldapadd command.

```
Dn: cn=ivelkova,ou=users,dc=example,dc=com
objectClass: top
objectClass: person
objectClass: organizationalPerson
objectClass: inetOrgPerson
cn: ivelkova
sn: Velkova
givenName: Ivona Velkova
```

```
userPassword: {SHA}nU4fzW93lj8q3I3TQDwBsjJREO
mail: ivelkova @example.com
```

Authentication and authorization policies can be configured using the OpenLDAP Access Control Language (ACL).

```
access to * by dn="cn=admin,dc=example,dc=com" write by * read
```

LDAP authentication for Hadoop can be enabled by editing the system core-site.xml configuration file of Hadoop. After securing access to the system, attention should be paid to the software business security at this level. Software business security refers to the protection of business activities, infrastructure, and data. Software business security tools are suitable for comprehensive data processing from various sources. This is important because it is considered to provide higher business awareness of potential threats in the online space and processing them in near real-time would help to provide a contemporary response to them. Here, we use the NiFi and IDOL tools, with the latter also used for result visualization. Processes related to data processing for extraction from sources, storage in repositories, processing for information extraction, and analysis of the information that can help in decision-making are included. By using these tools in an extended Lambda architecture and secure system access, organizations can build stable and scalable data processing channels that can handle large volumes of data and support real-time data processing and analysis.

Third-level security management of the Operational Security Information Center architecture

Security information and event management (SIEM) systems are used for third-level security management to aggregate and consolidate data from multiple sources, including networks, security, servers, and databases. IBM Qradar is such a system that uses advanced algorithms for analysis and machine learning to detect and prioritize security threats in real-time. Micro Focus Sentinel collects data from various sources and normalizes them to a common format, which makes analysis and correlation easier. When Qradar identifies a potential threat, it generates a warning that can be personalized based on severity and type. Micro Focus IDOL is used for analysis and

visualization of data collected from log files. IDOL uses text analysis and natural language processing techniques to extract suitable information from log files, and its set of tools includes management dashboards, diagrams, and graphs for exploring and analyzing log data. When potential issues or anomalies are detected, Qradar issues a notification for the potential threat. Together, Qradar and IDOL enable businesses to analyze and visualize security data in real-time to identify potential security threats.

Verification of the defined method using architectural solutions for notification of a potential threat

The proposed Information Security Operations Center uses a variety of software tools to monitor and protect networks and data. A case study is presented to demonstrate the effectiveness of the system. The first process involves retrieving a post from a social media platform, analyzing the data, and storing the results in the Hadoop repository – Figure 25.

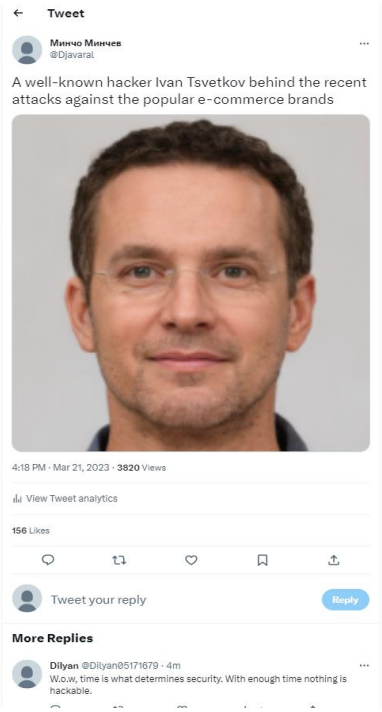


Figure 25. A social media post about a known hacker.

The second process involves extracting data from a video stream from a camera, performing analysis from the Media Server and face recognition - Figure 26, Figure 27.

