Jukka-Pekka Kunnari

KONEOPPIMISEN HYÖDYNTÄMISMAHDOLLISUUDET SIEM-JÄRJESTELMISSÄ

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2022

TIIVISTELMÄ

Kunnari, Jukka-Pekka
Koneoppimisen hyödyntämismahdollisuudet SIEM-järjestelmissä
Jyväskylä: Jyväskylän yliopisto, 2022, 52 s.
Tietojärjestelmätiede, pro gradu –tutkielma
Ohjaajat: Lehto, Martti, Hämäläinen, Timo

System Information ja Event Management, eli SIEM-järjestelmistä on tullut viime vuosina organisaatioiden kyberturvallisuusvalvonnan keskeinen ratkaisu. Järjestelmä kerää ja varastoi loki-, eli tapahtumatietoa organisaation tietojärjestelmästä täyttäen paitsi lainsäädännölliset vaatimukset tapahtumatietojen säilyttämisestä, mutta mahdollistaen myös tietojärjestelmän toiminnan valvonnan ja esimerkiksi haitallisen toiminnan havaitsemisen, koska kyberhyökkääjien yleisesti käyttämistä tekniikoista jää jälkiä järjestelmän lokitietoihin. SIEM-Järjestelmien haasteena kuitenkin on, että tapahtumatietoa kertyy nopeasti hyvin suuria määriä, ja esimerkiksi kyberhyökkäyksen valmistelun merkkien havaitseminen suuresta tietomäärästä on haastavaa. Tässä pro gradu -tutkielmassa tarkastellaan mahdollisena ratkaisuna SIEM-järjestelmän toiminnan tehostamiseksi ja SIEM-järjestelmää hyödyntävien henkilöiden työn helpottamiseksi yhden tekoälyn muodon, koneoppimisen, hyödyntämistä osana järjestelmän toimintaa. Tutkimuksen päätutkimuskysymys oli, miten koneoppimista voidaan hyödyntää SIEM-järjestelmissä.

Tutkimuksessa selvitettiin tunnettuja, SIEM-järjestelmissä hyödynnettyjä koneoppimisratkaisuja sekä konstruktiiviseen (DSRM; design science research methodology) tutkimusmenetelmään perustuen toteutettiin luonnollisen kielen prosessointia hyödyntävä koneoppimistoiminnallisuus, joka integroitiin Splunk Enterprise -sovellukseen perustuvaan SIEM-järjestelmään analysoimaan valvottavan järjestelmän Linux-palvelinten lokitietoja.

Tutkimuksen perusteella koneoppimisen integroimiseen osaksi SIEM-järjestelmää on useita mahdollisia ratkaisuja. Tutkimuksessa toteutetun esimerkkiratkaisun avulla suuri lokimäärä voitiin jakaa niiden tekstisisällön perusteella omiin ryhmiinsä, sekä erottelemaan tapahtumien joukosta muista tapahtumista selvästi poikkeavat tapahtumat reaaliajassa rajaten kyberuhkien havaitsemisen kannalta kiinnostavat tapahtumat pienemmäksi ryhmäksi niiden tarkemman analysoinnin helpottamiseksi.

Koneoppimisen integroiminen Splunkiin on melko yksinkertaista, koska tarvittavat lisäosat on saatavilla sovellukseen. Koneoppimismallin kehittäminen ja optimointi vaativat kuitenkin useita toistoja ja tulosten jatkuvaa validointia sopivien parametrien löytämiseksi. Tulokset kuitenkin osoittavat koneoppimisen hyödyntämispotentiaalin SIEM-järjestelmien tiedonlouhinnassa.

Asiasanat: SIEM-järjestelmä, koneoppiminen, loki, lokienhallinta, kyberuhka

ABSTRACT

Kunnari, Jukka-Pekka
Potential machine learning solutions in SIEM systems
Jyväskylä: University of Jyväskylä, 2022, 52 pp.
Information Systems, Master's Thesis
Supervisors: Lehto, Martti, Hämäläinen, Timo

During last few years, System Information and Event Management systems have become the backbone solution for organizations' cyber situational awareness monitoring. SIEM system collects and stores event or log information from organization's IT infrastructure to meet not only legal requirements of log management, but giving a tool to monitor the IT infrastructure, and to detect possible signs of cyber threats, as most of the techniques and tactics commonly used by adversaries leaves traces in the system logs. However, a common defect in SIEM systems is the massive amount of log data generated in every minute, making it very challenging to detect the signs of potential threats. This master's thesis studies potential machine learning applications in order to enhance the SIEM systems' capabilities, and to make SIEM system more user-friendly. The main research question of this study was "How could machine learning be utilized in SIEM systems?"

In this research, commonly known applications of machine learning were studied, and an example solution based on natural language processing techniques was developed. The function was integrated into Splunk Enterprise SIEM system for log mining from the Linux servers, following the design science research methodology (DSRM) for IT systems research process.

The results show that there are multiple possible solutions to utilize machine learning in SIEM systems. By using the solution proposed in the study, an extensive amount of log data could be divided into own groups and the potentially interesting log data could be separated and categorized for further analysis. Utilizing machine learning in a system like Splunk is relatively uncomplicated, as all the add-on modules are downloadable for all users. On the other hand, developing and optimizing a machine learning model is a long process, requiring multiple iterations and validations to find optimal parameters for the model. The results, however, point out the potential of machine learning, especially for data mining in the SIEM systems.

Keywords: SIEM system, machine learning, log management, cyber threat

KUVIOT

SISÄLLYS

1 JOHDANTO

System Information and Event Management (SIEM) -järjestelmistä on tullut kyberturvallisuusvalvonnan keskeinen työkalu monille organisaatiolle. Järjestelmä toimii keskitettynä lokienhallintajärjestelmänä, tuottaa tietoa järjestelmän toiminnasta ja tapahtumista, auttaa vianselvityksessä sekä auttaa havaitsemaan kyberuhkia (Viestintävirasto, 2016). Sen toimintoihin kuuluu tyypillisesti tapahtumatietojen, eli lokitietojen, keräys valvottavasta järjestelmästä sekä datan esikäsittely tallennus ja esittäminen mukaan lukien raportointi sekä mahdolliset kyberuhkat ilmaisevat hälytykset. Tyypillisesti järjestelmän tuottamaa tietoa hyödyntää ja sen tuottamiin hälytyksiin reagoi kyberturvallisuusvalvontahenkilöstö, joka työskentelee organisaation kyberturvallisuusvalvomossa, SOC:ssa (Security Operations Center, (Rasche, 2013) (Feng, C. ym., 2017).

Vaikka SIEM-järjestelmän avulla voidaan parantaa organisaation näkymää valvottavaan järjestelmään ja havaita kerätyn lokitiedon perusteella haitallinen toiminta järjestelmässä sekä täyttää muun muassa laissa määritellyt vaatimukset tapahtumatietojen keräämisestä ja tallentamisesta (Viestintävirasto, 2016), liittyy sen toimintaan ja käyttöön useita haasteita. Järjestelmän tulee kerätä ja käsitellä suuri määrä useista lähteistä tulevaa tietoa. Ensimmäisten sukupolvien SIEM-järjestelmät käsittelevät tietoa ja tunnistavat mahdollisia uhkia ennalta määrättyjen sääntöjen perusteella, jolloin esimerkiksi järjestelmän tuottamissa hälytyksissä on suuri määrä vääriä positiivisia, mikä kuormittaa valvontahenkilöstöä. Lisäksi kyberuhkatoimijoiden toiminta kehittyy jatkuvasti monimutkaisemmaksi ja valvontajärjestelmän avulla on haitallisen toiminnan havaitseminen jo hyökkäyksen valmisteluvaiheessa haitallisen toiminnan estämiseksi, on haastavaa (Iklody, A., ym. 2018). (Feng, C., Wu, S. & Liu, N., 2017)

Kyberturvallisuusvalvonnassa kasvavana trendinä ja ratkaisuna edellä mainittuihin haasteisiin on nähty koneoppimisen hyödyntäminen tehostamaan kyberturvallisuusvalvontajärjestelmien toimintaa ja vastaamaan muun muassa SIEM-järjestelmiin liittyviin haasteisiin (Vähäkainu, P. & Lehto, M., 2019). Tässä tutkimuksessa tarkastellaan koneoppimisen hyödyntämismahdollisuuksia SIEM-järjestelmissä. Tutkimus on toteutettu Jyväskylän yliopiston digipalveluiden tarpeesta saada lisätietoa käyttämäänsä Splunk-ohjelmistoon perustuvan

SIEM-järjestelmän mahdollisuuksista. Splunk on kaupallinen tiedonhallintatyökalu, joka voi toteuttaa kaikki SIEM-järjestelmän tehtävät ja siinä voidaan hyödyntää laajasti sisäänrakennettuja sekä erikseen saatavilla olevia työkaluja ja lisäosia (Splunk Inc., 2021).

Aikaisempaa tutkimusta koneoppimisen hyödyntämisestä kyberturvallisuuden alalla on tehty laajasti muun muassa tunkeutumisenestojärjestelmiin liittyen (Buczak, A. ym., 2016). SIEM-järjestelmiin liittyviä opinnäytetöitä on tehty useita suomalaisissa korkeakouluissa, esimerkiksi «Toward Cyber Situational Awareness with Open Source Software», jossa suunnittelutieteellistä menetelmää hyödyntäen suunniteltiin ja toteutettiin avoimeen lähdekoodiin perustuva SIEM-järjestelmä (Teriö, J., 2017). Näissä ei kuitenkaan ole tutkittu tai toteutettu koneoppimistoiminnallisuuksia.

Koneoppimista on hyödynnetty useissa kaupallisissa SIEM-järjestelmissä, joiden tarkasta toiminnasta on kuitenkin vaikea saada yksityiskohtaista tietoa. Siksi Splunk-ympäristön hyödyntäminen tutkimuksessa tuottaa uusia näkökulmia koneoppimisen käytännön toteutuksesta. Tutkimuksen päämääränä onkin havainnollistaa, saadaanko koneoppiminen liitettyä käyttäjän tai ylläpitäjän toimilla osaksi kaupallista tietokanta- ja tiedonhallintasovellusta.

Työn päätutkimuskysymys on: *Miten koneoppimista voidaan hyödyntää SIEM-järjestelmissä haitallisen toiminnan havaitsemiseksi?* Asian selvittämiseksi päätutkimuskysymyksestä on johdettu seuraavat alatutkimuskysymykset:

- Mitä ovat SIEM-järjestelmät ja miten ne toimivat?
- Mitä ovat kyberuhkat ja miten kyberuhka voidaan havaita?
- Mitä koneoppiminen on ja mitä sen hyödyntäminen vaatii?
- Miten koneoppimista on jo hyödynnetty SIEM-järjestelmissä?

Tutkimus on luonteeltaan konstruktiivinen tutkimus Design Science Research Methdology for information systems research (DSRM) tutkimusmenetelmään hyödyntäen (Peffers, K. ym., 2007). Tutkimuksessa toteutettiin Splunk-ohjelmistoon perustuvaan SIEM-järjestelmään ohjaamatonta koneoppimista käyttävä toiminnallisuus, joka käsittelee valvottaman järjestelmän Linux-palvelinten lokitietoja siten, että lokitapahtumat jaetaan sisältönsä perusteella omiin joukkoihin klusteroinnin avulla mahdollisesti haitalliseen toimintaan viittaavien lokitapahtumien havaitsemisen helpottamiseksi. Tutkimus osoittaa, että koneoppiminen voidaan liittää osaksi SIEM-järjestelmän tiedonkäsittelytoimintoja ja järjestelmän tuottamat havainnot voidaan tarkastella reaaliajassa ja esittää SIEM-järjestelmän näkymien avulla.

Tutkimuksen seuraavassa luvussa kuvataan työssä käytetty tutkimusmenetelmä. Luvut kolme ja neljä muodostavat kirjallisuuskatsauksen, jossa tarkastellaan, miten SIEM-järjestelmä toimii, mitä tarkoitetaan kyberuhkilla ja miten niitä voidaan havaita sekä selvitetään, millaisia käyttötapauksia koneoppimisen hyödyntämistä SIEM-järjestelmissä on toteutettu. Tämän jälkeen luvussa 5 tarkastellaan Splunk-esimerkkijärjestelmän suunnittelu ja toteutus sekä esimerkkijärjestelmässä käytettävät koneoppimistoiminnallisuudet yksityiskohtaisesti,

mikä muodostaa tutkimuksen laajimman kokonaisuuden. Esittely järjestelmän toiminnasta kohdeympäristössä esitetään luvussa 6. Järjestelmän toiminnan arviointi ja toiminnasta tehdyt havainnot, jotka muodostavat merkittävimmät tutkimustulokset ovat luvussa 7. Tutkimuksen johtopäätökset ja samalla yhteenveto on esitetty luvussa 8.

2 TUTKIMUSMENETELMÄ

Tutkimuksen keskeinen tarkoitus on tehdä käytännön havaintoja koneoppimisesta SIEM-järjestelmissä, joten tutkimus päädyttiin tekemään konstruktiivisena tutkimuksena. Konstruktiivisessa tutkimuksessa pyritään esittämään ratkaisu johonkin käytännön ongelmaan sekä toteuttamaan ja demonstroimaan ratkaisun tekninen toteutus käytännössä. Työn eteneminen noudattaa konstruktiiviseen tutkimusmenetelmään erityisesti informaatioteknologian alalla luotua Design Science Research Methdology for information systems research (DSRM) tutkimusmenetelmää. Menetelmä luotiin tarpeesta kehittää informaatioteknologian tutkimukseen systemaattinen ja yhtenäinen suunnittelutieteellinen malli. Oleellinen osa menetelmää on tutkimusongelman ja ratkaisuesityksen osoittaminen tutkimuksessa käytännössä toteutettavan esimerkkiratkaisun tai järjestelmän avulla. Menetelmä soveltuu erityisesti informaatioteknologian tutkimukseen. Menetelmä on kuusivaiheinen prosessi. Prosessin vaiheet ja niiden sisältö tässä tutkimuksessa on seuraava (Peffers, K. ym., 2007):

KUVIO 1 DSRM tutkimusprosessi

1. Problem identification & motivation. Ongelman kuvaus ja motivointi.
Prosessi käynnistyy ongelman asettelulla. Vaiheessa kuvataan aiheeseen liittyvä ongelmakenttä ja tutkimusongelma sekä arvioidaan tutkimuksen avulla saatavien tulosten mahdollista hyötyä käsiteltävään aihealueeseen. Ongelman asettelun merkitys on suuri erityisesti tutkijoille asian merkityksen ymmärtämiseksi ja tutkimusmotivaation säilyttämiseksi läpi tutkimuksen. (Peffers, K., Tuunanen, T., Gengler, C., Rossi, M., Hui, W., Virtanen, V. & Bragge, J., 2006)

Tässä tutkimuksessa prosessin ensimmäinen vaihe muodostuu työn johdannosta sekä kirjallisuuskatsausosuudesta, jossa pureudutaan tarkasteluympäristöön, kyberturvallisuusvalvontaan, siihen liittyvään ongelmakenttään sekä esitetään ja joitakin esimerkkejä toteutetuista koneoppimisratkaisuista SIEM-järjestelmissä. Ensimmäisessä vaiheessa vastataan seuraaviin alatutkimuskysymyksiin:

- Mitä ovat SIEM-järjestelmät ja miten ne toimivat?
- Mitkä ovat SIEM-järjestelmien merkittävimmät haasteet kyberturvallisuusuhkien havaitsemisessa?
- Miten koneoppimista on jo hyödynnetty SIEM-järjestelmissä?

Kirjallisuuskatsaus perustuu selvitykseen aiheeseen liittyvsitä tieteellisistä julkaisuista sekä avoimesti julkaistuihin kaupallisten SIEM-järjestelmien, mukaan lukien Splunkin, dokumentaatioon erilaisista käyttötapauksista koneoppimisen hyödyntämiseen. Kirjallisuuskatsaus käsittää tutkimuksen luvut 3 ja 4.

2. Definition of the objectives of a solution. Tavoitteiden määrittely ongelman ratkaisemiseksi.
Prosessin toisen vaiheen tarkoitus on määritellä yksityiskohtaiset tavoitteet, mitä tutkimuksessa kehitettävällä ratkaisulla tai järjestelmällä tulee saavuttaa. Tavoitteet tulee perustua tutkimusongelmaan ja ne voivat olla joko määrällisesti tai laadullisesti mitattavia ominaisuuksia. (Peffers, K., Tuunanen, T., Gengler, C., Rossi, M., Hui, W., Virtanen, V. & Bragge, J., 2006)

Tutkimusprosessin toinen vaihe käsittää tässä tutkimuksessa työn tilaajan kanssa yhteisesti asetetut tavoitteet ja vaatimukset toteutettavalle järjestelmälle sekä käytännön reunaehdot järjestelmän suunnittelua ja toteutusta ajatellen. Pohjana määritellylle on tutkimuksen ensimmäisessä osiossa laaditun kirjallisuuskatsauksen havainnot. Järjestelmän tavoitteiden ja vaatimusten määrittely käsitellään tutkimuksen luvussa 5.2.

