

Stevon Keel

LOKIDATAN HYÖDYNTÄMISEN MAHDOLLISUUDET JA HAASTEET PÄÄTÖKSENTEOSSA

Informaatioteknologian ja viestinnän tiedekunta
Kandidaattitutkielma
Joulukuu 2025

TIIVISTELMÄ

Stevon Keel: Lokidatan hyödyntämisen mahdollisuudet ja haasteet päätöksenteossa
Kandidaattitutkielma
Tampereen yliopisto
Tieto- ja sähkötekniikan tutkinto-ohjelma
Joulukuu 2025

Lokidata on yksi tärkeimmistä tiedonlähteistä yrityksen IT-järjestelmissä, josta voi saada tietoa esimerkiksi siitä, kuinka hidas järjestelmä on, milloin järjestelmä on kaatunut ja minkälaista kommunikaatiota järjestelmässä on tapahtunut milloinkin. Lokidatan analysointiin löytyy monia eri menetelmiä, kuten tilastolliset menetelmät, neuroverkot, ryväs-täminen, ohjattu ja ohjaamaton oppiminen. Lokidatan analysoinnissa hyödynnetään usein poikkeamien tunnistamista, mikä edistää lokidatan analysointia siten, että loki-datan näytearvoista löydetään suoraan mahdollisia poikkeuksia.

Lokidatan määrä ja merkitys on viime aikoina lisääntynyt suuresti. Yritykset nykypäivinä kärsivät informaatiotulvasta, eli lokidataa syntyy huomattavasti enemmän, kuin voidaan käsitellä. Tutkimuksissa on todettu, että kielimallit ja koneoppimisen menetelmät ovat osoittautuneet hyödylliseksi, vaikka lokidatassa koneoppimisen menetelmät ovat edel-leen varhaisessa vaiheessa. Lokidatasta on tutkittu useita hyviä sekä haittapuolia. Yksi yleisimmistä haittapuolista lokidatan käsittelyssä on huononlaatuiset lokit. Lokeja on yk-sinkertaisesti liikaa tai lokeista on vaikea hahmottaa, mitkä lokimerkinnät ovat oleellisia. Tutkimuksissa on todettu, että visualisoinnin puuttuminen on hankaloittanut lokidatan analysointia. Poikkeamien tunnistamista ja visualisointia voi hyödyntää yhdessä samaan aikaan, jolloin on mahdollista nähdä esimerkiksi kaikki järjestelmän aikaviitteet kuukau-den ajalta. Lisäksi se auttaa havaitsemaan poikkeavat arvot nopeasti ilman, että täytyy manuaalisesti tutkia lokimerkintöjä yksi kerrallaan. Visualisoinnin on todettu olevan kriit-tinen osa datan analysointia. Visualisoinnin avulla on helpompi saada datan kokonais-kuva. On todettu, että visualisoinnin hyödyntäminen parantaa päätöksentekoa. Loki-datan ja visualisoinnin yhdistämisellä on mahdollista ennaltaehkäistä mahdollisten virhe-tilanteiden syntymistä tai on mahdollista puuttua virhetilanteisiin heti tapahtuma hetkellä.

Avainsanat: lokidata, Splunk, poikkeamien tunnistaminen, visualisointi

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck –ohjelmalla.

TEKOÄLYN KÄYTTÖ OPINNÄYTTEESSÄ

Opinnäytteessäni on käytetty tekoälysovelluksia:

- Ei
- Kyllä

Ilmoitukseni mukaan olen käyttänyt opinnäytteessäni tutkielmaprosessin aikana seuraavia tekoälysovelluksia:

Tekoälysovellusten nimet ja versiot:

-ChatGPT-5

-Scopus AI

Käyttötarkoitus: Tekoälyä ChatGPT-5 on käytetty ideoimiseen ja lauseiden sujuvampaan ilmaisuun. Scopus AI:ta on käytetty ainoastaan lähteiden etsimiseen.

Osiot, joissa tekoälyä on käytetty: luvuissa 1, 3 ja 4

Olen tietoinen siitä, että olen täysin vastuussa koko opinnäytteeni sisällöstä, mukaan lukien osat, joissa on hyödynnetty tekoälyä, ja hyväksyn vastuun mahdollisista eettisten ohjeiden rikkomuksista.