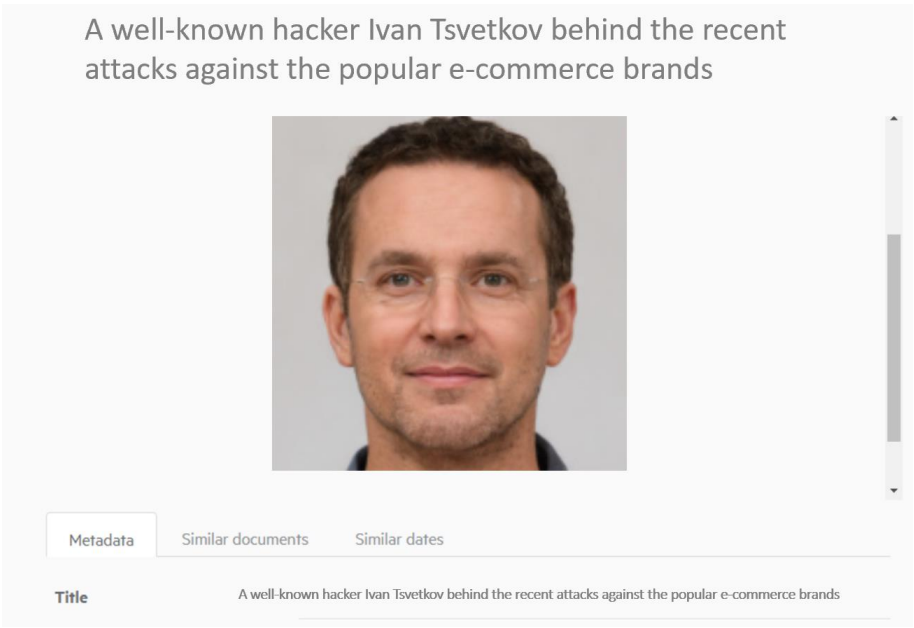


Figure 26. Visualize the retrieved social media post.

Figure 27. Source code from the Media Server with data of the corresponding recognized entity.

The media server then matches the identified person with the data stored in Hadoop, and if there is a match, QRadar generates an SMS alert to the OSIC team. Working together, these software tools provide organizations with a more efficient and effective way to monitor and protect their networks and data – Figure 28.

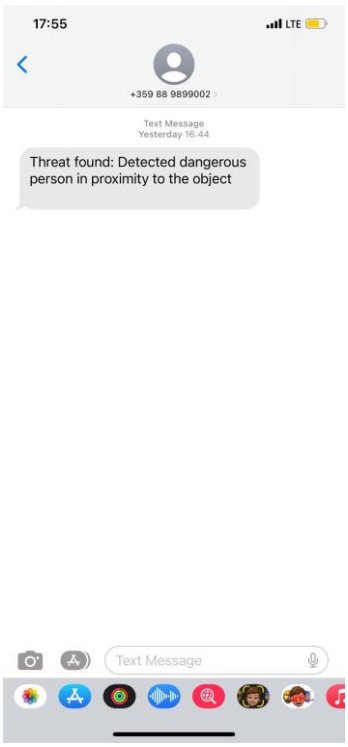


Figure 28. SMS alarm for potential danger in the vicinity of an object.

The collaborative use of these software tools can lead to a broader and more effective solution to IT and cybersecurity problems, which in turn allows organizations to better monitor, analyze, and protect their networks and data.

Conclusion

An assessment has been made of the proposed architecture for a Security Information and Event Management (SIEM) system in a big data environment, demonstrating the advantages of software solutions working together. Micro Focus Sentinel and macmon Network Access Control are used for first-level management, gathering protected data from devices and systems, and monitoring security threats. Micro Focus IDOL and NiFi are used for secure extraction and storage of data in a big data environment, while LDAP and OpenLDAP for multi-factor authentication provide secure access to the system. BPM and the extended Lambda architecture are used as one process with NiFi, IDOL, and PowerBI to generate graphics and diagrams for business conclusions. IDOL is used for visualization of log file analysis, and QRadar is used for real-time notifications and log file analysis.

5. Conclusion

Providing information security in big data systems is a complex task that requires effective security operation centers capable of detecting and responding to threats in a timely manner. In this regard, we propose a method for designing an information security operation center consisting of three management levels that utilize principles and methods from the world of information and cyber security, as well as controls from ISO 27001. The first level includes adaptive security and intelligent processing, the second ensures internal security in the Hadoop environment and focuses on software business security, while the third level utilizes a SIEM system for monitoring and analyzing security events in real-time. To validate the proposed method, several suggested tools were used, which led to a higher level of security.