3.Design development. Toteutuksen suunnittelu ja kehittäminen.
Prosessin kolmannessa vaiheessa pureudutaan varsinaisen järjestelmän suunnitteluun ja toteutukseen. Suunnitellussa määritellään tehtävän toteutuksen toiminta ja arkkitehtuuri yksityiskohtaisesti, minkä pohjalta käytännön toteutus rakennetaan. Toteutus edellyttää kyseisen aihealueen teoriapohjan ymmärrystä ja soveltamiskykyä, mitä tutkimusprosessin aikaisemmat vaiheet tukevat. (Peffers, K., Tuunanen, T., Gengler, C., Rossi, M., Hui, W., Virtanen, V. & Bragge, J., 2006)

Tämän tutkimuksen osalta prosessin kolmannessa vaiheessa avataan tarkemmin koneoppimisen perusteita sekä tarkastelen koneoppivien järjestelmien suunnitteluun liittyvät periaatteet sekä siihen liittyviä haasteita. Samalla vastataan alatutkimuskysymykseen: Mitä koneoppiminen on ja mitä sen hyödyntäminen vaatii? Tässä vaiheessa tutkimuksen kirjallisuuskatsauksessa tehtyjä havaintoja SIEM-järjestelmien toiminnasta sekä tunnistetuista käyttötapauksista sovelletaan käytäntöön Splunk-sovellusympäristössä. Vaiheen tulokset on dokumentoitu tutkimuksen luvussa 5.

4. **Demonstration**. Ratkaisun esittely.
Demonstraatio-vaiheessa esitellään prosessin edellisessä vaiheessa luodun ratkaisun käytännöntoiminta sekä havainnollistetaan, miten toteutettu ratkaisu toimii alkuperäiseen ongelmaan. Ratkaisun esittelyssä tulee kuvat yksityiskohtaisesti, kuinka konstruktio toimii ja kuinka sitä käytetään kuvattuun ongelmaan (Peffers, K., Tuunanen, T., Gengler, C., Rossi, M., Hui, W., Virtanen, V. & Bragge, J., 2006). Tässä tutkimuksessa toteutettu Toteutetun järjestelmän toiminta kuvataan yksityiskohtaisesti luvussa 6 tutkimuksen tilaajan järjestelmässä ja sinne kerättävällä lokidatalla.

5. **Evaluation**. Arviointi.
Järjestelmän toimivuus ja tehokkuus evaluoidaan prosessin viidennessä vaiheessa. Kehitetyn ratkaisun arvioinnissa tarkastellaan järjestelmän esittelyssä havainnollistettuja ominaisuuksia ja tuloksia sekä arvioidaan, saavutetaanko sillä toteutukselle asetettuja, tutkimusprosessin toisessa vaiheessa, asetettuja vaatimuksia. Arviointi voidaan toteuttaa monella tavalla, esimerkiksi toteuttamalla järjestelmälle suorituskykymittauksia, teettää asiakastutkimuksia järjestelmän käyttäjille tai keskittyä empiiriseen arvioon. (Peffers, K., Tuunanen, T., Gengler, C., Rossi, M., Hui, W., Virtanen, V. & Bragge, J., 2006)

Tässä tutkimuksessa arviointi toteutetaan tutkimuksen tilaajan ympäristössä toteutettua demonstraatiota tarkastellen. Keskeisessä roolissa on tutkimuksen tilaajilta saadut havainnot ja palaute järjestelmän toiminnasta. Järjestelmän arviointi muodostaa tutkimuksen keskeisimmät tulokset ja ne on dokumentoitu luvussa 7.

6. Communication. Kommunikaatio.
Tutkimusprosessin kuudes vaihe käsittää vuoropuhelun tutkimustulosten merkityksestä ja tehokkuudesta, mikä tämän tutkimuksen osalta muodostuu tästä kirjallisesta raportista sekä erityisesti luvusta 8, johtopäätökset, johon tiivistyy tutkimuksen keskeisimmät havainnot ja pohdinta.

3 SYSTEM INFORMATION AND EVENT MANAGEMENT -JÄRJESTELMÄT

SIEM on tietoturvallisuuden alalla vakiintunut termi, joka usein ymmärretään yksittäisenä tietoturvatuotteena tai ohjelmistona, mutta todellisuudessa kyse on laajemmasta kokonaisuudesta ja SIEM voidaankin nähdä myös prosessina tai tietoturvallisuuden hallinnan tapana. Oleellinen osa SIEM-järjestelmän toimintaa on joka tapauksessa SIEM-järjestelmän rooli organisaation tietojärjestelmästä kerättävän tiedon kokoamisen yhteen paikkaan analysoitavaksi, jolloin yksittäisistä pienistä tapahtumista voidaan muodostaa käsitys isommista tapahtumista, havaita ongelmia sekä nopeuttaa ongelmiin reagointia. (Viestintävirasto, 2016)

Tässä luvussa pureudutaan tarkemmin SIEM-järjestelmän toimintaan ja vastataan alatutkimuskysymyksiin:

- Mitä ovat SIEM-järjestelmät ja miten ne toimivat?
- Mitä ovat kyberuhkat ja miten kyberuhka voidaan havaita?

SIEM-järjestelmän toiminnan kuvaamiseksi luvussa tarkastellaan myös, mitä tarkoitetaan SIEM-järjestelmässä käsiteltävällä lokitiedolla.

3.1 Loki ja lokienhallinta

SIEM-järjestelmän yksi keskeinen tehtävä on kerätä, käsitellä ja varastoida valvottavasta järjestelmästä lokitietoja. Lokitiedolla tarkoitetaan tiettynä ajanhetkenä kirjattua tallennetta tapahtumasta ja sen aiheuttajista. Lokitieto kertoo, mitä, miksi ja milloin jotakin tapahtui ja sen perusteella voidaan seurata tietojärjestelmissä olevien tietojen käyttöä, teknisiä virheitä sekä myös havaita järjestelmälle ja organisaatiolle haitallinen toiminta. (Viestintävirasto, 2016)

Lokitapahtumalla tarkoitetaan järjestelmän, sovelluksen tai laitteen muodostamaa tapahtumatietoa ja lokilähteellä tarkoitetaan järjestelmää, sovellusta tai laitetta, joka tuottaa lokitapahtumia (Valtionhallinnon tietoturvallisuuden johtoryhmä, 2009). Lokeja on monen tyyppisiä eri käyttötarkoitusten mukaan. Esimerkiksi Viestintäviraston lokiohje määrittelemiä lokityyppejä ovat: ylläpitoloki, käyttöloki, muutosloki, virheloki, viestintäloki, haltijaloki ja pääsynvalvontaloki. (Viestintävirasto, 2016)

Organisaation näkökulmasta on oleellista tunnistaa lokitietojen merkitys organisaation toiminnalle sekä huomioida toimialaa koskevat erityismääräykset. Koska lokien kerääminen tuottaa arkaluontoista tietoa järjestelmän ja sen käyttäjien toiminnasta, lokien keräämille tulee olla perusteltu tarve (Viestintävirasto, 2016). Tämä edellyttää organisaatiolta lokienhallinnan suunnittelua, mikä käsittää esimerkiksi organisaation henkilöstön roolit ja vastuut lokienhallinnassa, lokipolitiikan määrittelyn, lokienhallintainfrastruktuurin ja

lokienkäsittelyprosessien suunnittelun (Souppaya, M., Scarfone, K., 2006). Tämä tutkimuksen kannalta pääasiallinen tarkastelunkohde on lokienhallintainfrastruktuurin tarkastelu.

ISO 27001 on merkittävin ja käytetyin standardi tietoturvallisuuden alalla (Susanto, H., ym., 2011). ISO 27001 ja ISO 27002 määrittelevät joukon vaatimuksia ja suosituksia myös organisaation lokienhallinnalle ja lokituksen tekniselle toteutukselle. Esimerkiksi Katakri 2020 ja Valtionvarainministeriön suosituskokoelma tiettyjen tietoturvallisuussäädösten soveltamisesta suosittelevat keskitetyn ja vahvasti suojatun lokienhallinnan käyttöä tehokkaamman seurannan ja analysoinnin mahdollistamiseksi (Kansallinen turvallisuusviranomainen, 2020) (Tiedonhallintalautakunta, 2020). SIEM-järjestelmällä on siten merkittävä rooli organisaation työkaluna, ei pelkästään eri tapahtumien havaitsemisesta lokitiedoista, vaan myös lokien varastoinnissa, koska SIEM-järjestelmässä määritellään organisaation lokipolitiikan mukaisesti muun muassa lokitietojen elinkaari eri lokityypeille, lokitiedon varmuuskopiointi ja pääsyoikeudet lokitietoon.

3.2 Kyberuhkat ja niiden havaitseminen

Yksi SIEM-järjestelmän tehtävä on tunnistaa lokitietojen perusteella organisaatioon kohdistuvia kyberuhkia. Kyberuhkalla tarkoitetaan mahdollisuutta kybertoimintaympäristöön vaikuttavaan tekoon tai tapahtumaan, joka toteutuessaan vaarantaa jonkin organisaation kybertoimintaympäristöstä riippuvaisen toiminnon. Kyberuhkat ovat tietoturvauhkia, jotka toteutuessaan vaarantavat organisaation tietojärjestelmän oikeanlaisen toiminnan. (Turvallisuuskomitean sihteeristö, 2013)

Kyberuhkat voidaan luokitella esimerkiksi hyökkäyksen tarkoitusta tai päämäärää, lainsäädännöllistä näkökulmaa, vakavuutta (aktiivinen vs. passiivinen), laajuutta tai kohdejärjestelmän tyyppiä, johon hyökätään (Uma, M., Padmavathi, G., 2013).

Kyberuhkien havaitsemisen kannalta erilaisten uhkien tarkastelunäkökulmana voidaan käyttää kyberhyökkääjän toimintatapoja ja tekniikoita. Erityisesti Advanced Persistence Threat (APT) -hyökkäysten kuvaamisen on laadittu viitemalleja, joiden tarkoitus tarkoitus kuva kyberhyökkäyksen vaiheita ja toimintaa kohdejärjestelmässä. Tällaisia viitemalleja ovat muun muassa Attack Lifecycle, Cyber Kill Chain, Unified Kill Chain ja ATT&CK (The MITRE Corporation, 2021). Malleissa on useita samankaltaisuuksia, mutta erityisesti Mitren ATT&CK framework -mallissa kuvataan yksityiskohtaisesti tunnettuja hyökkäystekniikoita, vastustajan toimintaa kohdejärjestelmässä sekä toiminnan havaitsemis- ja suojautumiskeinoja hyvin yksityiskohtaisesti. Helmikuussa 2021 mallissa oli kuvattuna 14 pääluokkaa ja 206 erilaista hyökkäystekniikkaa, jotka liittyvät esimerkiksi hyökkäyksen kohdejärjestelmän tiedusteluun sekä hyökkäyksen valmisteluun, ja voivat toimia indikaattoreina tulevasta hyökkäyksestä, minkä vuoksi toimien havaitseminen varhaisessa vaiheessa on oleellista (Bowers, B., ei pvm). Taulukossa 1 on esitetty Mitre ATT&CK -mallin hyökkäystekniikiden pääluokat sekä

kuvaus niihin kuuluvien taktiikoiden tavoitteesta hyökkääjälle. (The MITRE Corporation, 2021)

TAULUKKO 1 Mitre ATT&CK framework pääluokat

Pääluokka	**Kuvaus**
Reconnaissance	Kohteen tiedustelu
Resource Development	Omien resurssien kehittäminen
Initial Access	Jalansijan muodostaminen kohdejärjestelmään
Execution	Haitallisen koodin suorittaminen kohdejärjestelmässä
Persistence	Pysyvyyden varmistaminen
Privilege Escalation	Oikeuksien kasvattaminen
Defense Evasion	Vastatoimien välttteleminen
Credential Access	Käyttäjäoikeuksien hankkiminen
Discovery	Kohteen tunnistaminen
Lateral Movement	Liikkuminen kohdejärjestelmässä
Collection	Tiedon keräys
Command and Control	Komentokanavan luominen
Exfiltration	Datan kotiuttaminen
Impact	Vaikuttaminen kohdejärjestelmässä

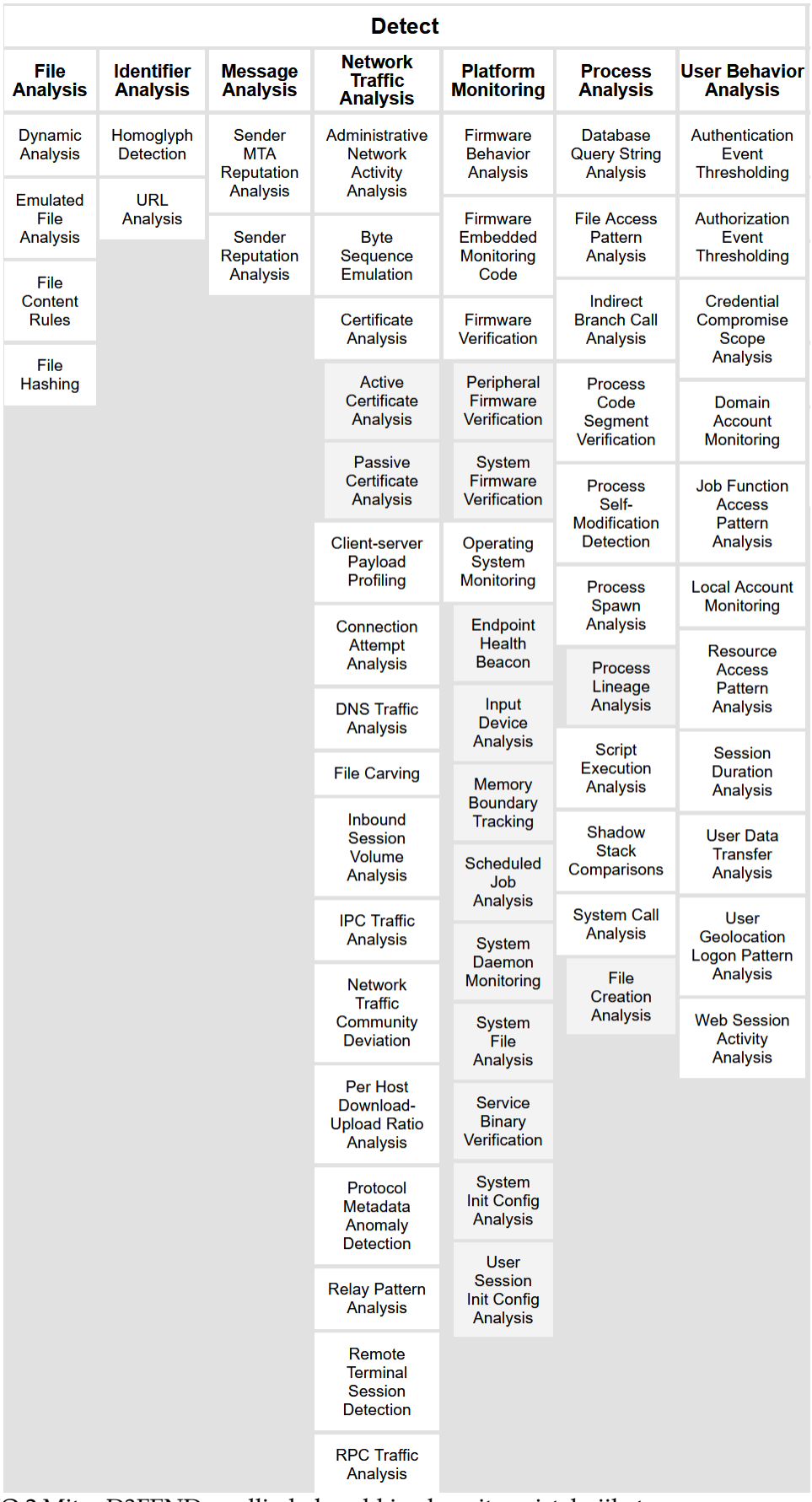
ATT&CK viitemallin yhteydessä on kuvattu myös datalähteet, joiden tietoa seuraamalla puolustajan on mahdollista havaita tiettyjä hyökkäystekniikoita. Esimerkiksi datalähteillä DS0008 (Kernel) ja DS0015 (Application log) viitataan lokitietojen keräämiseen esim. Linux käyttöjärjestelmän ytimestä, Kernelistä, sekä eri sovelluksilta ja palveluilta, jotka voivat toimia havainnointikeinona laskentatavan mukaan vähintään kolmeenkymmeneen eri hyökkäystekniikkaan.

ATT&CK viitemallin rinnalla ylläpidetään Mitren toimesta myös vastaavasti puolustajan toimenpiteiden kuvaamiseksi vielä melko varhaisessa vaiheessa olevaa D3FEND -mallia, joka perustuu juuri ATT&CK-mallissa kuvattuihin uhkiin. Malli kuvaa mahdollisia tietojärjestelmien puolustuskeinoja muun muassa järjestelmän koventamiseksi, uhkien havaitsemiseksi, järjestelmän eristämiseksi, hyökkääjän harhauttamiseksi sekä hyökkääjän häätämiseksi järjestelmästä.