SISÄLLYSLUETTELO

1	Johdanto	1
2	Tutkimusmenetelmä ja aineisto	3
2.1	Systemaattinen kirjallisuuskatsaus	3
2.2	Tiedonhaun prosessi	4
2.3	Aineiston seulonta ja esittely	6
3	Teoreettinen tausta	10
3.1	Lokidatan määritelmä ja tärkeys	10
3.2	Lokidata osana big dataa	11
3.3	Lokien jäsentely ja analysointi	11
3.4	Visualisointi ja poikkeamien tunnistaminen	12
4	Tulokset ja keskustelua	16
4.1	Tutkimuskysymys 1: Mitä hyviä ja huonoja puolia lokidatan hyödyntämisellä voi olla päätöksenteossa?	17
4.2	Tutkimuskysymys 2: Miten lokien visualisointi ja poikkeamien tunnistaminen tukevat päätöksenteon prosessia?	19
4.3	Tutkielman luotettavuus	20
5	Yhteenveto	21
	Lähdeluettelo	23

1 Johdanto

Nykyisissä digitaalisissa liiketoimintaympäristöissä yritysten menestys riippuu yhä enemmän kyvystä hallita ja hyödyntää lisääntyvää datamäärää. Yritykset ovat usein paremmissa asemassa kuin kilpailijansa, kun dataa hyödynnetään tehokkaasti (Alshawwreh et al., 2024). Shen ja muut (2022) selvittävät tietotekniikan roolia lentoyhtiöiden kilpailuedun tukemisessa, operatiivisen tehokkuuden lisäämisessä ja tiedon hyödyntämisessä päätöksenteossa. Zhao ja muut (2025) totesivat tutkimuksessaan, että pienimmätkin järjestelmien suoriutumisen poikkeamat voivat johtaa huomattaviin taloudellisiin tappioihin, häiriöihin, sekä ne voivat heikentää käyttäjäkokemusta. Tutkimuksessa havaittiin, että Amazonissa 0,1 sekunnin viive tietokannan latauksessa aiheuttaa noin 1 % lisätappion tuloissa. Tutkimuksessa myös mainittiin, että Alibaba Cloud menettää miljardeja dollareita vuosittain satunnaisten hidastuneiden kyselyiden vuoksi. Niinpä lokidata tarjoaa yksityiskohtaista tietoa järjestelmien ja asiakkaiden toiminnasta ja käyttäytymisestä, koska lokidata generoituu IT-järjestelmistä, palvelimista ja digitaalisista palveluista. Kun lokidataa käsitellään oikein, niin poikkeamat voidaan havaita heti alussa, ja mahdolliset vikatilanteet voidaan korjata tai ennaltaehkäistä.

Lokidataa syntyy valtava määrä, minkä vuoksi tarvitaan myös tehokkaita menetelmiä, jotka auttavat hahmottamaan tietoa nopeasti ja intuitiivisesti. Suurten tapahtumalokien visualisointi visuaalisen analytiikan avulla tukee tarkempaa päätöksentekoa sekä lisää datan ymmärrettävyyttä ja läpinäkyvyyttä (Sitova & Pecerska, 2020). Visualisoinnin avulla voidaan tunnistaa datasta erilaisia trendejä, poikkeamia ja riippuvuuksia, joita olisi vaikea havaita pelkästään yksittäisiä lokimerkintöjä seuraamalla. Lokidatan analysoinnin tukena on visualisoinnin lisäksi poikkeamien tunnistaminen. Huang ja muut (2025) ovat todenneet tutkimuksessaan, että poikkeamien tunnistamisen avulla voidaan automaattisesti havaita tapahtumia, jotka poikkeavat normaalista tai voidaan huomata käyttäytymismalleja suurissa ja monimutkaisessa tietoaaineistossa. Tutkimuksessa myös todettiin, että menetelmiä on kehitetty sekä tilastollisiin analyysiin, että kone- ja syväoppimiseen ja ne tukevat esimerkiksi järjestelmien valvontaa ja virheiden havaitsemista. Poikkeavia näytteitä kutsutaan yleisimmin anomaleiksi (engl. anomaly).