6. List of publications on the topic of the dissertation work

1. Ivona Velkova, Security challenges for big data platforms, 10TH INTERNATIONAL CONFERENCE ON APPLICATION OF INFORMATION AND COMMUNICATION TECHNOLOGY AND STATISTICS IN ECONOMY AND EDUCATION ICAICTSEE – 2020, November 27 – 28th, 2020, University of National and World Economy, Sofia, Bulgaria, ISSN 2367-7635 (PRINT), ISSN 2367-7643 (ONLINE), available at: https://icaictsee.unwe.bg/past-conferences/ICAICTSEE-2020.pdf, pp. 482-488

2. Ivona Velkova, Mariana Kovacheva, Digitalization in Bulgarian Higher Education – Present and Future Opportunities, 4th International Academic Conference on Education, 2021, 10th December 2021, Barcelona, Spain, ISBN: 978-609-485-239-8), available at: https://www.dpublication.com/proceeding/4th-iaceducation/#Table-of-Contents, pp. 1-13

3. Ivona Velkova, Unstructured Data Processing and Analysis Using Artificial Intelligence, Automatics and informatics magazine, ISSN 0861-7562 (Print), ISSN

2683-1279 (Online), Year LV No. 4/2022, available at: https://sai-bg.com/wp-content/uploads/2023/01/AI-4-2022.pdf, pp. 27-30

4. Ivona Velkova, Unstructured social media data processing with artificial intelligence, VIII INTERNATIONAL SCIENTIFIC CONFERENCE HIGH TECHNOLOGIES. BUSINESS.SOCIETY 2023, 06-09.03.2023, BOROVETS, BULGARIA, ISSN 2535-0005(PRINT), ISSN 2535-0013 (ONLINE), available at: http://hightechsociety.eu/sbornik/2023.pdf, p. 45-48

***Pending publication**: Ivona Velkova, Approaches to higher security level for Hadoop environment*

7. Literature

[1] “Total data volume worldwide 2010-2025,” *Statista*. https://www.statista.com/statistics/871513/worldwide-data-created/ (accessed Oct. 09, 2022).

[2] K. Stefanova and D. Kabakchieva, “Challenges and Perspectives of Digital Transformation,” *Conferences of the department Informatics*, no. 1, pp. 13–23, 2019.

[3] G. Sriram, “SECURITY CHALLENGES OF BIG DATA COMPUTING,” p. 8, Jan. 2022.

[4] InfoSec (www.infosec.gov.hk), “InfoSec: Core Security Principles,” *InfoSec*. https://www.infosec.gov.hk/en/knowledge-centre/core-security-principles (accessed Mar. 11, 2023).

[5] “7 SecOps roles and responsibilities for the modern enterprise | TechTarget,” *Security*. https://www.techtarget.com/searchsecurity/feature/7-SecOps-roles-and-responsibilities-for-the-modern-enterprise (accessed Jan. 29, 2022).

[6] D. Hain, R. Jurowetzki, S. Lee, and Y. Zhou, “Machine learning and artificial intelligence for science, technology, innovation mapping and forecasting: Review, synthesis, and applications,” *Scientometrics*, Jan. 2023, doi: 10.1007/s11192-022-04628-8.

[7] “The Global State of Digital in July 2022 — DataReportal – Global Digital Insights.” https://datareportal.com/reports/digital-2022-july-global-statshot (accessed Feb. 15, 2023).

[8] “Number of internet users worldwide 2022 | Statista.” https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/ (accessed Mar. 29, 2023).

[9] B. Scalzo, “Securing Structured Data,” *Stealthbits Technologies*, Apr. 24, 2019. https://stealthbits.com/blog/securing-structured-data/ (accessed Mar. 11, 2023).

[10] P. Lo Giudice, L. Musarella, G. Sofo, and D. Ursino, “An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake,” *Information Sciences*, vol. 478, pp. 606–626, Apr. 2019, doi: 10.1016/j.ins.2018.11.052.

[11] MicroFocus, “Saba Cloud Pro: IDOL01WBT - IDOL Essentials 12.6 Digital Learning for Administrators with Specialist Exam.”

[12] “What Is Unstructured Data?,” *MongoDB*. https://www.mongodb.com/unstructured-data (accessed Mar. 29, 2023).

[13] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, “Uncertainty in big data analytics: survey, opportunities, and challenges,” *Journal of Big Data*, vol. 6, no. 1, p. 44, Jun. 2019, doi: 10.1186/s40537-019-0206-3.

[14] “Internet and social media users in the world 2023,” *Statista*. https://www.statista.com/statistics/617136/digital-population-worldwide/ (accessed Feb. 27, 2023).

[15] “The Latest Facebook Statistics: Everything You Need to Know — DataReportal – Global Digital Insights.” https://datareportal.com/essential-facebook-stats (accessed Mar. 01, 2023).

[16] “10 Google Search Statistics You Need to Know in 2023 | Oberlo,” Jan. 13, 2023. https://www.oberlo.com/blog/google-search-statistics (accessed Mar. 02, 2023).