Kuviossa 2 on D3fend -mallin Detect-taktiikan alatekniikat, jotka sisältävät SIEM-järjestelmän toiminnan näkökulmasta useita keinoja, joita voidaan hyödyntää uhkien havaitsemiseen SIEM-järjestelmän avulla, esimerkiksi havaitsemistekniikka: D3-OSM operating system monitoring sisältää muun muassa seuraavia alatekniikoita:

- Päätelaitteiden tilan seuranta
- Järjestelmän ajastettujen toimintojen seuranta
- Järjestelmäprosessien seuranta

- Järjestelmätiedostojen muutosten seuranta
- Järjestelmän käynnistysmääritysten seuranta
- Järjestelmään kytkettävien laitteiden seuranta



KUVIO 2 Mitre D3FEND -mallin kyberuhkien havaitsemistekniikat

3.3 SIEM-järjestelmän hyödyntäminen organisaatiossa

SIEM-järjestelmä ei olemassaololollaan vielä ratkaise organisaation tietoturvallisuusongelmia tai estä haitallista toimintaa. SIEM-järjestelmään kerättyä ja järjestelmän käsittelemää tietoa hyödynnetään yleensä organisaation tietoturvallisuusvalvomon (engl. Security Operations Center, SOC) henkilöstö: SOC-operaattorit tai -analyytikot. SOC voi olla organisaation oma ulkoisena palveluna ostettu. Muita nimityksiä vastaavalle toiminnallisuudelle ovat: Computer Security Incident Response Team (CSIRT), Computer Incident Response Team (CIRT), Computer Security Incident Response Capability (CSIRC), Network Operations and Security Center (NOSC), Cyber Security Operations Center (CSOC). Toisaalta SOC-toiminnallisuus voi muodostua organisaation tietojärjestelmän ylläpitäjistä. Joka tapauksessa SIEM-järjestelmää käyttävän henkilöstön tulee toimia läheisesti järjestelmän ylläpitäjien kanssa ja SIEM-järjestelmällä voidaan monitoroida myös. (Zimmerman, C., 2014)

Tästä syystä SIEM-järjestelmän kehityksessä onkin oleellista huomioida, kuka järjestelmän tuottamaa tietoa hyödyntää ja miten ja kuinka nopeasti poikkeamiin tulee reagoida. Vaikka järjestelmässä olisikin koneoppimista hyödyntäviä toimintoja ja järjestelmä on ihmisen rooli varsinaisten toimenpiteiden ja päätöksentekijänä kuitenkin keskeinen myös oikeudellisesta näkökulmasta. (Viestintävirasto, 2016)

3.4 SIEM-järjestelmän toiminta

3.4.1 SIEM-toiminnalisuudet Splunkissa

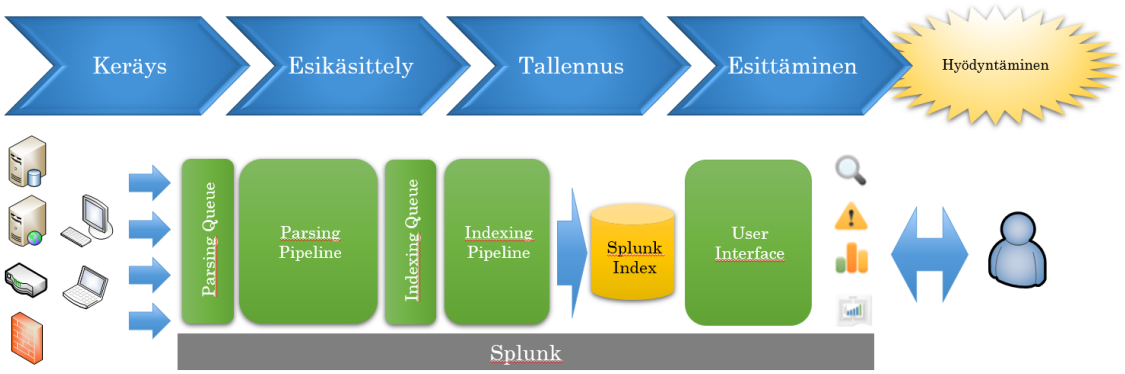
Tutkimuksen yhtenä tarkoituksena on suunnitella ja toteuttaa koneoppimista hyödyntävä SIEM-järjestelmä käyttäen kaupallista Splunk-sovellusta. Splunk-sovellus on Splunk-yhtiön merkittävin tuote ja sen kaupallinen, asiakkaan omalle palvelimelle asennettava versio, johon tässä tutkimuksessa viitataan, on nimeltään Splunk Enterprise (Splunk Inc., 2021).

SIEM-toiminnallisuudet ovat vain yksi osa Splunkin ominaisuuksista. Se on suuren tietomäärän käsittelyyn soveltuva tietojenkäsittelyalusta, jonka avulla voidaan toteuttaa monipuolisesti datan käsittely aina datalähteestä käyttäjälle esitettävään näkymään. Lisäksi sovelluksen toimintoja voi lisätä merkittävästi erikseen ladattavien liitännäisten avulla, joihin lukeutuu muun muassa lukuisia koneoppimista hyödyntäviä lisäosia.

Sovelluksen muita käyttötarkoituksia ovat esimerkiksi järjestelmien monitorointi, kaupankäynti ja business-analytiikka. Tuotteen merkittävimpänä asiakasryhmänä ovat yritykset ja yhteisöt, mutta rekisteröityneen käyttäjän on mahdollista tutustua järjestelmään rajoitetun evaluointiversion avulla. Itse sovellus on palvelinohjelmisto, joka on asennettavissa useille Linux-jakeluille tai

Windowsille, ja se käyttö tapahtuu selainkäyttöliittymän avulla. Splunk instanssi voidaan asentaa yhteen fyysiseen tai virtuaalipalvelimeen tai kuorman jakamiseksi eri komponentit voidaan hajauttaa useisiin instansseihin. (Splunk Inc., 2021)

Tässä alaluvussa SIEM-järjestelmän toimintaa tarkastellaan tiedonkäsittelyn päävaiheiden perusteella. Kuviossa 3 on esitetty SIEM-järjestelmän tiedonkäsittelyvaiheet ja toiminta Splunkissa, joita ovat: lokien keräys, esikäsittely, tallennus ja esittäminen. Lisäksi olennaisena osana tarkastellaan SIEM-järjestelmän esittämän tiedon hyödyntämistä osana SOC: n toimintaa.



KUVIO 3 Splunk-järjestelmän toimintaperiaate

3.4.2 Keräys

Organisaation tietojärjestelmän laitteet toimivat lokilähteinä. Lähtökohtaisesti lähes kaikki tietoverkkoon liitettävät laitteet tuottavat paikallisesti tallennettavaa lokitietoa, mutta lokien keskitetty hyödyntäminen SIEM-järjestelmässä vaatii niiden siirtämisen sinne. Lokien keräyksessä on huomioitava, mitkä lokitiedot, ja miltä laitteilta, siirretään SIEM-järjestelmään. Yleisesti tärkeimmät lähteet ovat autentikointitiedot (kirjautumisloki) sekä palomuuri- ja IDS-järjestelmien lokitiedot (Todd, B., 2017). Kerättävällä datalla on vaikutus SIEM: n toiminnan kannalta, koska se määrittelee, mitä tietoa on käytettävissä analyysin pohjana.

Lokien siirtämiseen SIEM: lle on useita vaihtoehtoja. Usein verkkolaitteet ja palvelinsovellukset voidaan konfiguroida lähettämään lokitietonsa ulkoiselle lokipalvelimelle syslog-muodossa. Toinen yleinen teknologia on jo 1980-luvulla verkonvalvontaan kehitetty Simple Network Management Protocol (SNMP), joka on edelleen käytössä.

Mikäli laitteistossa ei ole suoraa tukea tapahtumatiedon lähettämistä tukeviin protokolliin tai niitä ei haluta käyttää, voidaan tapahtumatiedonsiirto toteuttaa päätelaitteelle asennettavan agenttiohjelmiston avulla. Splunkissa ei ole erityisiä rajoituksia sisään tulevalle datalle ja se tukee useita tiedonsiirtomenetelmiä, dataformaatteja sekä tarjoaa käyttäjilleen myös Universal Forwarder -agenttiohjelmistoa lokidatan siirtoon. (Splunk Inc., 2021)

3.4.3 Esikäsittely

Lokitiedon esikäsittelyllä tarkoitetaan niitä toimenpiteitä, mitä valvottavasta ympäristöstä kerätylle lokitiedolle tehdään ennen kuin se tallentuu SIEM-järjestelmään. Näitä ovat esimerkiksi lokitiedon suodatus, normalisointi, rikastaminen ja korrelointi. Koska eri valmistajien laitteiden tuottaman lokitiedon rakenteessa on eroavaisuuksia, esikäsittelyn avulla varmistetaan, että SIEM-järjestelmään tallennettava tieto on yhdenmukaista ja järjestelmän ymmärtämässä muodossa.

Lokitiedosta poistetaan tarpeettomat osat suodatusvaiheessa. Poistettavia osuuksia ovat esimerkiksi yleistason kuvaus lokitapahtumasta, joka toistuu samanlaisena kaikissa vastaavissa lokitapahtumissa lähteestä riippumatta. Toimenpiteen avulla vähennetään lopullisen lokin tilantarvetta tietovarannossa. (Todd, B., 2017)

Normalisoinnissa eri lähteistä tuleva tieto yhdenmukaistetaan. Usein esimerkiksi aikatieto voi olla lokeissa eri aikaformaatissa, mikä korjataan normalisointiprosessissa (Todd, B., 2017). Rikastamisella tarkoitetaan puolestaan ulkoisesta lähteestä haettavan tiedon liittämistä lokitietoon. Yleisimpiä tapoja rikastaa lokitietoa on liittää lokitapahtumassa olevaan IP-osoitteeseen sille kuuluva DNS-nimi ja geoip-tieto, eli maantieteellinen sijainti, johon kyseinen IP-osoite on rekisteröity (Henderson, J., Hubbard, J., 2017).

Korreloinnissa eri lähteistä tuleva lokitieto yhdistetään samaksi tapahtumaksi, mikä auttaa haitallisen toiminnan tunnistamisessa. Korrelointia voidaan toteuttaa SIEM-järjestelmässä joko reaaliajassa datan esikäsittelyvaiheessa tai vasta haettaessa järjestelmästä tietoja esitettäväksi. Haasteena tiedon eri esikäsittelyvaiheissa on, että toimenpiteet tehdään yleensä sääntöperusteisesi, mikä aiheuttaa jatkuvaa ylläpitoa. (Stroeh, K, Madeira, E. & Goldenstein, S., 2013)

3.4.4 Tallennus

Esikäsittelyn jälkeen lokit tallentuvat SIEM-järjestelmän tietovarantoon. Tallennustavassa on tuotekohtaisia eroja. Lokien tallennuksessa tulee huomioida kunkin lokityypin elinkaari, joka tulisi määritellä organisaation lokipolitiikassa. Esimerkiksi Splunkissa määritellään sisään tulevalle datalle indeksi, johon tieto tallennetaan. Keskitetty lokien tallennus mahdollistaa myös lokien varmuuskopioinnin ja poistamisen elinkaaren päätyttyä.

3.4.5 Tiedon esittäminen

SIEM-järjestelmän käsittelemä tieto esitetään sovelluksen käyttöliittymässä. Samassa käyttöliittymässä hallitaan yleensä kaikkia SIEM:in tiedonkäsittelyn vaiheita ja tarkastellaan dataa raakamuodossa. Tiedon hyödyntämisen kannalta keskeistä on käyttöliittymässä esitettävät näkymät (engl. dashboardit), hälytykset ja raportit.

Visualisointien avulla järjestelmän suuresta tietomäärästä esitetään käyttäjälle olennaisimmat tiedot, hälytykset mukaan lukien, erilaisin kuvaajin ja

taulukoin havainnoinnin helpottamiseksi. Näkymät ovat täysin käyttäjän muokattavissa. Splunkissa visualisointien tekemiselle ja muokkaamiselle on useimpien SIEM-ratkaisuiden tavoin rajattomat mahdollisuudet. Splunkin applikaatioiden kautta valmiita näkymiä on lisäksi saatavilla, usein tiettyyn käyttötarkoitukseen. Yksi tarkastelunäkökulma on käyttää Mitren ATT&CK-viitemalliin tehtyä visualisointia, joka esittää tapahtumat suhteessa kyseisen viitemallin kategorioihin ja hyökkäystekniikoihin (Stoner, J., 2019).

Näkymille tai näkymissä esitettävälle tiedolle ei ole kuitenkaan erityisiä standardeja (Kokkonen, T., 2016), vaan niiden hyödyntäminen riippuu käyttäjän tarpeista ja mieltymyksistä ja usein tarpeeksi kattavien, mutta riittävän tarkkojen ja helppokäyttöisten näkymien luominen on haastavaa. Toisaalta visualisointien avulla käyttäjä voi ymmärtää paremmin SIEM-järjestelmän dataa sekä muun muassa käytettävien koneoppimismenetelmien toimintaa ja tuloksia. Kuviossa 4 on esimerkki SIEM-järjestelmän tiedon esittämiseen tehdystä näkymästä, joka on yleiskuvaus järjestelmään kerätystä tiedosta muodostetuista hälytyksistä ajan suhteen. (Vielberth, M., Böhm, F., Fichtinger, I., Pernul, G., 2020)

KUVIO 4 Esimerkki Splunkin SIEM-käyttöön räätälöidystä näkymästä

3.4.6 Tiedon hyödyntäminen

SIEM-järjestelmä ei itse toteuta aktiivisia toimia havaitessaan poikkeamia, vaan tiedon hyödyntämisessä ja vastatoimissa on avainasemassa SOC: n henkilöstö. SOC-analyytikon tehtävänä on tunnistaa SIEM: n esittämän tiedon perusteella kiinnostavat tapahtumat (engl. event of interest), ja muodostaa (engl. escalate) niistä tapauksia (engl. case), mikäli kyseinen tapahtuma tai tapahtumat voivat liittyä haitalliseen toimintaan. Tarkemman analyysin perusteella tapauksista voi paljastua meneillään oleva, mennyt tai mahdollisesti tuleva tietoturvaloukkaus (engl. incident), joka johtaa tapahtumaan reagointiin (engl. incident response). (Zimmerman, C., 2014)

Perinteisen tapahtumien korrelointi ja hälytysten tuottaminen SIEM-järjestelmässä tapahtuu sääntöperusteisesti, jolloin haasteena on joko suuri väärien positiivisten määrä hälytyksissä tai toisaalta liian tiukat tai huonosti toteutetut säännöt eivät tuota hälytyksiä oikeasti haitallisista tapahtumista. Suuri väärien positiivisten määrä kuormittaa SOC: n henkilöstöä tehden työstä monotonista (Vielberth, M., Böhm, F., Fichtinger, I., Pernul, G., 2020) johtaen jatkuvaan sääntöjen muuttamiseen (Kokulu, F. B., Shoshitaishvili, Y., Ziming Zhao, A. S., Ahn, G. J., Bao, T., Doupé, A., 2019).

Järjestelmästä, joka toteuttaa aktiivisia vastatoimia uhkia havaitessaan, kutsutaan Security Orchestration, Automation and Response (SOAR)-järjestelmäksi, jonka tarkoituksena on vähentää SOC: n työkuormaa (Watts, S., 2020). Sen toiminta perustuu koneoppimisen hyödyntämiseen tapahtumien analysoinnissa. SOAR ja SIEM eivät ole sama asia, eivätkä toisiaan poissulkevia, koska SOAR-järjestelmä ei toteuta kaikkia SIEM-järjestelmän perustehtäviä kuten lokidatan keräystä, esikäsittelyä ja tallentamista. Esimerkiksi Splunk-yhtiöllä on erillinen SOAR-tuote Splunk Phantom, joka tarvitsee tapahtumatietoa esimerkiksi organisaation SIEM-järjestelmältä (Splunk Inc., 2021).

Tutkimuksessa tarkastellaan koneoppimisen hyödyntämismahdollisuuksia SIEM-järjestelmissä, joten SOAR-toiminnallisuudet on rajattu tutkimuksen ulkopuolella ja pääpaino on tunnistaa menetelmiä SIEM-järjestelmään liittyvien haasteiden, kuten oleellisen tiedon havaitsemisen ja väärien positiivisten tuottaman työkuorman, ratkaisemiseen.

4 KÄYTTÖTAPAUKSIA KONEOPPIMISEN HYÖDYNTÄMISEEN SIEM-JÄRJESTELMISSÄ

Tutkimuksen ensimmäisessä osassa tarkastellaan koneoppimisen hyödyntämismahdollisuuksia SIEM-järjestelmissä selvittämällä, millaisiin käyttötapauksiin koneoppimista on jo hyödynnetty SIEM-järjestelmissä sekä avataan koneoppimisen käsitettä. Tässä luvussa vastataankin tutkimuksen seuraaviin alatutkimuskysymyksiin.

- Mitä koneoppiminen on?
- Miten koneoppimista on jo hyödynnetty SIEM-järjestelmissä?

Koska ala kehittyy kovaa vauhtia ja uusia toteutuksia syntyy jatkuvasti, tutkittua tietoa SIEM-järjestelmien käyttötapauksista on suhteellisen vähän. Tutkimuksessa päädyttiin tarkastelemaan käyttötapauksia edellisessä luvussa kuvatun SIEM-järjestelmän käsittelyketjun eri vaiheiden mukaisesti.

Mikäli tarkastellaan SIEM-järjestelmän lokien käsittelyketjua kokonaisuudessaan, voidaan koneoppimista hyödyntää useissa lokien käsittelyn vaiheissa, kuten keräysvaiheessa hyödyntämällä koneoppimista jo lokilähteessä, lokien esikäsittelyssä sekä tiedonlouhinnassa SIEM-järjestelmästä, mikä on tutkimuksen havaintojen perusteella yleisin ja potentiaalisin koneoppimisen hyödyntämiskohde.