Lokidatan teknisiä puolia on tutkittu paljon, mutta verrattain vähän erilaisista päätöksenteon näkökulmista, vaikka lokidatan määrä ja merkitys esimerkiksi yrityksissä on lisääntynyt suuresti viime vuosina. Etenkin Splunk-työkalun hyödyntämistä tieteellisessä kontekstissa on käsiteltyä hyvin vähän, vaikka se on yritysmaailmassa yksi yleisimmistä lokien analytiikkatyökaluista. Tässä tutkielmassa siis vastataan siihen, mitä mahdollisuuksia ja haasteita lokidatan hyödyntämisellä voi olla päätöksenteossa. Tutkimuskysymys on jaoteltuna kahteen alakysymykseen: *“Mitä hyviä ja huonoja puolia lokidatan hyödyntämisellä voi olla päätöksenteossa?”* ja *“Miten lokien visualisointi ja poikkeamien tunnistaminen tukevat päätöksenteon prosessia?”*. Tutkielma suoritetaan kirjallisuuskatsauksena. Lisäksi hyödynnetään Splunk-työkalua. Aiheen valinta perustuu aikaisempaan työkokemukseen, josta kertyi käytännönsaamista yrityksen lokidatan käsittelystä, hyödyntämisestä ja sen vaikeuksista.

Tässä tutkielmassa asioita ei käydä läpi niin yksityiskohtaisesti kuin voitaisiin käydä. Esimerkiksi visualisoinnissa tai sen hyödyntämisessä ei oteta kantaa erilaisiin käytettävyyteen liittyviin asioihin, kuten värisokeuteen. Tässä tutkielmassa selitetään mitä visualisointi on, ja miten sitä hyödynnetään lokidatassa. Tutkielmassa puhutaan koneoppimisesta, koska on syytä huomioida erilaiset koneoppimisen menetelmät ja kuinka niitä hyödynnetään, mutta tutkielmassa ei selitetä koneoppimisen perusteita, kuten ryvästämistä tai ohjattua oppimista. Splunk-työkalua käytetään esimerkinomaisesti, mutta tutkielman tarkoituksena ei ole selittää työkalun perusteita ja kuinka sillä luodaan kuvaajia. Toisena rajauksena on, että vaikka aihe perustuu aikaisempaan työkokemukseen, niin kyseisen yrityksen lokidataratkaisuja ja päätöksentekomalleja ei käydä läpi tai mainita tässä tutkielmassa. Kaikki tutkielmaan liittyvät aiheet perustuvat tieteellisiin tutkimuksiin, joita havainnollistetaan Splunk-työkalun avulla.

Tutkielma on jaettu viiteen eri lukuun. Luvussa 2 käsitellään tutkimusmenetelmiä ja tutkimusaineistoa. Luvussa 3 käsitellään teoreettista taustaa aiheesta ja avataan lokidatan käsitettä, ja lokidatan suhdetta big dataan. Lisäksi selitetään visualisoinnin sekä poikkeamien tunnistamisen hyödyntämistä lokidatan analysoinnissa. Luvussa 4 käsitellään tutkielman tulokset ja vastataan varsinaiseen tutkimuskysymykseen. Luvussa 5 esitellään yhteenveto tutkielmasta, ja pohditaan tulosten merkitystä sekä jatkotutkimusmahdollisuuksia.

2 Tutkimusmenetelmä ja aineisto

Seuraavissa alaluvuissa on kerrottu mahdollisimman tarkasti, miten aineistoja on kerätty, millaisia kriteereitä on käytetty. Lisäksi niissä on kerrottu, mistä tutkielma koostuu. Systemaattisella kirjallisuuskatsauksella pyritään siihen, että muiden on mahdollista toistaa tutkielma, sekä siihen, että tutkielman toteutusta ja tuloksien pätevyyttä voidaan arvioida kriittisesti.