[17] “The 42 V’s of Big Data and Data Science,” *Elder Research*. https://www.elderresearch.com/blog/the-42-vs-of-big-data-and-data-science/ (accessed Mar. 09, 2023).

[18] К. Стефанова and С. Йорданова, “ПРЕДИЗВИКАТЕЛСТВАТА НА ГОЛЕМИТЕ ДАННИ СЪЩНОСТ, ХАРАКТЕРИСТИКИ И ТЕХНОЛОГИИ.” https://docplayer.bg/166167087-предизвикателствата-на-големите-данни-същност-характеристики-и-технологии.html (accessed Nov. 20, 2022).
[19] Techreviewer, “The Most Popular Big Data Frameworks in 2022 | Techreviewer Blog.” https://techreviewer.co/blog/the-most-popular-big-data-frameworks-in-2022 (accessed Oct. 09, 2022).
[20] W. Inoubli, S. Aridhi, H. Mezni, M. Maddouri, and E. Mephu Nguifo, “An experimental survey on big data frameworks,” *Future Generation Computer Systems*, vol. 86, pp. 546–564, Sep. 2018, doi: 10.1016/j.future.2018.04.032.
[21] Y. Filaly, N. Berros, H. Badri, F. E. mendil, and Y. E. B. EL Idrissi, “Security of Hadoop Framework in Big Data,” in *Artificial Intelligence and Smart Environment*, Y. Farhaoui, A. Rocha, Z. Brahmia, and B. Bhushab, Eds., in Lecture Notes in Networks and Systems. Cham: Springer International Publishing, 2023, pp. 709–715. doi: 10.1007/978-3-031-26254-8_103.
[22] “Apache Hadoop.” https://hadoop.apache.org/ (accessed Mar. 09, 2023).
[23] IBM Cloud Education, “What is Apache Spark?,” Sep. 14, 2022. https://www.ibm.com/cloud/learn/apache-spark (accessed Oct. 09, 2022).
[24] “What is Apache Hive? | AWS.” https://aws.amazon.com/big-data/what-is-hive/ (accessed Mar. 11, 2023).
[25] MicroFocus, “Powered By Idol,” *Micro Focus*. https://content.microfocus.com/idol-analytics-21/powered-by-idol-sdks (accessed Oct. 09, 2022).
[26] N. Millman, “8 considerations when selecting big data technology,” *Computerworld*, Feb. 27, 2014. https://www.computerworld.com/article/2475840/8-considerations-when-selecting-big-data-technology.html (accessed Feb. 24, 2023).
[27] “7 Criteria for Choosing Big Data Developers - N-iX,” *Software Development Company - N-iX*. https://www.n-ix.com/7-criteria-choosing-big-data-developers/ (accessed Sep. 28, 2021).
[28] “Apache NiFi Overview.” https://nifi.apache.org/docs/nifi-docs/html/overview.html (accessed Mar. 11, 2023).
[29] “Supervised vs. Unsupervised Learning: What’s the Difference?,” Nov. 15, 2022. https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning (accessed Mar. 05, 2023).
[30] “What is Artificial Intelligence (AI)? | Definition from TechTarget,” *Enterprise AI*. https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence (accessed Mar. 11, 2023).
[31] “What is Natural Language Processing? | IBM.” https://www.ibm.com/topics/natural-language-processing (accessed Mar. 11, 2023).
[32] N. James, “160 Top Cybersecurity Statistics 2023: Figures, Facts & Trends,” Dec. 19, 2022. https://www.getastra.com/blog/security-audit/cyber-security-statistics/ (accessed Mar. 06, 2023).
[33] “Critical cybersecurity areas worldwide 2023,” *Statista*. https://www.statista.com/statistics/1292944/critical-cybersecurity-area-worldwide/ (accessed Mar. 06, 2023).
[34] “What is Information Security | Policy, Principles & Threats | Imperva,” *Learning Center*. https://www.imperva.com/learn/data-security/information-security-infosec/ (accessed Mar. 29, 2023).