4.1 Koneoppiminen

Tekoäly on yläkäsite laajalle kirjolle menetelmiä ja sovellutuksia, jotka kykenevät suorittamaan ainakin osittain autonomisesti älykkäitä toimintoja ympäristönsä huomioiden. Tekoälyn määrittely on haastavaa ja termin käyttö on levinnyt laajalle ja tekoäly-termiä käytetäänkin myös asioiden yhteydessä, jotka eivät todellisuudessa ole tekoälyä. (Boucher, P, 2020)

Koneoppiminen on yleisin tekoälyn käytännön sovellus. Sillä tarkoitetaan menetelmiä, jossa laskennassa käytettävä algoritmi on harjoitettu esimerkkidatan avulla. Se on koneen kykyä oppia ilman, että se ohjelmoidaan. (Boucher, P, 2020) Koneoppimiseen puolestaan liittyy alakäsitteitä, kuten syväoppinen. Syväoppimisella tarkoitetaan menetelmiä, jonka algoritmit toimivat ihmisaivojen tavoin. (Sindhu, V., Nivedha, S., Prakash, M., 2020)

Koneoppiminen yhteyteen lisätään usein myös koneoppimisen opetusmuoto, joita ovat ohjattu- (engl. supervised learning), puoliohjattu- (engl. semi-supervised), ohjaamaton- (engl. unsupervised learning) sekä vahvistusoppiminen (engl. reinforcement learning). Ohjatussa oppimisessa koneoppimismallin opetuksessa käytetään valmiiksi luokiteltua tietoa, jonka perusteella

koneoppimismalli muodostetaan. Ohjaamattomassa oppimisessa dataa ei ole luokiteltu valmiiksi. (Buczak, A. & Guven, E., 2016)

Esimerkkejä koneoppimisen käytännön sovellutuksista ovat esimerkiksi Optical Character Recognition (OCR), eli tekstintunnistus, Natural Language Processing (NLP), luonnollisen kielen prosessointi, Human Language Technology (HLT), jota hyödynnetään puheentunnistuksessa, kuvantulkinta sekä kasvojentunnistus.

4.2 Koneoppimisen hyödyntäminen lokilähteissä

Koneoppimista voidaan hyödyntää havaitsemaan kyberuhkia jo lokilähteissä, jolloin SIEM-järjestelmälle lähetetään uhkiin liittyvät hälytykset. Tunkeutumisenhavaitsemis-, eli IDS (Intrusion Detection System) -järjestelmät ovat yksi keskeinen lokilähde SIEM-järjestelmälle, koska niiden pääasiallinen tehtävä on havaita organisaation tietoverkkoon kohdistuva haitallinen toiminta. Päätelaitteissa toimivasta tunkeutumisenhavaitsemisjärjestelmästä käytetään termiä HIDS (Host-based Intrusion Detection System) ja verkkoliikennettä tarkkailevasta termiä NIDS (Network-based Intrusion Detection System). (Buczak, A. & Guven, E., 2016)

Vaikka IDS-järjestelmien tarkoitus onkin havaita kyberuhkia, eivät ne tässä suhteessa korvaa SIEM-järjestelmää. IDS-järjestelmien näkymä on rajattu vain kyseiseen päätelaitteeseen tai siihen kohtaan tietoverkkoa, mihin järjestelmä on sijoitettu, eivätkä ne SIEM-järjestelmän tavoin käsittele useista lähteistä tulevaa tietoa. Perinteisesti IDS-järjestelmien toiminta perustuu anomalioiden tunnistamiseen ja/tai etukäteen määritettyihin haittaohjelmatunnisteisiin ennalta määrättyjen sääntöjen perusteella. Anomalioihin perustuvassa tunnistamisessa haasteena on IDS-järjestelmän tuottamien väärien positiivisten suuri määrä ja tunnisteisiin perustuvassa haasteena on jatkuva tunnisteiden päivitystarve. Muun muassa näitä haasteita on pyritty lieventämään koneoppimista hyödyntämällä. Käytettyjen koneoppimismenetelmien vertailussa ei kuitenkaan havaittu, että jonkin tietty menetelmä toimisi toisia paremmin, mikäli huomioidaan eri tekijät, kuten tehokkuus, tarkkuus, monimutkaisuus, opettamiseen vaadittava aika. (Buczak, A. & Guven, E., 2016)

Koneoppimisen eri muotoja päätelaitteiden haittaohjelmien tunnistamisessa hyödyntäviä käytännön toteutuksia on useita, esimerkiksi Deep Instinct, SparkCognition DeepArmor ja Vectra Networks Cognito (Vähäkainu, P. & Lehto, M., 2019). Yhteistä näissä sovelluksissa on, että itse sensori toimii päätelaitteessa, mutta algoritmin opettaminen tapahtuu palveluntarjoajan palvelimella, missä syväoppimisen keinoin opetetaan sovellus tunnistamaan erityyppisiä haittaohjelmia ja viruksia suuresta määrästä näytteitä.

Etuna lokilähteissä tehtävässä uhkien havainnoinnissa on, että SIEM-järjestelmään siirrettävän tapahtumatiedon määrää voidaan vähentää, mikä säästää SIEM-järjestelmän ja tiedonsiirtoverkon kapasiteettia ja täten vähentää organisaation kuluja.

4.3 Lokitiedon esikäsittely koneoppimisen avulla

Yksi SIEM-järjestelmän haasteista on tapahtumatiedon monimuotoisuus. Lokietietoa saadaan eri lähteistä, tieto voi tulla eri formaateissa ja samasta tapahtumasta voidaan saada havaintoja useista eri lähteistä. Eri lokien tunnistamista samaan tapahtumaan liittyväksi on toteutettu koneoppimista hyödyntäen, millä on onnistuttu parantamaan joidenkin hyökkäystyyppien tunnistamista ja vähentämään väärien positiivisten määrää (Stroeh, ym., 2013).

Splunkin tapauksessa koneoppisen hyödyntäminen jo esiprosessointivaiheessa on mahdollista Machine Learning Toolkitin avulla. Siihen sisäänrakennetut algoritmit perustuvat pääsääntöisesti Scikit-learn python-kirjastoon (Splunk, 2021) (Scikit-learn project, 2021) .

Rikastamisessa lokitapahtumalle luodaan konteksti lisäämällä siihen tietoa ulkoisesta lähteestä. Koneoppimista on hyödynnetty ohjelmistossa, joka hakee automaattisesti tietoa kyberuhkista niin kutsutusta dark netistä ja deep netistä foorumiviesteistä ja verkossa tapahtuvasta haittaohjelmakaupankäynnistä (Nunes, E., ym, 2016). Kyseistä tietoa voitaisiin hyödyntää SIEM-järjestelmässä, vaikka Arizonan yliopiston tutkimusryhmä, joka on toteuttanut kyseisen sovelluksen, ei näin tehnytkään.

4.4 Lokidatan analysointi SIEM-järjestelmässä koneoppimisen avulla

SIEM-järjestelmään kerättävä data muodostaa niin kutsutun big data -tietokannan. Suuri osa koneoppimista SIEM-järjestelmässä hyödyntävistä menetelmistä liittyy järjestelmän tietokantaan tallennetun tiedon automaattiseen analysointiin ja tiedonlouhintaan kyberuhkiin viittaavaan toiminnan tunnistamiseksi. Tällä pyritään muun muassa järjestelmää käyttävän analyytikon työmäärän vähentämiseen sekä tunnistamaan lokitapahtumista asioita, joiden havaitseminen pelkästään ihmisen toimesta olisi todella haastavaa tai mahdotonta. Ideaalitilanteessa haitallisen toiminnan jäljille voidaan päästä ennen kuin haittatoimija on saanut vielä vahinkoa aikaiseksi. Tällöin lokitapahtumien nopealla analysoinnilla ja tunnistamisella on merkitystä. Tässä luvussa on kuvattu tarkemmin joitakin lokidatan analysoinnissa käytettyjä koneoppimismenetelmiä.

4.4.1 Numeerinen poikkeamahavainnointi

Kohtuullisen helposti toteutettavia koneoppimista hyödyntäviä tiedonlouhinnan menetelmiä ovat ohjaamattomaan oppimiseen perustuvat poikkeamahavainnointimenetelmät, joiden avulla numeerista dataa voidaan analysoida koneoppimisen avulla poikkeamien havaitsemiseksi. Yksi käytännön sovellus poikkeamahavainnoinnista SIEM-järjestelmässä on valvottavaan järjestelmään

tehtävien kirjautumistapahtumien tai niiden yritysten käsittely ohjaamatonta koneoppimista hyödyntäen. Analyysissa hyödynnetään Learning with Label Proportions (LLP) menetelmää kirjautumistapahtumien historian perusteella odotetun vaihteluvälin määrittämiseksi, mihin normaalitilanteessa kirjautumistapahtumien määrän tulisi mahtua. Mikäli kirjautumistapahtumien määrä on vaihteluvälin ulkopuolella, tulkitaan se poikkeavaksi tapahtumaksi. Tällä pyritään tunnistamaan poikkeavaa toimintaa järjestelmässä. Menetelmän etuna on, että poikkeava kirjautumismäärä voidaan määritellä tilastollisesti perustuen historiaan sen sijaan, että järjestelmän käyttäjä joutuisi itse määrittämään raja-arvot poikkeavien tapahtumien määrälle.

Tietoturvauhkien tunnistamisessa ainoastaan kirjautumistapahtumien määrään analysointiin käytettäessä menetelmä on ongelmallinen, koska varsinkin suuressa valvottavassa järjestelmässä tapahtumien määrän vaihtelu ajan suhteen voi olla suurta ihan syystäkin. Esimerkiksi, mikäli menetelmässä käytetään historiatietona edellisen kuukauden tilannetta, voikin esimerkiksi ainoastaan vuosittain esiintyvä juhlapyhä tai muu harvinainen tapahtuma aiheuttaa luonnollisen poikkeaman kirjautumisten määrään, mitä koneoppimisen keinoin ei voida ennustaa. Lisäksi määrällisen poikkeaman esiintyminen ei paljasta vai tapahtuman juurisyytä, vaan se tulee selvittää aina erikseen. Kuviossa 5 on esimerkki numeerisesta poikkeamahavainnoinnista järjestelmään tehdyistä kirjautumistapahtumista niiden lukumäärän perusteella Splunkissa. Vihreät pisteet ilmaisevat tunnit, jolloin kirjautumisten määrä poikkeaa koneoppimisen avulla määritellystä poikkeama-arvon (engl. residual) odotetusta vaihteluvälistä (vaalen sininen värjätty alue). Esimerkkidatassa on kirjautumistapahtumat yhden kuukauden ajalta.

KUVIO 5 Splunkin visualisointi numeerisesta poikkeamahavainnoinnista

Yksi tapa hyödyntää määrällistä poikkeamahavainnointia on käsitellä jo jollakin muulla menetelmällä käsiteltyä tietoa. Esimerkiksi IDS järjestelmän tuottamien hälytysten lukumäärän analysointi yllä kuvatulla menetelmällä voisi auttaa SOC-operaattoria rajaamaan kiinnostavat tapahtumat niiden esiintyvyyden perusteella.

4.4.2 UEBA

Esimerkki kehittyneemmästä koneoppimisen hyödyntämisestä on käyttäjätoiminnan ennustamiseen pyrkivät User and Entity Behaviour Analytics (UEBA) koneoppimismenetelmä. Menetelmän tarkoituksena on analysoida järjestelmän käyttäjien toiminnasta SIEM-järjestelmään tallennettavia lokitapahtumia, esimerkiksi järjestelmään kirjautumisia, ja tunnistaa käyttäjän toiminnasta poikkeamat, jotka voivat viitata tunnusten päätymistä haittatoimijan käsiin tai sisäisen uhkan mahdollisuudesta, eli järjestelmän käyttäjä esimerkiksi vie luvattomasti dataa organisaation järjestelmästä. Mitre Defend -mallissa käyttäjätoiminnan havainnointi muodostaa yhden luokan D3-UBA, johon kuuluu kuviossa 6 esitetyt 12 eri havaitsemismenetelmää (The MITRE Corporation, 2022).

Technique Subclasses

There are 12 countermeasure techniques in this category, ***User Behavior Analysis****.*

Name	ID	Definition
User Behavior Analysis	D3-UBA	Analysis of user behavior and patterns for the purpose of detecting unauthorized user activity.
- Job Function Access Pattern Analysis	D3-JFAPA	Detecting anomalies in user access patterns by comparing user access activity to behavioral profiles that categorize users by role such as job title, function, department.
- Local Account Monitoring	D3-LAM	Analyzing local user accounts to detect unauthorized activity.
- Authentication Event Thresholding	D3-ANET	Collecting authentication events, creating a baseline user profile, and determining whether authentication events are consistent with the baseline profile.
- Authorization Event Thresholding	D3-AZET	Collecting authorization events, creating a baseline user profile, and determining whether authorization events are consistent with the baseline profile.
- Credential Compromise Scope Analysis	D3-CCSA	Determining which credentials may have been compromised by analyzing the user logon history of a particular system.
- Domain Account Monitoring	D3-DAM	Monitoring the existence of or changes to Domain User Accounts.
- Resource Access Pattern Analysis	D3-RAPA	Analyzing the resources accessed by a user to identify unauthorized activity.
- Session Duration Analysis	D3-SDA	Analyzing the duration of user sessions in order to detect unauthorized activity.
- User Data Transfer Analysis	D3-UDTA	Analyzing the amount of data transferred by a user.
- User Geolocation Logon Pattern Analysis	D3-UGLPA	Monitoring geolocation data of user logon attempts and comparing it to a baseline user behavior profile to identify anomalies in logon location.
- Web Session Activity Analysis	D3-WSAA	Monitoring changes in user web session behavior by comparing current web session activity to a baseline behavior profile or a catalog of predetermined malicious behavior.

KUVIO 6 UEBA-menetelmässä hyödynnettäviä havaitsemistaktiikoita (The MITRE Corporation, 2022)

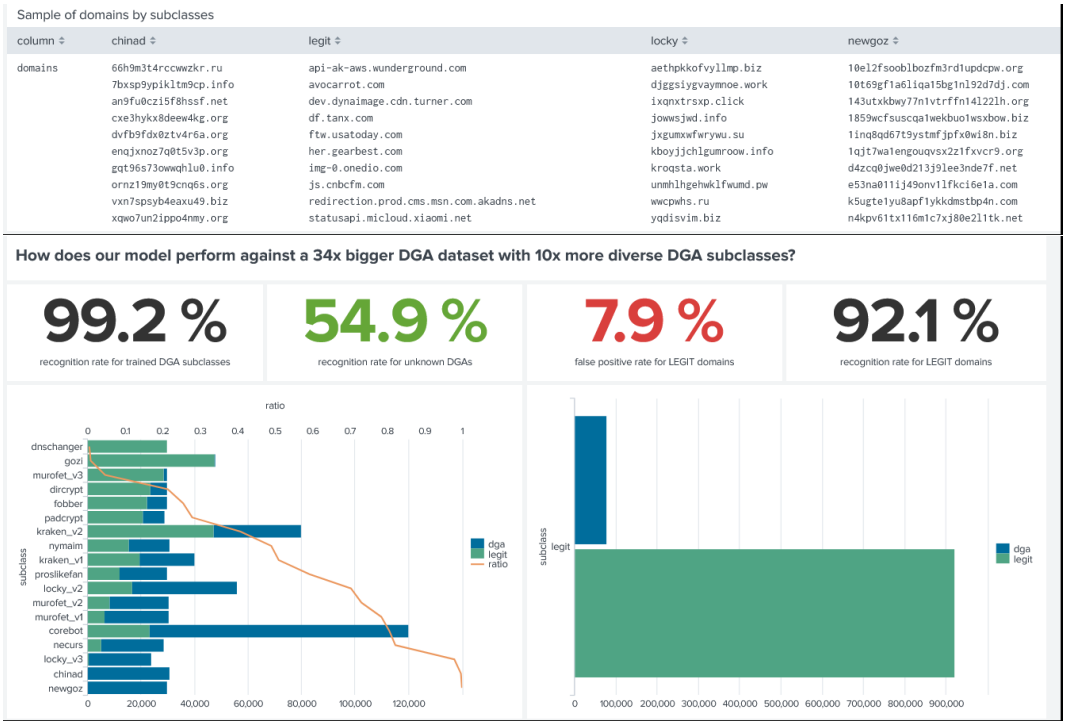
UEBA voidaan toteuttaa esimerkiksi siten, että jokaiselle lokitapahtumalle määritellään koneoppimisen avulla riskiarvo perustuen käyttäjän aikaisempaan toimintaan. Sitä korkeampi riskiarvo, mitä enemmän lokitapahtuman tiedot poikkeavat käyttäjän aikaisemmista tapahtumista esimerkiksi järjestelmään kirjautumisajankohdan tai kirjautumis-IP-osoitteen maantieteellisen sijainnin perusteella. UEBA tuotteita on saatavilla useisiin SIEM-järjestelmiin, myös Splunkiin, mutta niiden tarkasta toimintaperiaatteesta tai käytetyistä koneoppimismenetelmistä ei ole juurikaan tietoa saatavilla. (Logpoint, 2022)

4.4.3 DGA havainnointi

Esimerkki ohjatun koneoppimisen hyödyntämisestä kyberuhkien havainnoinnissa on Domain Generation Algorithm (DGA) -tunnistus. DGA: lla tarkoitetaan

koneellisesti luotua (generoitua) domain-tunnusta, joissa sijaitsevalle haitalliselle verkkosivulle voidaan uhriksi joutuneet käyttäjät ohjata esimerkiksi komentoyhteyden luomiseksi uhrikoneelle. Kyseinen hyökkäystaktiikka on Mitre Attack-mallissa ID T1637.001 ja kuuluu komentokanavan muodostamistaktiikoiden alaluokkaan (The MITRE Corporation, 2021).