2.1 Systemaattinen kirjallisuuskatsaus

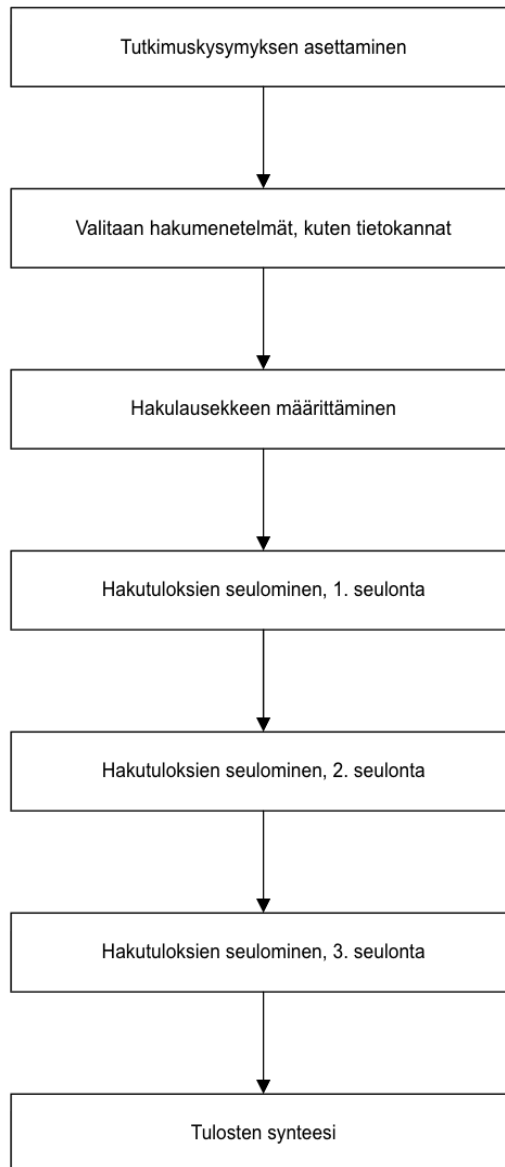
Tutkielma suoritetaan systemaattisena kirjallisuuskatsauksena. Systemaattinen kirjallisuuskatsaus on menetelmä, jossa tavoitteena on kerätä ja analysoida aiempaa tutkimusta järjestelmällisesti sekä arvioida sen laatua kriittisesti. Kirjallisuuskatsauksen avulla kootaan ja analysoidaan ajantasainen tieto tietystä tutkimusaiheesta, jolloin lukijalle muodostuu käsitys tämänhetkisestä tutkimustilanteesta. Tutkielmassa on tavoitteena tarkastella ja arvioida olemassa olevaa tietämystä kriittisesti, tunnistaa sen keskeiset vahvuudet ja puutteet, sekä osoittaa mahdollisia uusia tutkimussuuntia tulevia tutkimuksia varten. (Carrera-Rivera et al, 2022)

Systemaattisen kirjallisuuskatsauksen tarkoituksena on olla metodologisesti tarkka, kun arvioidaan tutkimustuloksia. Systemaattisen kirjallisuuskatsauksen tarkoitus ei ole pelkästään kaikkien sopivien näytteiden keräämistä tietystä tutkimuskysymyksestä, vaan tukea näyttöön perustuvien ohjeistusten kehittämistä. (Kitchenham et al., 2009) Systemaattisessa kirjallisuuskatsauksessa on standardeja, joita Kitchenham (2007) on kehittänyt ja myöhemmin päivitetty vuonna 2009 useamman tutkijan avulla.

Kitchenham kertoo tutkimuksessaan (Kitchenham, 2007), että suurin osa tutkimuksista noudattaa jonkinlaista kirjallisuuskatsausta, mutta jos kirjallisuuskatsaus ei ole tasapuolinen tai perusteellinen, on tieteellinen arvo vähäinen. Tutkimuksessa todetaan, että perusteellisuus ja tasapuolisuus on keskeinen peruste systemaattiselle kirjallisuuskatsaukselle. Katsauksessa on tarkoituksena käyttää ennalta määritettyjä hakulauseita, jotta voidaan arvioida haun kattavuutta. Lisäksi tutkimuksessa mainitaan, että tutkijan tulee pyrkiä tunnistamaan ja raportoimaan tutkimuksia, jotka tukevat tutkimushypoteesia, mutta myös niitä tutkimuksia, jotka eivät tue sitä

2.2 Tiedonhaun prosessi

Tutkielmassa on noudatettu mallia, joka on näkyvässä kuvassa 1. Systemaattisessa kirjallisuuskatsauksessa on pyritty asettamaan sopiva tutkimushypoteesi tai tutkimuskysymys, jonka lisäksi valitaan sopivat tietokannat ja hakulausekkeet. Sen jälkeen on oleellista seuloa tutkimuksia kriteereiden mukaisesti, ja suorittaa tulosten synteesi.