[35] Y. Soo, “Discussion and Comparison of Several Hadoop Security Tools | by Yinyi Soo | Medium.” https://ysoo23.medium.com/discussion-and-comparison-of-several-hadoop-security-tools-b4532a8c67f9 (accessed Feb. 10, 2022).
[36] Cloudera, “Apache Ranger,” *Cloudera*. https://www.cloudera.com/products/open-source/apache-hadoop/apache-ranger.html (accessed Feb. 11, 2022).
[37] A. Javed, “3 Basic A’s of Identity and Access Management -Authentication, Authorization, and Accounting.” https://www.xorlogics.com/2019/04/15/3-basic-as-of-identity-and-access-management-authentication-authorization-and-accounting/ (accessed Feb. 10, 2022).
[38] EC-Council, “Understanding the Role of a Security Operations Center,” *Cybersecurity Exchange*, Apr. 28, 2022. https://www.eccouncil.org/cybersecurity-exchange/security-operation-center/responsibilities-security-operations-center-soc-team/ (accessed Mar. 29, 2023).
[39] M. Vielberth, F. Bohm, I. Fichtinger, and G. Pernul, “Security Operations Center: A Systematic Study and Open Challenges,” *IEEE Access*, vol. 8, pp. 227756–227779, 2020, doi: 10.1109/ACCESS.2020.3045514.
[40] S. MS, “Useful KPIs for a Security Operation Center (SOC),” Dec. 19, 2019. https://www.cloudcybersafe.com/useful-kpi-elements-for-a-security-operation-center-soc/ (accessed Mar. 29, 2023).
[41] LogRhythm, “7 Steps to Building A Security Operations Center (SOC),” *LogRhythm*, Jun. 16, 2020. https://logrhythm.com/blog/7-steps-to-build-your-security-operations-center/ (accessed Jan. 18, 2023).
[42] “How to Build a Security Operations Center (SOC): Peoples, Processes, and Technologies,” *Digital Guardian*. https://www.digitalguardian.com/blog/how-build-security-operations-center-soc-peoples-processes-and-technologies (accessed Mar. 29, 2023).
[43] “The Evolution of Security Operations and Strategies for Building an Effective SOC,” *ISACA*. https://www.isaca.org/resources/isaca-journal/issues/2021/volume-5/the-evolution-of-security-operations-and-strategies-for-building-an-effective-soc (accessed Mar. 29, 2023).
[44] A. T. A. J. Kienzle, “What Is a SOC? Top Security Operations Center Challenges,” *IIoT World*, Jan. 28, 2022. https://www.iiot-world.com/ics-security/cybersecurity/top-challenges-soc-are-facing/ (accessed Mar. 29, 2023).
[45] “5G/SOC: SOC Generations -HP ESP Security Intelligence and Operations Consulting Services - Business white paper”.
[46] E. R. https://www.emergenresearch.com, “SOC as a Service Market Size, Share | Industry Forecast by 2030.” https://www.emergenresearch.com/industry-report/security-operations-center-as-a-service-market (accessed Apr. 17, 2023).
[47] “Atos opens new global next-gen Security Operations Center in Bulgaria and strengthens its sovereign security offering in Europe,” *Atos*, Mar. 22, 2022. https://atos.net/en/2022/press-release_2022_03_22/new-global-security-operations-center-in-bulgaria (accessed Apr. 17, 2022).
[48] “Next Gen CR - PROCYB srl,” Oct. 06, 2021. https://procyb.io/en/next-generation-soc-en/ (accessed Apr. 17, 2023).
[49] “ISO 27001 Annex A.5 - Information Security Policies,” *https://www.isms.online/*. https://www.isms.online/iso-27001/annex-a-5-information-security-policies/ (accessed Apr. 23, 2023).