Tunnistamalla valvottavasta järjestelmästä yhteyden DGA-osoitteeseen, voidaan päästä mahdollisen hyökkäyksen tai sen yrityksen jäljille. DGA-osoitteiden havaitsemiseksi valvottavasta järjestelmästä tulisi kerätä tapahtumatietoa järjestelmästä tehtävistä DNS-kyselyistä. Kuviossa 7 on esitettynä DGA-tunnistuksessa käytettävää opetusdataa Splunkin DGA Analysis -liitännäisessä. Näkymässä on esimerkkejä haitallisista luokista, jotka on nimetty: chinad, locky ja newgoz. Turvalliset domainit on merkitty luokkaan legit.



KUVIO 7 DGA tunnistuksessa käytettävää opetusdataa

4.4.4 Haitallisten sähköpostien tunnistaminen

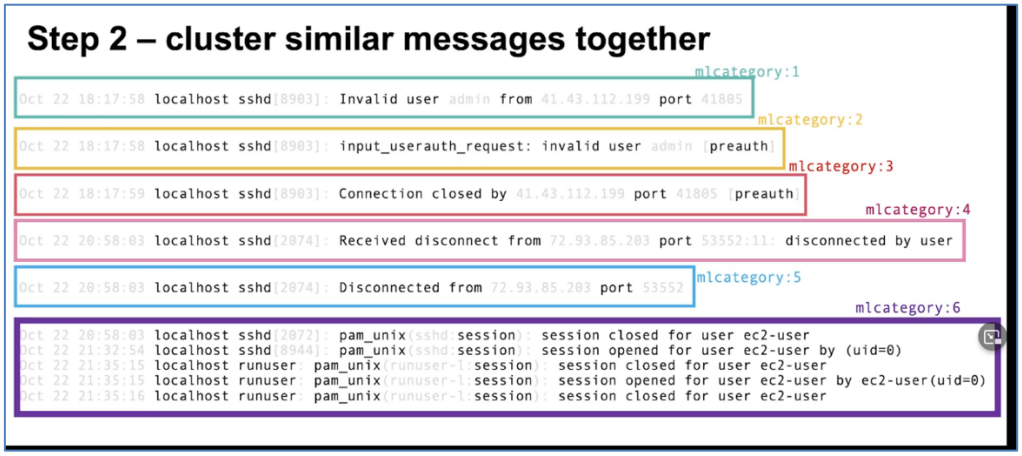
DGA havainnoinnin lisäksi tekstimuotoisen tiedon analysointia koneoppimisen avulla voidaan hyödyntää esimerkiksi roskapostien tai muuten haitallisten sähköpostien tunnistamisessa (Crawford M., Khoshgoftaar Taghi M., Prusa J. D. ,Richter A. N., Al-Najada H., 2015).

Tällöin opetusdatana käytetään aineistoa sisältäen esimerkkidataa roskaposteista sekä ei-roskaposteista. Opetusdatassa näiden tapahtumien luokka (spam tai non-spam) tulee olla merkittynä. Vaikka menetelmän toteuttaminen teknisesti olisikin toteutettavissa SIEM-järjestelmän, esimerkiksi Splunkin, toiminnallisuuksin, menetelmä on ongelmallinen, koska järjestelmän käyttäjien sähköpostiliikenne on luottamuksellista, eikä viestien sisällön näkymistä SIEM-

järjestelmässä tulisi sallia. Myös tunnistetun haitallisen sähköpostin välityksen estäminen tai poistaminen on käyttäjien yksityisyyden suojan vuoksi haastavaa.

4.4.5 Lokianalyysi NLP-menetelmällä

DGA tunnistuksessa sekä haitallisten sähköpostien tunnistamisessa hyödynnetään luonnollisen kielen prosessoinnin (NLP) menetelmiä. Vastaavasti NLP-menetelmällä voidaan käsitellä SIEM-järjestelmään tallentuvien lokitapahtumien tekstisisältöä mahdollisesti haitallisen toiminnan tunnistamiseksi. Järjestelmään tunkeutuvan hyökkääjän toiminnasta jää usein lokimerkintöjä järjestelmään. Haastavaa lokianalyysissa kuitenkin on, että itse lokin tekstisisällöstä ei voi sanoa, liittyykö kyseinen tapahtuma järjestelmässä haitalliseen toimintaan, vai onko kyse tarkoituksellisesta toiminnasta. Tästä syystä merkittyä opetusdataa haitallisista lokeista on haastava tuottaa. Lokianalyysia voidaan toteuttaa kuitenkin ohjaamatonta koneoppimista hyödyntäen luomalla koneoppimismalli järjestelmään tallennetuista lokitapahtumista perustuen siinä olevien sanojen esiintyvyyteen ja merkitsemällä tapahtumille luokka sanayhdistelmien esiintyvyyden mukaan. Tällä tavoin pyritään jakamaan samankaltaiset tapahtumat omiin luokkiinsa klusterointialgoritmia hyödyntäen. Luokkien, eli klustereiden lokitapahtumien lukumäärän muutoksia seuraamalla voidaan havainnoida suuresta lokimäärästä järjestelmässä mahdollisesti tapahtuvia muutoksia. Menetelmään voidaan yhdistää numeerinen poikkeamatunnistus, jolloin kunkin klusterin lokitapahtumien määrän muutosten analysoinnissa voidaan hyödyntää myös koneoppimista.



KUVIO 8 Esimerkki tekstimuotoisen lokidatan jakamisesta luokkiin Elastic- sovelluksessa Machine Learning -toimintoa hyödyntäen

5 KONEOPPIMISTA HYÖDYNTÄVÄN SIEM-JÄRJESTELMÄN SUUNNITTELU JA TOTEUTUS

Tutkimuksen ensimmäisessä osassa tarkasteltiin esimerkkejä erilaisista koneoppimistoiminnallisuuksista SIEM-järjestelmiin. Kirjallisuudesta ei ole tutkimuksen perusteella saatavilla yksityiskohtaisia kuvauksia sopivan koneoppimismallin toteutuksesta esimerkiksi Splunkiin, joten oman konstruktion, koneoppimista hyödyntävän SIEM-järjestelmän, kehitys toteutettiin vaihe vaiheelta tiedonlouhinnan ja koneoppimisen suunnittelu- ja käyttöönottoprosesseja hyödyntäen. Tässä luvussa kuvataan yksityiskohtaisesti tutkimuksessa kehitetyn konstruktion kehityksen perusteet ja mallin kehitys vaiheittain.

5.1 Järjestelmän suunnittelu- ja kehitysprosessi

Tiedonlouhinnan ja data-analyysin kehitysprosessin kuvaamiseksi on laadittu malleja kuvaamaan data-analyysiprosessia hyödynnettävän datan tunnistamisesta aina tuotantoon saakka ja mahdollistamaan mahdollisimman sujuva ja yhtenäinen kehitysprosessi. (Wiley, 2015) on yhdistänyt eri tiedonlouhinnan ja data-analyysin prosesseja, muun muassa CRISP-DM prosessia ja todennut, että yleisesti data-analyysiprosessiin kuuluu vaiheet: tarpeiden tunnistaminen, datan valmistelu, mallin suunnittelu, mallin kehitys, tulosten arviointi, mallin hyödyntäminen.

KUVIO 9 Data-analyysiprosessi (Wiley, J. and Sons, 2015)

Prosessia ja sen eri vaiheiden sisältöä mukaillen tämän tutkimuksen esimerkkijärjestelmän suunnittelussa edettiin seuraavien vaiheiden mukaan:

1. Tavoitteiden ja reunaehtojen asettelu
2. Datan valmistelu
3. Koneoppimismallin luominen
4. Tulosten evaluointi ja mallin kehittäminen
5. Menetelmän hyödyntäminen

5.2 Tavoitteet ja reunaehdot toteutettavalle konstruktiolle

Data-analyysiprosessin ensimmäisessä vaiheessa tunnistetaan ja määritetään tarpeet data-analyysiprojektille sekä muodostetaan näkemys, mitä projektilla halutaan saada aikaiseksi. (Wiley, J. and Sons, 2015) Tässä tutkimuksessa toteutettavan konstruktion yksityiskohtaisessa tavoitteiden määrittelyssä käytettiin pohjana tutkimukseen laadittua kirjallisuuskatsausta koneoppimisen hyödyntämismahdollisuuksista, mitä tarkasteltiin yhdessä tutkimuksen toimeksiantajan kanssa. Potentiaalisimmaksi ja kiinnostavimmaksi käyttötapaukseksi poikkeavien lokitapahtumien louhiminen koneoppimisen menetelmiä hyödyntäen. Organisaation Linux-palvelinten lokidataa on saatavilla todella paljon, eikä sen

analysoinnille ole ollut juurikaan resursseja, joten kyseisestä menetelmästä arvioitiin saavutettavan hyötyä.

Aikaisemmin luvussa 3.2 käsiteltiin haittatoimijan taktiikoita, joita on mahdollista havaita mm. järjestelmälokien perusteella. Näin ollen kehitettävän konstruktion päämääränä on tukea lokidatan analyysia ja auttaa erottamaan lokitapahtumista ne mahdolliset tapahtumat, jotka voivat indikoida haitallista toimintaa. Muiden vaatimusten osalta rajattiin jo lähtökohtaisesti, että toteutuksen tulee olla toteutettavissa Splunkin omilla työkaluilla ilman mittavia uudelleenasennuksia tai konfiguraatioita itse Splunkin perusasennukseen.

Lisäksi toteutuksessa tulee huomioida koko lokidatan käsittelyketju aina tiedon esittämiseen saakka Splunkissa, mistä menetelmällä käsiteltävä tieto voidaan tulkita ja hyödyntää. Lisäksi järjestelmään tulevaa lokidataa tulee käsitellä reaaliajassa ja mallin tulee toimia testijärjestelmän palvelinresursseilla, jotta se olisi toteuttamiskelpoinen.

Kokonaisuudessaan vaatimukset toteutettavalle esimerkkijärjestelmälle ovat seuraavat:

1) Menetelmä on Splunkin työkaluilla ja resursseilla toteutettavissa
2) Käsitellään järjestelmään tuotavaa dataa ilman muutoksia
3) Tulokset on esitettävissä Splunkin näkymissä
4) Järjestelmä käsittelee tulokset reaaliajassa
5) Järjestelmän resurssit riittävät koneoppimismallin käsittelyyn

Koneoppimismallin toiminnan arviointi esimerkiksi kyberuhkien tunnistamistarkkuuden perusteella todettiin lähtökohtaisesti haastavaksi toteutettavassa testiympäristössä. Ennustetarkkuuden mittaaminen vaatisi, että käytössä olisi tieto aidosta haitallisesta toiminnasta, mikä edellyttäisi hyökkäysten tekemistä kohdejärjestelmään, mutta käytetyssä kohdeympäristössä tämä ei ole käytännössä mahdollista. Järjestelmän arvioinnissa tullaankin keskittymään laadulliseen arvioon, jossa tarkastellaan menetelmän tekninen toimivuus kohdeympäristössä sekä järjestelmän hyödynnettävyys kyberuhkien havaitsemisessa.

5.3 Datan esivalmistelu

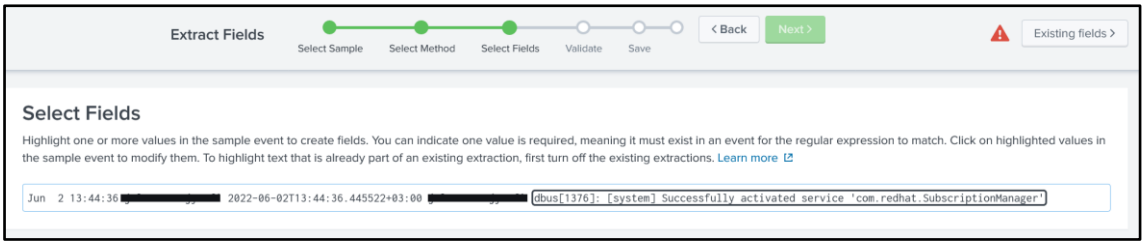
Esimerkkitoteutuksessa käytettävälle tutkimuksen tilaajan Splunk alustalle oli jo valmiiksi toteutettu automaattinen Linux lokien keräys valvottavan ympäristön Linux-palvelimilta. Kyseiset syslog-tapahtumat tallentuvat samaan indeksiin, mutta data on valmiiksi luokiteltu esimerkiksi palvelimen nimen tai lokeja tuottavan Linux -prosessin perusteella. Esimerkkijärjestelmän datavolyymi on n. 1 500 lokitapahtumaa minuutissa.

Kuviossa 10 on esimerkki Linux-palvelimen tuottamasta lokitapahtumasta, joka kertoo SSH-hallintayhteyden muodostumisesta Linuxin ssh-palveluun käyttäjälle « test » tiettynä ajanhetkenä.

_time	process	message
2022-07-01 05:02:26	sshd(pam_unix)	sshd(pam_unix)[21692]: session opened for user test by (uid=509)

KUVIO 10 Esimerkki lokitapahtumasta

Oleellinen osa data esivalmistelua koneoppimista varten on erottaa koneoppimista varten käsiteltävä lokikenttä. Tässä tapauksessa haluttiin yksittäisen tapahtuman koko tekstimuotoinen merkkijono käsiteltäväksi. Splunkin ominaisuuksiin kuuluu mahdollisuus erottaa koneoppimisen avulla käsiteltävä kenttä esimerkiksi säännöllisen lausekkeen avulla.



KUVIO 11 Käsiteltävän lokikentän määrittely Splunkissa

Kuviossa 11 on esitetty käsiteltävän lokikentän määrittely Splunkissa automaattisen työkalun avulla, joka muodostaa säännöllisen lausekkeen lokitapahtuman tekstikentän parsimiseksi.

5.4 Koneoppimismallin suunnittelu ja kehitys

Seuraavassa, kolmannessa vaiheessa, koneoppimismallin kehityksessä kartoitetaan, mitä koneoppimismenetelmiä, kuten luokittelua, klusterointia tai regressiota, työssä voidaan hyödyntää tavoitteiden saavuttamiseksi. Samalla määritetään, käytetäänkö tehtävän toteuttamisessa ohjattua vai ohjaamatonta koneoppimista. (Wiley, 2015)

Tutkimuksessa haluttiin toteuttaa koneoppimista hyödyntävä menetelmä lokitapahtumien sisällön analysoimiseksi edellisessä luvussa esitetyn esimerkkikäyttötapauksen mukaisesti, joten kehityksessä keskityttiin luonnollisen kielen prosessoinnin mahdollisuuksiin Splunkissa.

Splunkiin on toteutettu käyttäjäyhteisön toimesta liitännäisiä, joista yksi on nimeltään NLP Text Analytics, joka on Splunkin Enterprise -lisenssiin kuuluvaa Machine Learning Toolkit:a (MLTK), hyödyntävä -liitännäinen, joka on kehittäjänsä oma tutkimusprojekti (Worsham, 2018). Liitännäinen sisältää käyttöliittymän Splunkiin, jonka avulla Splunkiin tallennettua tekstimuotoista dataa voidaan tarkastelle erilaisia luonnollisen kielen prosessointimenetelmiä hyödyntäen.

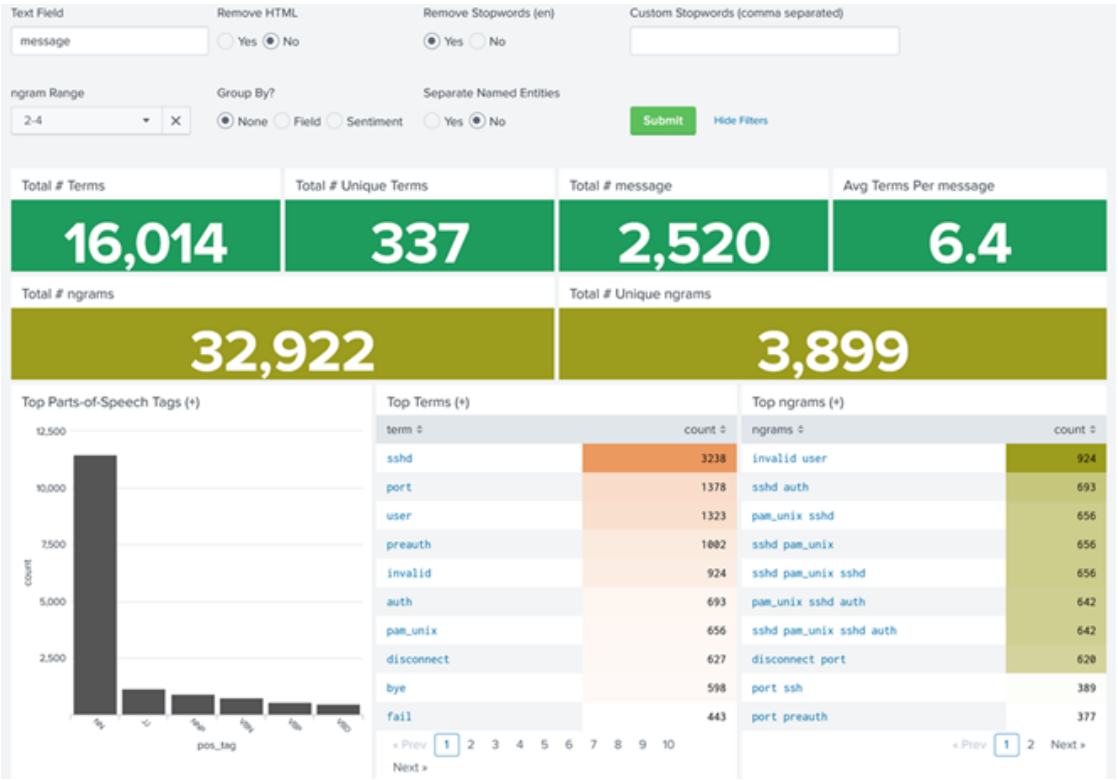
Tarkempi tarkastelu osoittaa, että NLP-liitännäisessä tekstin luokittelussa hyödynnetään muun muassa klusterointia, joka on yksi ohjaamattoman koneoppimisen menetelmä, jonka avulla teksti voidaan jakaa ryhmiin, missä sisällöltään

samankaltaiset tekstikentät jaetaan omiin ryhmiinsä (Wiley, 2015). Vaikka NLP-liitännäinen ei sellaisenaan sovellu SIEM-käyttöön, voidaan samoja koneoppimismenetelmiä hyödyntää Splunkissa ilman kyseistä liitännäistä tai sen käyttöliittymää, jolloin menetelmä voidaan integroida osaksi Splunkin SIEM-toimintoja ja -näkymiä.