Kuva 1: Tutkielman systemaattisen tiedonhaun prosessi.

Tutkielmaan on kerätty tieteellisiä vertaisarvioituja lähteitä tietokantahakujen avulla tunnetuilta tieteellisiltä foorumeilta. Lähteitä on pyritty keräämään tutkielman aiheen hyvistä sekä huonoista puolista, jotta aihetta voidaan kriittisesti vertailla eri näkökulmista. Tarkoituksena on ollut kerätä hyvä kokonaiskuva tutkimuskysymyksen tilanteesta ja pohtia

mahdollisia tietoaukkoja sekä niiden jatkotutkimusmahdollisuuksia. Aineistohakuun on hyödynnetty Scopus AI tekoälyä sekä seuraavia tietokantoja: Computer Science Database (ProQuest), IEE Xplore – IEE Electronic Library (EEL) ja ScienceDirect. Scopus AI haun perusteella aineistoja on valittu myös muista tietokannoista ja julkaisualustoista. Tutkielman lähteinä on ainoastaan tieteellisiä ja vertaisarvioituja tutkimuksia. Tutkielman tieteellisten lähteiden sisäänotto- ja poissulkukriteerit on ilmoitettu taulukossa 1.

Taulukko 1: Tutkielman tieteellisten lähteiden sisäänotto- ja poissulkukriteerit.

Sisäänottokriteeri	Poissulkukriteeri
Tieteellinen ja vertaisarvioitu.	Ei ole saatavilla kokotekstinä tai se on rajoitetusti saatavilla.
Julkaistu vuosien 2019–2025 aikana.	On julkaistu ennen vuotta 2019.
Englanninkielinen.	Muu kuin englanninkielinen.
Käsittelee aihetta tai tuottaa tuloksia, jotka ovat olennaisia tämän tutkielman tutkimuskysymysten kannalta.	Ei käsittele lokidataa, poikkeamien tunnistamista tai päätöksentekoon liittyviä sovelluksia tai niitä ei voi yhdistää yritys kontekstiin.
Peräisin tunnetulta tieteelliseltä foorumilta.	On blogi, uutisartikkeli, tiivistelmä, esitys tai muu vastaava lähde.
Sisältää riittävät tiedot menetelmistä ja tuloksista, jotta sen sisältöä voidaan arvioida kriittisesti.	On suppea, ei sisällä riittävästi tietoa tutkimusmenetelmistä tai tuloksista analysointia varten.

Tiedonhaku suoritettiin syyskuun ja maaliskuun välillä vuonna 2025, ja hakulauseina on toiminut:

1. ("Data Mining" AND "Data Analysis" AND "Data visualization") OR ("data-driven culture" AND "Data-Driven Transformation" AND "challenges" AND "success") OR ("Business Intelligence" AND "Knowledge Engineering")
2. ("Event logs" OR "Log Data") AND ("Behavior-Based" AND "Anomaly Detection") OR ("IQR-Based" AND "Detection")
3. ("Log analysis" OR "system logs") AND ("survey" OR "Research" OR "study") AND ("reconstruction" OR "unstructured data" OR "Splunk")

Scopus AI:n deep researchiä on hyödynnetty sopivien lähteiden etsinnässä samoilla kriteereillä, ja hakulauseena on toiminut "Scientific articles peer-reviewed between the years 2019 and 2025 about ("process mining" AND "event logs" AND "data quality" AND "preprocessing" AND ("performance analysis" OR "process insights")) OR Business intelligence data-driven culture or its success and challenges in analysis". Tutkielmassa on käytetty useita hakulausekkeita, siitä syystä, että joissain tietokannoissa on boolean ehto rajoitteita, kuten ScienceDirectissä, jossa rajoitteena on maksimissaan kahdeksan ehtoa. Kuitenkin on tärkeää rajata hakulausekkeet mahdollisimman tarkkaan, jotta saadaan sopiva määrä tuloksia ja erilaisia näkökulmia tutkimuskysymyksen kannalta. Tutkimuskysymyksen vastaamiseen on pyritty neljällä hakulausekkeella saamaan tutkimuksia datan visualisoinnista ja sen hyödyntämisestä, anomalian tunnistamisesta ja lokien analysoinnista. Scopus AI:ta hyödynnettiin vain vähän, mutta kokeiluna. Tarkoituksena oli kokeilla, kuinka hyödyllinen kyseinen tekoäly on, ja löytääkö sen avulla hyödyllisiä lähteitä. Kuitenkin tuli havaittua, että käyttämällä samaa hakulauseketta, niin Scopus AI myös tarjoaa samantyyppisiä lähteitä uudestaan.