[50] D. Kosutic, “What are the 11 new security controls in ISO 27001:2022?” https://advisera.com/27001academy/explanation-of-11-new-iso-27001-2022-controls/ (accessed Apr. 23, 2023).
[51] “ISO 27001 Annex A.9 Access Control - Your Step-by-Step Guide,” *https://www.isms.online/*. https://www.isms.online/iso-27001/annex-a-9-access-control/ (accessed Apr. 24, 2023).
[52] “ISO 27001 Annex A.16 - Information Security Incident Management,” *https://www.isms.online/*. https://www.isms.online/iso-27001/annex-a-16-information-security-incident-management/ (accessed Apr. 24, 2023).
[53] “Adaptive architecture: Key to True Cybersecurity | Kaspersky official blog.” https://www.kaspersky.com/blog/asa-key-to-true-cybersecurity/6678/ (accessed Dec. 10, 2022).
[54] “Designing an Adaptive Security Architecture for Protection From Advanced Attacks,” *Gartner*. https://www.gartner.com/en/documents/2665515 (accessed Mar. 28, 2023).
[55] “Set Up Containerize and Test a Single Hadoop Cluster using Docker and Docker compose,” *Engineering Education (EngEd) Program | Section*. https://www.section.io/engineering-education/set-up-containerize-and-test-a-single-hadoop-cluster-using-docker-and-docker-compose/ (accessed Mar. 29, 2023).
[56] K. Miao, J. Li, W. Hong, and M. Chen, “A Microservice-Based Big Data Analysis Platform for Online Educational Applications,” *Scientific Programming*, vol. 2020, p. e6929750, Jun. 2020, doi: 10.1155/2020/6929750.
[57] “Enterprise Business Intelligence | Sentinel.” https://www.microfocus.com/en-us/cyberres/secops/sentinel (accessed Mar. 29, 2023).
[58] “Zero Trust Network Access.” https://www.macmon.eu/en/ (accessed Mar. 29, 2022).
[59] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, “Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda,” *International Journal of Information Management*, vol. 48, pp. 63–71, Oct. 2019, doi: 10.1016/j.ijinfomgt.2019.01.021.
[60] “Big data Hadoop and MapReduce solutions for CMS content management problems,” *TheServerSide.com*. https://www.theserverside.com/tutorial/How-big-data-solved-the-content-management-CMS-problem (accessed Mar. 02, 2023).
[61] KARDEN, “Authentication in Hadoop cluster: MIT Kerberos and Active Directory – DekarLab,” May 23, 2020. https://dekarlab.de/wp/?p=883 (accessed Mar. 29, 2022).
[62] “Set up Okta Verify (MFA) | eSolutions.” https://www.monash.edu/esolutions/phones/change-device-multi-factor-authentication (accessed Mar. 29, 2023).
[63] “Best Practices Guide for Systems Security Services Daemon Configuration and Installation - Part 1 - Cloudera Blog.” https://blog.cloudera.com/best-practices-guide-for-systems-security-services-daemon-configuration-and-installation-part-1/ (accessed Mar. 29, 2022).
[64] A. Luntovskyy and D. Gütter, “From Big Data to Smart Data: Best Practices for Data Analytics,” in *Highly-Distributed Systems: IoT, Robotics, Mobile Apps, Energy Efficiency , Security*, A. Luntovskyy and D. Gütter, Eds., Cham: Springer International Publishing, 2022, pp. 79–96. doi: 10.1007/978-3-030-92829-2_4.
[65] P. P. Sharma, “Securing Big Data Hadoop : A Review of Security Issues , Threats and Solution,” 2014. https://www.semanticscholar.org/paper/Securing-Big-Data-Hadoop-%3A-

A-Review-of-Security-%2C-Sharma/fedad0e2e8a2eb2a81a572679c233ee94b1a2bf8 (accessed Feb. 10, 2022).
[66] "Apache Hadoop Amazon Web Services support – Object Store Auditing." https://hadoop.apache.org/docs/current/hadoop-aws/tools/hadoop-aws/auditing (accessed Feb. 10, 2023).
[67] W. Rowe, "Introduction to Hadoop Security," *BMC Blogs*. https://www.bmc.com/blogs/hadoop-security/ (accessed Feb. 10, 2022).
[68] "Hadoop KMS – Hadoop Key Management Server (KMS) - Documentation Sets." https://hadoop.apache.org/docs/stable/hadoop-kms/index.html (accessed Mar. 29, 2023).
[69] R. Ross, M. McEvilley, and J. Carrier Oren, "Systems Security Engineering: Considerations for a Multidisciplinary Approach in the Engineering of Trustworthy Secure Systems," National Institute of Standards and Technology, NIST SP 800-160, Nov. 2016. doi: 10.6028/NIST.SP.800-160.
[70] "Use BPM Software to Improve Enterprise Security," *abas*. https://abas-erp.com/en/resources/erp-blog/improve-enterprise-security-bpm-software (accessed Mar. 29, 2023).
[71] "Artificial Intelligence: Transforming Business Security Strategies." https://www.securityinformed.com/insights/artificial-intelligence-transforming-business-security-strategies-co-3425-ga.1555480394.html (accessed Mar. 29, 2023).
[72] "What Is Lambda Architecture?" https://www.databricks.com/glossary/lambda-architecture (accessed Mar. 29, 2023).
[73] "Definition of SIEM - IT Glossary | Gartner." https://www.gartner.com/en/information-technology/glossary/security-information-and-event-management-siem (accessed Mar. 29, 2023).

8. List of figures

9. List of tables