5.4.1 TFIDF

Klusteroinnin etuna on muun muassa, että opetusdatana ei tarvita merkittyä opetusdataa, vaan järjestelmän käyttämä koneoppimismalli opetetaan järjestelmään aikaisemmin tallennetulla lokidatalla. Koska klusterointialgoritmilla ei voida kuitenkaan käsitellä suoraan tekstimuotoista dataa, tulee data käsitellä ensiksi term frequency-inverse document frequency (TF-IDF) menetelmällä. TF-IDF on tilastollinen menetelmä, jonka avulla tunnistetaan tekstistä sanat sekä annetaan sanoille numeroarvo niiden esiintyvyystiheyden mukaan. (Wiley, 2015)

Splunkissa toimenpide voidaan tehdä MLTK:aan sisältyvällä TF-IDF -ominaisuudella, joka hyödyntää Scikit-Learn-kirjaston TfidfVectorizer:a sanojen ja sanayhdistelmien pisteyttämiseksi niiden esiintyvyyden mukaan. Kuviossa 12 on TF-IDF algoritmin avulla selvitetyt yleisimmät sanat ja sanayhdistelmät esimerkkilokidatasta sanayhdistelmien rajauksen (engl. ngram range) ollessa 2–4



KUVIO 12 Yleisimmät sanat ja sanayhdistelmät esimerkkilokidatasta.

TFIDF-algoritmille on annettava parametreina käsiteltävien sanayhdistelmien määrä, ominaispiirteiden (featureiden) maksimimäärä. Lisäksi määritellään mallille nimi, jolla se tallennetaan Splunkiin käyttöönottoa varten. TF-IDF-mallin opetuskomento Splunkissa voisi olla esimerkiksi seuraava:

« fit TFIDF message max_features=1000 ngram_range=2-4 into mallinimi »

TF-IDF mallin sovittaminen tuottaa asetetun ominaispiirteiden määrän mukaisesti uudet lokikentät. Kukin lokitapahtuma saa kyseiseen kenttään arvon sen mukaan, kuinka hyvin tapahtuma sopii kyseisiin ominaispiirteisiin. Kuviossa 13 on esimerkki yhdestä TF-IDF-menetelmän määrittämästä uudesta kentästä *message_tfidf_227_22151 failed password for*. Tarkasteluaikana viisi kertaa toistunut tapahtuma, jossa esiintyy sanayhdistelmä *failed password for* sekä SSHD-prosessin numero *22151* saavat kyseiseen kenttään arvon 0.20733149566258222. Vastaavasti tapahtuma, jossa ei ole kyseistä sanayhdistelmää, saa arvon 0.

sourcetype	process	message	message_tfidf_227_22151 failed password for
syslog	sshd	sshd[22151]: Failed password for root from 61.177.173.40 port 40058 ssh2	0.20733149566258222
syslog	sshd	sshd[22151]: Failed password for root from 61.177.173.40 port 40058 ssh2	0.20733149566258222
syslog	sshd	sshd[22151]: Failed password for root from 61.177.173.40 port 40058 ssh2	0.20733149566258222
syslog	sshd	sshd[22151]: Failed password for root from 61.177.173.40 port 40058 ssh2	0.20733149566258222
syslog	sshd	sshd[22151]: Failed password for root from 61.177.173.40 port 40058 ssh2	0.20733149566258222
syslog	sshd	sshd[22822]: Disconnected from 138.197.138.123 port 54952 [preauth]	0.0

KUVIO 13 Esimerkki TF-IDF-käsittelyssä luodusta lokikentästä ja sille annetuista numeroarvoista.

5.4.2 Aihemallinnus

Ennen varsinaista klusterointia, voidaan TF-IDF-käsittelyn tuloksesta määritellä usein yhdessä esiintyvät sanat ja sanayhdistelmät aihemallinnuksen (engl. topic

modeling) avulla, jolloin klusteroinnin tuloksessa saadaan paremmin eroteltua samankaltaista sisältöä edustavat ryhmät varsinkin suuresta datamäärästä (Wiley, 2015). Aihemallinnukseen soveltuu esimerkiksi Latent Dirichlet Allocation (LDA) -algoritmi (Wiley, 2015). Tämän lisäksi

Splunkin NLP-liitännäisessä aihemallinnuksessa on hyödynnetty vaihtoehtoisesti myös TruncatedSVD (engl. Singular Value Decomposition) sekä Non-negative matrix factorization (NMF) -algoritmeja (Worsham, 2018). Algoritmilla pyritään ulottuvuuksien vähentämiseen (engl. dimension reduction) TF-IDF-menetelmällä käsitellystä, moniulotteisesta datasta. Splunkissa esimerkiksi truncatedSVD -algoritmi sovitetaan dataan seuraavalla komennolla, missä määritellään käsiteltävä datakenttä, nimi tallennettavalle mallille, sekä algoritmin k:n arvo, jolla määritellään ulottuvuuksien määräkäsiteltävälle datall , minkä tulee olla vähemmän kuin aikaisemmin TF-IDF :lle määritelty ominaisuuksien (max_features) määrä:

« fit TruncatedSVD k=500 "message_*" into mallinimi »

5.4.3 Klusterointi

Klusterointi on ohjaamattoman koneoppimisen menetelmä samanlaisen datan luokittelemiseksi omiksi ryhmikseen. Koneoppimisessa menetelmän päämääränä on muun muassa löytää vaikeasti havaittavia rakenteita tai säännönmukaisuuksia datasta. (Wiley, 2015)

Lokidatan luonnollisen kielen prosessoinnissa klusteroinnilla käsitellään aikaisemmin TF-IDF-menetelmällä sekä aihemallinnuksen menetelmin prosessoitu data, jolloin klusteroinnin tuloksena kullekin lokitapahtumalle määritellään käytettävän algoritmin perusteella klusterinumero, joka määrittelee, mihin ryhmään kyseinen tapahtuma kuuluu. Samaan klusterinumeron saavat tapahtumat ovat täten samankaltaisia ja lokitapahtumia voidaan tarkastella klusterinumeron perusteella aikaisemmin luvussa 4 esitetyn esimerkin tavoin.

Todennäköisesti suosituin klusterointimenetelmä on kMeans, jossa klusteroinille määritetään arvo k, jolla määritellään ennalta klustereiden määrä, johon data halutaan jaettavan (Wiley, 2015).

Splunkissa NLP-käsittelyssä voidaan hyödyntää kmeans-klusterointia, mutta vaihtoehtoisesti myös muun muassa Birch tai Spectral Clustering -algoritmeja, joille ei kmeans:n tavoin määritellä ennakolta klustereiden lukumäärä. Klusterointi voidaan tehdä myös määrittämättä ennakolta klusterimäärää käyttäen xmeans tai Density-Based Spatial Clustering of Applications with Noise (DBSCAN) -klusterointimenetelmiä, jotka arvioivat itse datasta muodostettavan optimaalisen klustereiden lukumäärän. Näiden suhteen on kuitenkin huomioitava, että dbscan algoritmia käytettäessä koneoppimismallin tallentaminen Splunkiin ei ole mahdollista, mikä rajoittaa sen käytännön hyödyntämistä. (Splunk Inc: Algorithms in the Machine Learning Toolkit, 2022)

Splunkissa klusterointi voidaan toteuttaa esimerkiksi seuraavien komentojen avulla, jolloin komennossa annetaan käytettävästä algoritmista riippuen k:n

arvo sekä esimerkiksi truncatedSVD-algoritmin avulla luodut lokikentät arvolla « SVD_* »:

« fit KMeans k=10 "SVD_*" into mallinimi »
« fit XMeans "SVD_*" into mallinimi »

5.5 Mallin toiminnan arviointi

Ennen koneoppimismallin käyttöönottoa, arvioidaan kehitetyn mallin toimivuus ja arvioidaan, toteuttaako se halutun tehtävän ja saavutetaanko sillä haluttu tavoite (Wiley, 2015). Kehitetty koneoppimismalli muodostuu kaikkiaan kolmesta vaiheesta:

1) Tekstikentän muuttaminen numeeriseksi dataksi TF-IDF-algoritmin avulla
2) Aihemallinnus
3) Klusterointi

Klusterointituloksen arvioinnissa ja mallin kehittämisessä voidaan hyödyntää seuraavia periaatteita (Wiley, 2015):

- Poikkeaako eri klustereiden tulokset selvästi toisistaan, eli klustereiden välillä on selvä ero?
- Jakaantuuko joihinkin klustereihin vain pieni määrä dataa?
- Onko joidenkin klustereiden tulokset (data) liian lähellä toisiaan?

Testatessa kehitettyä mallia esimerkkilokidatalle eri parametreilla ja käytettäessä klusterointialgoritmina xmeans:a, eli klusterimäärää ei määritelty etukäteen, vaan algoritmi määritti sen automaattisesti, jakaantui data lähes kaikissa tilanteissa kymmeniin tai satoihin klustereihin. Tulos ei tuottanut haluttua hyötyä, koska saadun tuloksen perusteella ei erotettu datasta säännönmukaisuuksia halutulla tavalla. Tämän vuoksi selvitettiin myös k-means klusterointialgoritmin toimintaa. Määrittämällä algoritmin parametriksi k=10, eli algoritmi muodostaa datasta kymmenen klusteria saavutettiin tulos, jossa yhdeksän klusteria muodostui tekstirakenteeltaan selvästi yhdenmukaisista lokitapahtumista ja yksi klustereista sisälsi selvästi muusta datasta poikkeavia tapahtumia ja tässä joukossa oli myös keskenään poikkeavia tapahtumia.

Kmeans klusteroinnilla saadut tulokset vastasivat melko hyvin päämäärää, mitä koneoppimismallilla haluttiin saavuttaakin: malli jakaa lokitapahtumat ryhmiin, joiden perusteella voitiin muodostaa kokonaiskuva käsiteltävästä datasta sekä saatiin eroteltua ryhmä, joka poikkeaa muusta datasta ja muodostaa joukon, josta todennäköisimmin voitaisiin havaita poikkeava toiminta.

5.6 Mallin käyttäminen

Koneoppimismallin kehitysprosessin viimeinen vaihe on mallin sovittaminen tuotantojärjestelmään ja mallin käyttöönotto. Käyttöönotossa tulee myös osoittaa järjestelmän toimivuus ja lisäarvo sen käyttäjille heidän ymmärtämällään tavalla (Wiley, 2015). Järjestelmän toteutuksen tavoitteiksi asetettiin aiemmin, että menetelmän tule käsitellä sisään tuleva lokidata reaaliajassa sekä esittää käsitelty data tulokset Splunkin näkymissä. Splunkissa koneoppimismallin käyttö toteutetaan liittämällä hakukomentoon datan rajausehdon lisäksi apply- komento, jolla viitataan aikaisemmin fit-komennolla opetettuihin malleihin, esimerkiksi alla oleva komento käsittelee valitun aikarajuksen mukaisesti kaikki syslog-tyyppiset lokit aikaisemmin opetetuilla kolmella koneoppimismallilla:

« sourcetype=syslog | apply tfidf_model | apply svd_model | apply kmeans_model »

Apply -komento voi käyttää kuten mitä tahansa hakua Splunkissa. Splunkin visualisoinnissa kullekin kuvaajalle, taulukolle tms, määritellään hakukomento siinä esitettävälle datalle, joten apply -komento voidaan siten upottaa myös visualisointeihin ja esittää koneoppimismallin käsittelemä data.

6 JÄRJESTELMÄN DEMONSTROINTI

Tutkimuksen päämääränä on osoittaa esimerkkiratkaisun tavoin, kuinka koneoppimista voidaan hyödyntää SIEM-järjestelmissä. Tässä luvussa esitellään tutkimuksessa kehitetyn menetelmän toiminta yksityiskohtaisesti esimerkein, mikä vastaa DSRM tutkimusmenetelmän vaihetta 4. Demonstraatiossa käytettävä järjestelmä on toteutettu Jyväskylän yliopiston Splunk -koulutusinstanssissa keväällä 2022 käyttäen järjestelmään kerättyä lokidataa, joka sisältää tapahtumia joistakin yliopiston käytössä olevilta palvelimilta.

Esimerkissä luonnollisen kielen prosessointia hyödyntävää koneoppimismenetelmää käytetään selvittämään Splunkin tietokantaan tallentuvista lokitapahtumista niiden tekstisisällön perusteella poikkeavat tapahtumat, jotka voivat mahdollisesti merkitä järjestelmän poikkeavaa toimintaa tai haitallista toimintaa.

6.1 Koneoppimismallin opettaminen

Menetelmän käyttöönoton ensimmäisessä vaiheessa opetetaan ja tallennetaan koneoppimismalli halutusta datatyypistä valitulta aikajaksolta. Koneoppimismallin opettaminen tehdään Splunkin fit-hakukomennolla ja sen lisämääreillä. Kokonaisuudessaan komento on seuraavanlainen:

```
sourcetype=syslog
| fit TFIDF message max_features=1000 ngram_range=1-4 into tfidf_model
| fit TruncatedSVD k=100 "message_*" into svd_model
| fit KMeans k=10 "SVD_*" into kmeans_model
```

Komennon «sourcetype » arvo määrittää käsiteltävän lokityypin, tässä tapauksessa kaikki Splunkin syslog-tyyppiset lokitapahtumat. Tarvittaessa tässä osiossa voidaan rajata käsiteltävää tapahtumajoukkoa esimerkiksi tiettyihin lokilähteisiin.

Komennon fit TFIDF-osalla käsitellään datan « message » -kenttä TFIDF-algoritmillä ja määritellään ominaisuuksien määräksi 1000 sekä käsiteltävien sanayhdistelmien määräksi, eli ngram-arvoksi 1-4. into -komento puolestaan määrittää, että algoritmin tulos tallennetaan malliksi Splunkiin ja tässä tapauksessa nimetään « tfidf_model ». fit TruncatedSVD -osalla käsitellään TFIDF-mallin tuottamat tulokset, jotka ovat tässä vaiheessa message_* -kentissä, TurncatedSVD algoritmillä k:n arvolla 100 ja tallennetaan malli nimellä « svd_model ».

Viimeinen osa, fit kmeans käsittelee TFIDF ja TruncatedSVD -algoritmeillä käsitellyn datan kentistä « SVD_* » kmeans klusterointialgoritmillä ja tallentaa mallin nimellä « kmeans_model ». Parametrilla k=10 määritellään klustereiden lukumääräksi kymmenen.

Samassa komennossa siis käytettiin peräkkäin kolmea eri koneoppimisalgoritmia. Mallin opetus vie opetusdatan määrästä, algoritmeille annetuista

parametreistä sekä palvelinresursseista riippuen sekunneista useisiin minuutteihin. Esimerkiksi n. 3000 tapahtuman käsittely kesti n. 5 minuuttia.

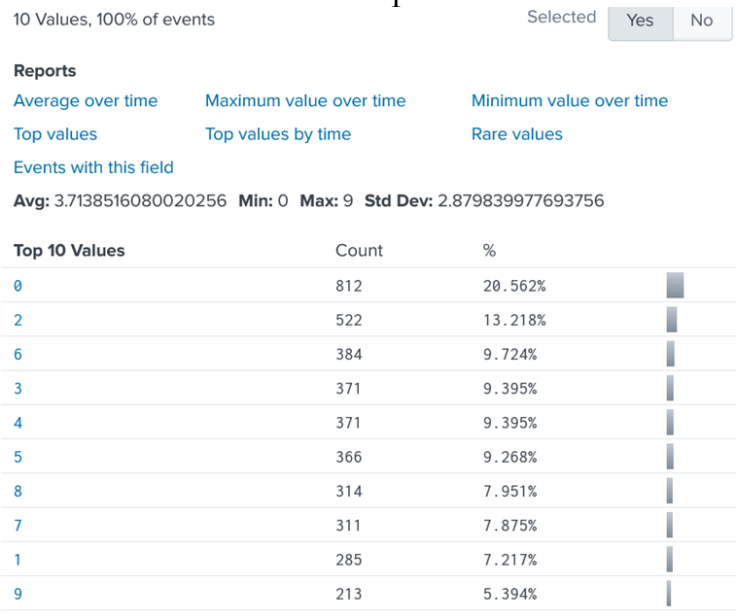
Splunkin fit komento palauttaa suorituksen päätteeksi mallin tulokset opetusdatasta tarkasteltavaksi Splunkin hakunäkymässä jakaen opetusdatana käytettävät tapahtumat klustereihin niiden sisällön perusteella Tämä mahdollistaa mallin kehittämisen ennen käyttöönottoa.