2.3 Aineiston seulonta ja esittely

Aihe on suosittu yritysmaailmassa ja tutkijoiden keskuudessa, minkä takia erilaisia tutkimuksia on tehty big datasta, lokidatasta, poikkeamien tunnistamisesta ja visualisoinnista. Tieteelliset lähteet on valittu työhön seulonnan avulla sillä perusteella, kuinka hyvin tieteellinen aineisto sopii aiheeltaan ja tutkimuksen tuloksiltaan tähän tutkielmaan. Tavoitteena on soveltaa aiheen ymmärrystä yritys kontekstiin tai päätöksentekoon. Seulonnassa on tarkoituksena löytää tutkimuksia, joita voidaan tulkita yleisellä tasolla tutkimuskysymyksen hyvistä tai huonoista puolista.

Lähteiden seulomiseen on käytetty kolmea eri seulontavaihetta. Ensimmäisessä seulonnassa on kerätty useita tutkimuksia, jotka vaikuttavat otsikon, tiivistelmän ja taulukon 1. kriteereiden perusteella sopivilta. Tässä vaiheessa on myös painotettu tutkimuksia, joissa on enemmän, kuin vain yksi tekijä. Toisessa seulonnassa vertaillaan tutkimuksia toisiinsa ja arvioidaan mitä lisäarvoa tutkimus tuo tähän tutkielmaan. Lähteistä on luettu johdanto ja johtopäätelmät, ja niistä on silmäillen tarkistettu tulokset, joiden perusteella lähteistä on valittu ainoastaan ne, jotka sopivat tutkielmaan parhaiten ja tuovat arvoa tutkimuskysymysten kannalta. Kolmannessa seulonnassa luetaan jokainen lähde alusta loppuun ja tehdään lopullinen päätös siitä, että otetaanko tutkimus tähän tutkielmaan

mukaan. Taulukossa 2 on esitetty tutkielman aineiston keruu, jossa mainitaan aineistohaun hakulauseke, miltä vuodelta, aineiston kieli ja monta kappaletta on löytynyt tiedonhaussa.

Taulukko 2: Tutkielman aineiston keruu.

Hakulauseke	("Data Mining" AND "Data Analysis" AND "visual analytics") OR ("Monitoring" AND "Processing" AND "Log data" AND "visualisation tool") OR ("Business Intelligence" AND "Knowledge Engineering")	("Event logs" OR "Log Data") AND ("Behavior-Based" AND "Anomaly Detection") OR ("IQR-Based" AND "Detection")	("Log analysis" OR "system logs") AND ("survey" OR "Research" OR "study") AND ("reconstruction" OR "unstructured data" OR "Splunk")
Vuosi	2019-2025	2019-2025	2019-2025
Kieli	englanti	englanti	englanti
Vertaisarvioitu	Kyllä	Kyllä	Kyllä
Tulosten määrä yhteensä (n)	892	228	168
1. seulonnan jälkeen	30	20	20
2. seulonnan jälkeen	5	5	3
3. seulonnan jälkeen	3	3	3
Yhteensä	3	3	3

Yhteensä tutkimuskysymysten vastaamiseen on valikoitunut 11 kappaletta tieteellistä lähdettä. Tietokannoista löytyi yhdeksän ja Scops AI:n avulla kaksi tutkimusta tekijöiltä Dakic ja muut (2023) sekä Storm ja Borgman (2020). Talukoissa 3a ja 3b on esiteltynä

tutkielman aineistot, joita käytetään luvussa 4 tulokset ja keskustelu -osiossa. Tutkielmassa on yhteensä 20 lähdettä, joista kaikki on tieteellisiä ja monia niistä käytetään seuraavassa luvussa teoreettisen taustan esittämisessä.