6.2 Koneoppimismallin hyödyntäminen ja uhkien havaitseminen

Koneoppimismallin hyödyntäminen toteutetaan Splunkin apply -hakukomennolla käyttämällä datan käsittelyssä opetusvaiheessa tallennettua kolmea koneoppimismallia. Komento on muotoa:

```
sourcetype=syslog | apply tfidf_model  | apply svd_model  | apply kmeans_model
```

Apply hakukomento toteutuu huomattavasti nopeammin, eikä palvelin resurssitarve ole niin suuri kuin opetusvaiheessa.



KUVIO 14 Opetusdatana käytettyjen tapahtumien jakaantuminen klustereihin klusterinumeron (0-9) perusteella

Tässä tapauksessa käsitellyn datan tarkempi tarkastelu osoitti, että klusteriin 1 jakaantui sellaiset tapahtumat, joissa ei ollut selkeää toisteisuutta, vaan ne olivat sisällöltään muista poikkeavia. muiden klustereiden osalta jakauma perustui Taulukossa 2 esitettyihin sanoihin ja sanayhdistelmiin. Niiden esiintyminen on säännöllistä ja osa järjestelmän normaalia toimintaa, esimerkiksi yhteydenmuodostukseen tai yhteyden päättämiseen liittyviä lokitapahtumia.

TAULUKKO 2 Klustereiden 2-9 ja 0 jakaantuminen sisällön poerusteella

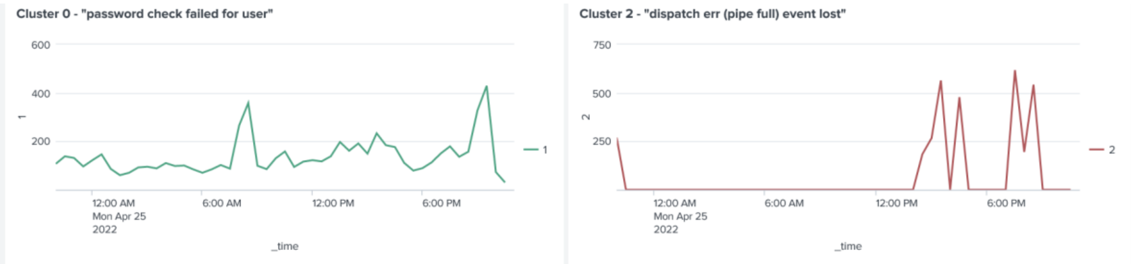
Klusterinumero	MKlusterissa esiintyvien lokien sisältöä
2	dispatch err (pipe full) event lost
3	Received disconnect from
4	authentication failure
5	Disconnect from
6	requirement "uid >= 1000" not met by user "root"
7	check pass; user unknown
8	Failed password
9	invalid user
0	password check failed for user

Tiedon hyödyntämiseksi Splunkiin tehdään näkymät, joissa tapahtumadataa voidaan tarkastella klustereittain. SIEM-järjestelmän dataa tarkasteleva analyytikko voi hyödyntää klusterointia tarkastelemalla eri tapahtumien lukumäärissä tapahtuvia muutoksia järjestelmän toiminnan havainnoimiseksi ja esimerkiksi mahdollisten järjestelmän häiriöiden havaitsemiseksi. Näkymien luominen tulee tehdä Splunkissa siten, että kunkin yksittäisen ikkunan esittämissä tuloksissa käytetään koneoppimismalleja apply-hakukomennon avulla sekä asettamalla hakusuodattimeksi esimerkiksi yhden minuutin, päivittyy näkymän tulokset yhden minuutin välein, mikä on SIEM-käytössä useimmiten riittävä. Täysin reaaliaikainen tulosten esittäminen on mahdollista lisäämällä hakuun komento rtorder, mutta reaaliaikainen näkymien päivitys kuluttaa enemmän palvelinresursseja.

KUVIO 15 eri klustereihin jakaantuvien lokitapahtumien esiintyvyyden tarkastelua visualisoinnin avulla

Mahdollisten kyberuhkien havaitsemisen kannalta hyödyllisintä on tarkastella poikkeavuuksia, jotka tässä tapauksessa kuuluvat klusteriin 1. Niiden sisältö ei ole ennakolta tiedossa, kuten muiden klustereiden, jotka on jo aiemmassa tarkastelussa todettu järjestelmän normaaliksi toiminnaksi. Kuviossa 16 on klustereihin 0 ja 2 jakaantunee lokitapahtumat, joissa on selvä toisteisuus. Kuviossa 17 on puolestaan klusteriin 1 jaetut tapahtumat, joissa on selvästi vaihtelevuutta. Kuviossa 18 on klusterista 1 tapahtumia, jotka ilmaisevat ajastettujen skriptien suorituksen. Kyseiset tapahtumat ovat tässä tapauksessa järjestelmän normaalia toimintaa ja tarkoituksellisia, mutta niiden suoritus ja täten lokitapahtumien esiintyminen on melko harvinaista, joten ne jakaantuivat mallin perusteella juuri klusteriin yksi. Samalla tavalla voitaisiin havaita myös haittatoimijan suorittamia

toimintoja tai haittatoimijan luomia ajastettuja tapahtumia, jotka ovat esimerkkejä kyberhyökkääjän taktiikoista, kuten luvussa 3.2. aikaisemmin kuvattiin.



Cluster 0 - "password check failed for user"

_time	message
2022-04-25 22:40:43	password check failed for user (root)
2022-04-25 22:40:37	password check failed for user (root)
2022-04-25 22:40:21	password check failed for user (root)
2022-04-25 22:40:20	password check failed for user (root)
2022-04-25 22:40:02	password check failed for user (root)
2022-04-25 22:39:54	password check failed for user (root)
2022-04-25 22:39:50	password check failed for user (root)
2022-04-25 22:38:53	password check failed for user (root)
2022-04-25 22:38:47	password check failed for user (root)
2022-04-25 22:38:47	password check failed for user (root)

« Prev 1 2 3 4 5 6 7 8 9 10 Next »

Cluster 2 - "dispatch err (pipe full) event lost"

_time	message
2022-04-25 20:51:20	dispatch err (pipe full) event lost
2022-04-25 20:51:20	dispatch err (pipe full) event lost
2022-04-25 20:51:20	dispatch err (pipe full) event lost
2022-04-25 20:51:20	dispatch err (pipe full) event lost
2022-04-25 20:51:20	dispatch err (pipe full) event lost
2022-04-25 20:51:20	dispatch err (pipe full) event lost
2022-04-25 20:51:20	dispatch err (pipe full) event lost
2022-04-25 20:51:20	dispatch err (pipe full) event lost
2022-04-25 20:51:20	dispatch err (pipe full) event lost
2022-04-25 20:51:20	dispatch err (pipe full) event lost

« Prev 1 2 3 4 5 6 7 8 9 10 Next »

KUVIO 16 esimerkkitapauksessa klustereihin 0 ja 2 jakaantuneita tapahtumia

Cluster 1

_time	message
2022-04-25 22:30:20	Connection closed by 130.234.0.9 port 38226 [preauth]
2022-04-25 22:30:02	(root) CMD (/usr/lib64/sa/sa1 1 1)
2022-04-25 22:30:02	(root) CMD (/root/bin/memory-usage-per-user.sh)
2022-04-25 22:29:56	EXIT: auth status=0 pid=71650 duration=0(sec)
2022-04-25 22:29:56	START: auth pid=71650 from=::ffff:130.233.252.241
2022-04-25 22:29:20	Connection closed by 130.234.0.9 port 38130 [preauth]
2022-04-25 22:28:20	Connection closed by 130.234.0.9 port 38032 [preauth]
2022-04-25 22:27:20	Connection closed by 130.234.0.9 port 37936 [preauth]
2022-04-25 22:27:13	pam_krb5[71408]: authentication fails for 'karim' (karim@AD.JYU.FI): Authentication failure (Preauthentication failed)
2022-04-25 22:27:13	User karim from 180.76.145.213 not allowed because none of user's groups are listed in AllowGroups

« Prev 1 2 3 4 5 6 7 8 9 10 Next »

KUVIO 17 Klusteriin 1 kuuluvia tapahtumia

```
(root) CMD (/root/bin/diskcheck.sh)
(root) CMD (for i in /etc/yum.repos.d/*; do if [ ! "$i" = "/etc/yum.repos.d/redhat.repo" ] && [ ! "$i" = "/etc/
(root) CMD (run-parts /etc/cron.hourly)
(root) CMD (/root/bin/report_packages.sh)
(root) CMD (/usr/lib64/sa/sa2 -A)
```

KUVIO 18 Esimerkkejä klusterissa 1 havaituista lokitapahtumista

6.3 Johtopäätökset järjestelmän käytöstä

Järjestelmän käyttö sisältää tarkasteltavan datajoukon määrittelyn, mallin opettamisen, mallin toiminnan arvioimisen sekä mallin käyttöönoton Splunkin näkymissä. Tiedon hyödyntäminen voidaan tehdä reaaliajassa tarkastelemalla klustereihin jakaantuvia tapahtumia sekä muutoksia niissä. Menetelmän avulla luodaan SIEM-järjestelmää käyttävälle analyytikolle kokonaiskuva muutoin epämääräisestä lokitapahtumajoukosta, josta yksittäisten havaintojen tekeminen esimerkiksi kyberuhkien tunnistamiseksi olisi haastava ja aikaa vievää. Huomioitavaa mallin käytössä on, että klustereiden numerointi muuttuu mallin opetuksen yhteydessä. Näin ollen jokaisen opetuskerran jälkeen on tarkasteltava malli uudelleen mm. sen selvittämiseksi, mihin klusteriin harvinaiset ja poikkeavat tapahtumat jakaantuvat.

Koneoppimismenetelmien käyttö on integroitu osaksi Splunkia, jolloin sen hyödyntäminen Splunkin käyttöliittymän avulla on helppoa. Teknisesti myös muiden koneoppimismenetelmien käyttö Splunkissa voidaan tehdä vastaavalla tavalla fit- ja apply-komennoilla, ja käytössä on laajasti mm. scicit-learn-kirjastoon sisältyviä koneoppimisalgoritmeja.

7 JÄRJESTELMÄN EVALUOINTI JA TULOKSET

DSRM Tutkimusmenetelmän vaiheessa 5 arvioidaan toteutetun järjestelmän toimivuus ja tehokkuus, mikä käsitellään tutkimuksen tässä luvussa. Luvussa kuusi demonstroitiin järjestelmän toiminta tutkimuksen tilaajan ympäristössä. Järjestelmän toiminnasta saatuja havaintoja tarkasteltiin yhdessä tutkimuksen tilaajan kanssa. Tähän on koottu keskeiset havainnot järjestelmän kehitysprosessista, järjestelmän teknisestä toimivuudesta sekä järjestelmän hyödyntämismahdollisuuksista. Tässä kuvataan samalla tutkimuksen tulokset sekä vastataan päätutkimuskysymykseen: « Miten koneoppimista voidaan hyödyntää SIEM-järjestelmissä haitallisen toiminnan havaitsemiseksi?

7.1 Järjestelmän kehitys

Koneoppimista hyödyntävän konstruktion kehittämisen keskeiseksi määrittäväksi tekijäksi muodostui selkeä näkemys järjestelmän kehitysprosessista sekä ennen kaikkea tavoitteenasettelusta: « Mitä koneoppimisen avulla halutaan saavuttaa? ». Tätä tukee myös koneoppimisen ja tiedonlouhinnan kehittämistä kuvaavat prosessit, kuten CRISP-DM, joka lähtee tarpeen ja tavoitteiden määrittelystä, koska tämä ohjaa myös sen, millainen koneoppimisen menetelmä soveltuu parhaiten kyseisen ongelman ratkaisemiseksi (Chapman P., Clinton J., Kerber R., Khabaza T. , Reinartz T., Shearer C., and Wirth R., 2000). Toinen iso tekijä koneoppimismallia hyödyntävän järjestelmän kehittämisessä on käytettävissä oleva data ja sen rakenne ja onko käytössä esimerkiksi merkittyä, rakenteellista, opetusdataa ohjatun koneoppimisen hyödyntämiseksi? Tässä tutkimuksessa ja esimerkkitoteutuksessa tavoitteen tarkempi rajaaminen yhdessä tutkimuksen tilaajan kanssa vaati useamman tarkastelun ennen lopullista määrittelyä, mutta konstruktion kehitysvaiheessa näkemys päämäärästä oli jo selvä, mikä helpotti työn etenemistä ja järjestelmä päästiin testaamaan ja edelleen kehittämään tutkimuksen aikana.

Kaikkiaan koneoppimista hyödyntävän toiminnallisuuden kehitys onkin monivaiheinen ja iteratiivinen prosessi. Joten tässä suhteessa voidaan todeta, että koneoppimisen hyödyntäminen SIEM-järjestelmässä vaatii jatkuvaa mallin toiminnan arviointia ja kehittämistä ennen kuin saavutetaan toivuttuja tuloksia. Tämänkin jälkeen kehitettävää ilmenee, varsinkin, mikäli toimintaympäristössä ja datassa tapahtuu merkittäviä muutoksia.

7.2 Järjestelmän tekninen toteutus

Menetelmän toteuttaminen on teknisesti suhteellisen yksinkertaista, eikä vaadi erityistoimia Splunkin suhteen tai esimerkiksi korkeampia käyttäjäoikeuksia.

Erityisesti koneoppimismalin opettaminen vaatii kuitenkin paljon laiteresursseja palvelimelta, erityisesti keskusmuistia. Splunkin koneoppimismalleille kohdennettavan muistimäärän muuttaminen voikin vaatia järjestelmänvalvojan toimia. Sekä datan valmistelussa, että mallin kehitysvaiheessa voidaan vaikuttaa paljon mallin tekniseen toimintaan ja resurssitarpeisiin. Käsiteltävien lokikenttien rakennetta kehittämällä, sekä suodattamalla opetusaineistona käytettävää historiadataa harkitusti, voidaan pienentää järjestelmän resurssitarvetta ja lyhentää mallin opetusaikaa. Myös vaihtoehtoisilla algoritmien valinnalla esimerkiksi klusterointiin voidaan saada mahdollisesti resurssitarvetta pienennettyä.

7.3 Järjestelmän hyödyntämismahdollisuudet

Tutkimuksessa toteutettua järjestelmää testattiin toimintaympäristössä, jonka lokitapahtumien sisällöstä ei tutkimuksen alkaessa ei ollut juurikaan tietoa. Tässä suhteessa tutkimuksessa toteutettu luonnollisen kielen prosessointia hyödyntävän menetelmän avulla saatiin käsiteltyä suurta tietomäärää ja muodostettua hyvä käsitys lokidatan sisällöstä.

Menetelmää ei kuitenkaan testattu ympäristössä, jossa olisi tiedetty varmasti olevan käynnissä haitallista toimintaa. Näin ollen koneoppimismallin toimintaa suoranaisesti kyberuhkien havaitsemiseksi ei validoitu, eikä mallin ennustetarkkuutta kyberuhkien tunnistamiseksi voitu määrittää. Tutkimuksen perusteella ei voida yksiselitteisesti sanoa, pystyykö menetelmällä tunnistamaan todellista haitallista toimintaa. Empiirinen tarkastelu kuitenkin osoitti, että ilman ennakkotietoa järjestelmään tehdystä toimista lokidatasta saatiin rajattua joukko sellaisia lokitapahtumia, jotka viittasivat järjestelmässä ajastetusti tai suunnitellusti järjestelmänvalvojien toimesta suoritettuihin prosesseihin ja skripteihin, mitkä kyberhyökkääjän yleisiä tekniikoita tarkastellen voisi huonommassa tapauksessa olla myös hyökkääjän järjestelmään luomaa haitallista toimintaa esimerkiksi pysyvyyden varmistamiseksi järjestelmässä.

Järjestelmän ylläpitäjien näkökulmasta saaduissa tuloksissa havaittiin paljon lokitapahtumia, jotka voitaisiin jo lähtökohtaisesti jättää huomioitta, eli koneoppimismenetelmällä käsiteltävää datajoukkoa voisi rajata tarkemmin, jotta malli ei käsittelisi lainkaan dataa, jonka merkitys on jo tiedossa, vaan mallin käyttö kohdennettaisiin selkeästi siihen osaan datasta, joka ei ole valmiiksi riittävän hyvin luokiteltua, ja jonka sisältö on tuntematonta, mikä säästäisi myös mallin opettamisessa tarvittavia resursseja.

Yksi mahdollisuus olisi jatkaa järjestelmän kehitysprosessia ja luoda ohjaamattoman oppimisen avulla opetusdataa ohjatun oppimisen menetelmien, esimerkiksi luokittelun (engl. classification) hyödyntämiseksi klusteroinnin sijaan, jolloin ohjaamattoman oppimisen avulla muodostettuun aineistoon voitaisiin nimetä klusterit niiden sisällön mukaan (Wiley, J. and Sons, 2015).

SIEM-järjestelmässä tiedon esittämistapa on oleellista tiedon hyödyntämisen ja mahdollisten vastatoimien vuoksi:

- Visualisointien rakentaminen on usein yksilöllistä ja tehtäväsidonnaista – valvonta vs. aktiivinen uhkien etsintä
- Menetelmän avulla esitetyn tiedon tarkastelu vaatii ihmisarviointia – tulokset eivät ole yksiselitteisiä
- Havaintojen hyödyntäminen vaatii hyvää tuntemusta ja tilannekuvaa valvottavasta järjestelmästä. ”Onko jokin tapahtuma esim. tarkoitukselliseen ylläpitotoimeen liittyvä”.

8 JOHTOPÄÄTÖKSET

SIEM-järjestelmä kerää, esikäsittelee, tallentaa lokitietoja valvottavasta järjestelmästä sekä esittää näkyminä järjestelmän toiminnassa tapahtuvat muutokset, havainnot mahdollisesta poikkeavasta toiminnasta sekä muut järjestelmän käyttäjän, esimerkiksi SOC-analyytikon haluamat asiat. SIEM-järjestelmä täyttää paitsi organisaation lakisääteiset velvoitteet tapahtumatietojen tallentamisesta ja säilyttämisestä, mutta mahdollistaa myös pahimmillaan organisaatiolle merkittävääkin haittaa aiheuttavien kyberhyökkäysten tunnistamisen, parhaimmillaan hyökkäyksen valmisteluvaiheessa, ennen kuin hyökkääjä saavuttaa omat tavoitteensa. Haasteena kyberuhkien tunnistamisessa SIEM-järjestelmään kerätystä lokidatasta on usein, uhkien tunnistamisen edellyttävien havainnointisääntöjen ylläpitoa ja kehittäminen, käsiteltävän tiedon suuri määrä. Huonoista havainnointisäännöistä johtuvien väärien positiivisten hälytysten suuri määrä.

Ratkaisuna edellä mainittuihin ongelmiin on nähty muun muassa koneoppimisen hyödyntämisen. Tutkimuksessa selvitettiinkin, ”Miten koneoppimista voidaan hyödyntää SIEM-järjestelmissä haitallisen toiminnan havaitsemiseksi?” Tutkimuksessa toteutetun kirjallisuuskatsauksen perusteella koneoppimista hyödyntäviä menetelmiä voidaan integroida SIEM-järjestelmän tiedonkäsittelyketjun eri vaiheisiin, millä pyritään parantamaan järjestelmän kykyä tuottaa käyttäjälleen tietoa mahdollisesta haitallisesta toiminnasta valvottavasta järjestelmästä. Kyberuhkien havaitsemisen kannalta potentiaalisimmat hyödyntämismahdollisuudet ovat tapahtumahavaintojen käsittelyssä tiedonlouhinnan keinoin, koska kyberhyökkääjän taktiikoihin lukeutuu useita toimia, joista jää jälkiä hyökkäyksen kohteena olevan järjestelmän lokitietoihin, ja toiminta voidaan havaita tunnistamalla nämä tapahtumat SIEM-järjestelmään kerätystä lokitiedoista.

Tutkimuksessa esitetty ohjaamatonta koneoppimista hyödyntävä, luonnollisen kielen prosessointia hyödyntävä menetelmä on helppo toteuttaa, koska menetelmän käyttö ei vaadi mittavia valmistelutoimia, eikä merkittyä opetusdataa, kunhan mallin opetuksen käytettävää historiadataa on käytössä paljon ja järjestelmän palvelimessa on riittävästi resursseja. Yksi haaste koneoppimismenetelmän hyödyntämisessä onkin sopivien mallin kehittäminen valvottavaan ympäristöön. Mallin kehittäminen on iteratiivinen prosessi, joka vaatii useita toistoja parametrien muuttamista halutun tuloksen saamiseksi. Koneoppimismallin käyttö ja kehittäminen edellyttää myös jatkuvaa toimivuuden arviointia sekä vuoropuhelua SIEM-järjestelmän dataa hyödyntävien henkilöiden kanssa, jotta menetelmä on edelleen käyttökelpoinen myös ympäristön ja datan mahdollisesti muuttuessa.

Vaikka koneoppimisen avulla ei ole realistista odottaa saatavan yksiselitteisiä havaintoja mahdollisesta haitallisesta toiminnasta, eikä se ratkaise täysin kyberuhkien havaitsemiseen tai SIEM-järjestelmien käyttöön liittyviä haasteita. Tutkimuksen havaintojen perusteella tutkimuksessa käytetty ratkaisu tuo se kuitenkin yhden työkalun SOC-analyytikoiden työkalupakkiin suuren tietomäärän käsittelyn tukemiseksi. Tälläkin menetelmällä saatavat havainnot potentiaalisesti kiinnostavista lokitapahtumista edellyttävät SOC-analyytikon, eli ihmisen tapauskohtaista arviointia tapahtuman merkityksestä ja vakavuudesta. Tutkimus toteutettiin tukemaan Jyväskylän yliopiston Digipalveluiden tarpeesta saada lisätietoa koneoppimisesta SIEM-järjestelmistä, ja nimenomaan Splunk -sovelluksesta. Tutkimuksessa toteutettu esimerkkiratkaisu lisäsi entuudestaan tutkimuksen tilaajan kiinnostusta ja luottamusta koneoppimistoteutuksia kohtaan. Koneoppimismenetelmien toteuttaminen Splunkissa on teknisesti helppoa sikäli, että järjestelmään kuuluvassa liitännäisessä on käytössä scikit learn -kirjaston koneoppimisalgoritmit sekä melko hyvä dokumentaatio, mikä mahdollistaa erilaiset koneoppimistoteutukset, kunhan toteuttajalla on selkeä näkemys ja tavoitteet, mitä menetelmällä halutaan saada aikaiseksi.

Lokitietojen käsittely toteutetun ratkaisun avulla tuotti usein tilanteita, jossa kiinnostavalta vaikuttava lokitapahtuma oli täysin tarkoituksellinen, järjestelmän ylläpitäjän toteuttama toiminta. Potentiaalisena jatkotutkimusaiheena voisi olla analyytikon palautteen perusteella oppivan koneoppimisjärjestelmän toiminta, jolloin väärän positiivisen havaittuaan analyytikko voisi merkitä tapahtuman harmittomaksi, jolloin koneoppimismalli kehittyy analyytikon palautteen perusteella. Tällaisesta toteutuksesta on saatu havainnointitarkkuuden näkökulmasta hyviä tuloksia (Veeramachaneni K., Arnaldo I., Cuesta-Infante A., Korrapati V. Bassias, Li K., 2016), mutta menetelmän käytännön sovittamisesta osaksi SIEM-järjestelmää voisi olla hyödyllinen lisätutkimuksen aihe.

LÄHTEET

Baier, L., Jöhren, F. & Seenbacher, S. (2019). Challenges in the Deployment and Operation of Machine Learning in Practice. *27th European Conference on Information Systems.* Stockholm.

Baier, L., Jöhren, F., Seebacher, S. (2019). CHALLENGES IN THE DEPLOYMENT AND OPERATION OF MACHINE LEARNING IN PRACTICE. *Twenty-Seventh European Conference on Information Systems (ECIS2019).* Stockholm-Uppsala, Sweden.

Boucher, P. (2020). *Artificial intelligence: How does it work, why does it matter, and what can we do about it?* Scientific Foresight Unit (STOA). Noudettu osoitteesta Scientific Foresight Unit (STOA): https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2020)641547

Bowers, B. (11. 6 2020). *How to improve threat detection and response with MITRE ATT&CK*. Noudettu osoitteesta https://blog.shi.com/solutions/how-to-improve-threat-detection-and-response-with-mitre-attck/

Buczak, A. & Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys and Tutorials, VOL. 18, NO. 2, second quarter 2016.*

Chapman P., Clinton J., Kerber R., Khabaza T. , Reinartz T., Shearer C., and Wirth R. (2000). *The CRISP-DM User Guide*. Noudettu osoitteesta https://www.the-modeling-agency.com/crisp-dm.pdf

Corporation, T. M. (2022). | *User Behavior Analysis*. Noudettu osoitteesta https://d3fend.mitre.org/technique/d3f:UserBehaviorAnalysis/

Crawford M., Khoshgoftaar Taghi M., Prusa J. D. ,Richter A. N., Al-Najada H. (05. October 2015). Survey of review spam detection using machine learning techniques. *Journal of Big Data 2, 23 (2015).*

Endsley, M. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, vol 37, no. 1*, ss. 32-64.

Feng, C., Wu, S. & Liu, N. (2017). A user-centric machine learning framework for cyber security operations center. *IEEE International Conference on Intelligence and Security Informatics (ISI).*

Finlex. (2019). *Laki julkisen hallinnon tiedonhallinnasta 906/2019 17 §*. Noudettu osoitteesta https://www.finlex.fi/fi/laki/alkup/2019/20190906

Franke, U., Brynielsson, J. (2014). Cyber situational awareness - A systematic review of the literature. *Computers & Security 46*, ss. 18-31.

Gregor, S. & Hevner, A. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly Vol. 37 No. 2*, 337-335.

Henderson, J., Hubbard, J. (2017). Modern Log Parsing and Enrichment with SIEM. *SANS Webcast 8.11.2017*. Noudettu osoitteesta https://github.com/SMAPPER/presentations

Iklody, A., Wagener, G., Dulaunoy, A., Mokaddem, S., Wagner, C. (2018). *Decaying Indicators of Compromise.* Luxembourg: CIRCL- Computer Incident Response Center .

Kansallinen turvallisuusviranomainen. (2020). *Katakri 2020 Tietoturvallisuuden auditointityökalu viranomaisille.* Traficomin julkaisusarja.

Keller, J. (18. November 2015). *Army Cyber Situational Awareness Innovation Challenge focuses on cyber threats at brigade leve*. Noudettu osoitteesta Military & Aerospace Electronics: https://www.militaryaerospace.com/computers/article/16713939/army-cyber-situational-awareness-innovation-challenge-focuses-on-cyber-threats-at-brigade-level

Kokkonen, T. (2016). Architecture for the Cyber Security Situational Awareness System. *O. Galinina, S. Balandin, Y. Koucheryavy (eds.) Internet of Things, Smart Spaces, and Next Generation Networks and Systems. Lecture Notes in Computer Science, vol. 9870*, ss. 294-302.

Kokulu, F. B., Shoshitaishvili, Y., Ziming Zhao, A. S., Ahn, G. J., Bao, T., Doupé, A. (2019). Matched and Mismatched SOCs: A Qualitative Study on Security Operations Center Issues. *CCS '19: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security November, 2019 .*

Logpoint. (2022. 09 2022). *Logpoint UEBA*. Noudettu osoitteesta https://docs.logpoint.com/docs/ueba-manual/en/latest/index.html

Neapolitan, R. & Jiang, X. (2018). *Artificial Intelligence : With an Introduction to Machine Learning, Second Edition.* CRC Press LLC.

Nunes, E., Diab, A., Gunn, A., Marin, E., Mishra, V., Paliath, V., Robertson, J., Shakarian, J., Thart, A., Shakarian, P. (2016). *Darknet and Deepnet Mining for Proactive Cybersecurity Threat Intelligence.* Arizona State University.

Peffers, K., Tuunanen, T., Gengler, C., Rossi, M., Hui, W., Virtanen, V. & Bragge, J. (2006). Design Science Research Process: A Model for Producing and Presenting Information Systems Research. *1st International Conference, DESRIST 2006 Proceedings.*, 83-106.

Rasche, G. (2013). Guidelines for Planning an Integrated Security Operations Center. EPRI Project.

Scikit-learn project. (2021). *scikit-learn Machine Learning in Python*. Haettu 15.2.2021 osoitteesta https://scikit-learn.org/stable/

Sindhu, V., Nivedha, S., Prakash, M. (2020). An empirical science research on bioinformatics in machine learning. *Journal of Mechanics of Continua and Mathematical Sciences.* ISSN (Online) : 2454 -7190. Noudettu osoitteesta https://www.journalimcms.org/wp-content/uploads/6-AN-EMPIRICAL-SCIENCE-RESEARCH.pdf

Souppaya, M., Scarfone, K. (2006). *Guide to Computer Security Log Management.* NIST Special Publication 800-92.

Splunk. (2021). *Splunk® Machine Learning Toolkit User Guide. Preprocessing your data using MLTK Assistants*. Noudettu osoitteesta https://docs.splunk.com/Documentation/MLApp/5.2.1/User/Preprocessing

Splunk Inc. (2021). *Splunk® Enterprise*. Noudettu osoitteesta https://www.splunk.com/en_us/software/splunk-enterprise.html

Splunk Inc. (2021). *Splunk® Universal Forwarder. Forwarder Manual*. Noudettu osoitteesta https://docs.splunk.com/Documentation/Forwarder/8.1.2/Forwarder/Abouttheuniversalforwarder

Splunk Inc. (ei pvm). *Product Feature Details: Splunk Phantom*. Noudettu osoitteesta https://www.splunk.com/en_us/software/splunk-security-orchestration-and-automation/features.html

Splunk Inc: Algorithms in the Machine Learning Toolkit. (2022). Noudettu osoitteesta https://docs.splunk.com/Documentation/MLApp/latest/User/Algorithms

Stoner, J. (8. 2 2019). *ATT&CK-ing the Adversary: Episode 3 – Operationalizing ATT&CK with Splunk*. Noudettu osoitteesta https://www.splunk.com/en_us/blog/security/att-ck-ing-the-adversary-episode-3-operationalizing-att-ck-with-splunk.html

Stroeh, K, Madeira, E. & Goldenstein, S. (2013). An approach to the correlation of security events based on machine learning techniques. Journal of Internet Services and Applications. Noudettu osoitteesta http://www.jisajournal.com/content/4/1/7

Strom, B.E., Applebaum, A., Miller, D.P., Nickels, K.C., Pennington, A.G., Thomas, C.B. (2018). *MITRE ATT&CK™: Design and Philosophy*. Noudettu osoitteesta https://www.mitre.org/sites/default/files/publications/pr-

Susanto, H., Almunawar, M., Tuan, Y. (2011). Information Security Management System Standards: A Comparative Study of the Big Five. *nternational Journal of Electrical & Computer Sciences IJECS-IJENS Vol: 11 No: 05*, 23-29.

The Industrial Control System Information Sharing and Analysis Center (ICS-ISAC). (ei pvm). *Situational Awareness Reference Architecture (SARA)*. Noudettu osoitteesta http://ics-isac.org/blog/sara/

The MITRE Corporation. (2021). *MITRE ATT&CK® Matrix for Enterprise*. Noudettu osoitteesta https://attack.mitre.org/

The MITRE Corporation. (2022). *D3FEND A knowledge graph of cybersecurity countermeasures*. Noudettu osoitteesta https://d3fend.mitre.org/

The MITRE Corporation. (ei pvm). *Cybersecurity, Situation Awareness*. Noudettu osoitteesta https://www.mitre.org/capabilities/cybersecurity/situation-awareness/

The MITRE Corporation; Mitre Att&ck | Dynamic Resolution: Domain Generation Algorithms. (2022). Noudettu osoitteesta https://attack.mitre.org/techniques/T1637/001/

The MITRE Corporation; Mitre Att&ck | Scheduled Task/Job. (2022). Noudettu osoitteesta https://attack.mitre.org/techniques/T1053/

Tiedonhallintalautakunta. (2020). *Suosituskokoelma tiettyjen tietoturvallisuussäädösten soveltamisesta. Valtiovarainministeriön julkaisuja 2020:21.* Helsinki: Valtionvarainministeriö.

Todd, B. (2017). *Creating a LoggingInfrastructure*. Noudettu osoitteesta SANS InstituteInformation Security Reading Room.

Turvallisuuskomitean sihteeristö. (2013). Suomen kyberturvallisuusstrategia. *Valtioneuvoston periaatepäätös 24.1.2013*.

Uma, M., Padmavathi, G. (2013). A Survey on Various Cyber Attacks and Their Classification. *International Journal of Network Security, Vol.15, No.5*, ss. 390-396.

Valtionhallinnon tietoturvallisuuden johtoryhmä. (2009). *VAHTI 3/2009 Lokiohje*. Valtionvarainministeriö .

Veeramachaneni K., Arnaldo I., Cuesta-Infante A., Korrapati V. Bassias, Li K. (2016). AI2: : Training a big data machine to defend. *IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)* (ss. 49-54). New York: IEEE.

Vielberth, M., Böhm, F., Fichtinger, I., Pernul, G. (31. 12 2020). Security Operations Center: A Systematic Study and Open Challenges. *IEEE Open Access Journal 1109/ACCESS.2020.3045514*.

Viestintävirasto. (2016). *Lokien keräys ja käyttö. Ohje lokitietojen tallentamiseen ja hyödyntämiseen.* Viestintäviraston Ohje 4/2016.

Vähäkainu, P. & Lehto, M. (2019). Artificial intelligence in the cyber security environment. *14th International Conference on Cyber Warfare and Security.*

Watts, S. (15. May 2020). *What Is Security Orchestration, Automation, and Response (SOAR)?* Noudettu osoitteesta Security & Compliance Blog: https://www.bmc.com/blogs/soar-security-orchestration-automation-response/

Wiley, J. and Sons. (2015). *Data science & big data analytics : discovering, analyzing, visualizing and presenting data.* Indianapolis, Indiana: John Wiley and Sons.

Worsham, Nathan. (2018). *NLP Text Analytics Splunk App*. Noudettu osoitteesta https://github.com/geekusa/nlp-text-analytics

Zimmerman, C. (2014). *Ten Strategies of a World-Class Cybersecurity Operations Center.* MITRE.