Taulukko 3a: lähteet 1–5, jotka löytyvät systemaattisen tiedonhaun kautta.

Tekijät	Vuosi	Aihe	Julkaisupaikka
Dakic ja muut	2023	Event Log Data Quality Issues and Solutions.	<i>Mathematics</i> , vol. 11, no. 13, pp. 2858.
Hany ja muut	2023	Framework for automatic detection of anomalies in DevOps.	Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 3, pp. 8–19.
He ja muut	2022	An empirical study of log analysis at Microsoft.	<i>ESEC/FSE 2022: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering</i> , pp. 1465–1476.
Li ja muut	2022	Research on Real-time Log Data Processing And Monitoring Scheme of Printing Equipment Based on Flink Framework.	<i>EITCE '22: Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering</i> , pp. 1096–1100.
Partovian ja muut	2023	Analysis of log files to enable smart-troubleshooting in Industry 4.0: a systematic mapping study.	<i>IEEE Access</i> , vol. 12, pp. 147640–147658.

Taulukko 3b: lähteet 6–11, jotka löytyvät systemaattisen tiedonhaun kautta.

Sitova ja Pec-erska	2020	Process Data Analysis Using Visual Analytics and Process Mining Techniques.	<i>2020 61st International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS)</i> , pp. 1–6.
Skopik ja muut	2023	Behavior-Based Anomaly Detection in Log Data of Physical Access Control Systems.	IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 4, pp. 3158–3175.
Song ja muut	2020	Self-Healing Event Logs.	IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 6, pp. 2750–2763.
Storm ja Borgman	2020	Understanding challenges and success factors in creating a data-driven culture.	Proceedings of the 53rd Hawaii International Conference on System Sciences, pp. 5399–5408.
Studiawan ja muut	2019	A survey on forensic investigation of operating system logs.	Digital Investigation, Elsevier Ltd, vol. 29, pp. 1–20.
Swami ja muut	2023	IQR-based approach for DDoS detection and mitigation in SDN.	Defence Technology, vol. 25, pp. 76–87.

Vanhin lähde, joka valikoitui tutkielmaan, on vuodelta 2019 Studiawan ja muiden (2019) tekemä tutkimus forensiikasta ja systeemilokeista.

3 Teoreettinen tausta

Luvussa käsitellään tutkimuksen kannalta oleellista taustaa lokidatasta. Luvussa määritellään keskeisimpiä käsitteitä sekä avataan lokidatan perusteita ja kokonaisuutta isoon dataan. Luvussa avataan tarkemmin visualisoinnin ja poikkeamien tunnistamisen merkitystä osana lokidatan hyödyntämistä. Luvun tarkoituksena on antaa selkeä lähtökohta lukuun 4, jossa käsitellään tutkimuskysymyksien tuloksia. Keskeisiä käsitteitä tutkielman kannalta ovat lokidata, lokimerkintä, big data, poikkeamien tunnistaminen, splunk ja visualisointi.

3.1 Lokidatan määritelmä ja tärkeys

Ma ja muut (2023) totesivat tutkimuksessaan, että keskeinen data, johon tallentuu ohjelmiston toimintaan liittyvää tietoa, kuten luotettavuuden varmistaminen, on ohjelmistojärjestelmien lokit. Tutkimuksessa myös todettiin, että lokeilla on todettu olevan erittäin tärkeä rooli teollisuudessa, kyberturvallisuudessa tunkeutumisen havaitsemisessa, forensiikassa ja tilannetietoisuudessa, sillä pienimmätkin poikkeamat voivat aiheuttaa merkittäviä häiriöitä. Ma ja muut (2023) löysivät tutkimuksessaan myös erilaisia tilanteita, joissa lokidatan analysointi olisi hyödyntänyt yrityksiä. Esimerkiksi heinäkuussa vuonna 2022 Microsoftin ja Googlen ohjelmistojärjestelmissä tapahtui poikkeustilanteita, joissa käyttäjät eivät päässeet palvelimeen yli tunnin ajan tai ohjelmistoja ei pystytty käyttämään normaalisti tallennuspalvelun ongelmien vuoksi.

Patil ja muut (2025) luokittelivat lokit eri kategorioihin: Järjestelmä-, tietoverkko- ja pilvilokeihin. Järjestelmälokit voivat pitää sisällään esimerkiksi ohjelmistojen tai käyttöjärjestelmien lokit. Tietoverkkolokit pitävät sisällään tietoliikenteeseen liittyvät lokit, kuten ruuhkat, tietoliikenteiden yhteydenotot sekä suorituskykyyn ja uhkaan liittyvät lokit. Pilviloikeissa on järjestelmien ja palveluiden lokitapahtumat. Jokaisesta kategoriasta voidaan jäsentää tai analysoida lokeja (engl. log parsing). Zhang ja muut (2023) totesivat tutkimuksessaan, että järjestelmälokit täytyy saada merkitykselliseen rakenteelliseen muotoon tavallisesta raakadatasta, minkä jälkeen lokimerkinnöistä voidaan tehdä hyödyllisiä tulkintoja.

3.2 Lokidata osana big dataa

Big data on tällä vuosikymmenellä suosittu aihe. Big dataa hyödyntävät monet suuret yritykset ja datasta on hyötyä tekoälyn kouluttamisessa. Esimerkiksi tieto ja ymmärrys omien järjestelmien toiminnoista ja käyttäjien käyttäytymisestä on yrityksille valtava apu, kun he pyrkivät parantamaan yrityksen sisäistä päätöksentekoa (Alshawawreh et al, 2024). Big data koostuu viidestä osa-alueesta: määrä (engl. volume), nopeus (engl. velocity), monimuotoisuus (engl. variety), totuudellisuus (engl. veracity) ja arvo (engl. value). Termit viittaavat syntyvään ja tallennetun datan määrään, datan suureen kasvunopeuteen, datan erilaisiin tyypeihin, muotoihin ja laatuun. (Zhang et al, 2023) Big data on määritelmältään laaja, minkä ansiosta lokidata on osa big datan kokonaisuutta tiedonmäärältään, kasvunopeudeltaan, tyypeiltään, muodoiltaan ja datan laadultaan.

Li ja muut (2025) totesivat, että big datan aikakauden myötä erilaiset teollisuuden koneet, kuten tietosensorit, koneet tai muut laitteet tuottavat käytön aikana suuria määriä lokitietoja. Lokitietojen keräämisestä, tarkkailusta ja hyödyntämisestä haastavaa tekee se, että lokitiedostot ovat laajamittaisia, monimutkaisia, jatkuvasti muuttuvia ja reaaliaikaisia. Lokitietojen analysoinnissa voidaan havaita piileviä arvoja koneiden ja laitteiden valvonnassa, hälytysten optimoinnissa tai yritysten laitteiden tehokkuuksien parantamisessa.

3.3 Lokien jäsentely ja analysointi

Patil ja muiden (2025) sekä Zhang ja muiden (2023) mukaan raakadatasta täytyy poimia tieto, mikä on yrityksen tai tutkimuksen kannalta oleellinen, ja epäoleelliset tiedot tulee rajata pois. Tutkimuksessa (Zhang et al., 2023) todettiin, että raakadata, joka voi olla yhdessä käyttötarkoituksessa hyödyllinen, ei välttämättä ole toisessa tutkimuskohteessa hyödyllinen. Patil ja muut (2025) totesivat, että lokien analysointia varten lokimerkinntät tulisi pilkkoa rakenteellisiin osiin, kuten aikaleimat, tapahtuma, virhekoodi tai käyttäjätunnus. Jäsentelyjen jälkeen lokimerkintä halutaan tiettyyn rakenteelliseen muotoon kuten yleisimmin JSON (JavaScript Object Notation) tai taulukkomuotoon. Kun lokimerkinnät on saatu rakenteelliseen muotoon ja epäoleellinen tieto on karsittu pois, voi niitä alkaa analysoida ja hyödyntämään päätöksenteossa. Kyseisiä lokimerkintöjä kutsutaan yleisimmin tapahtuma lokeiksi.

transaction_id	user_id_from	user_id_to	date	value
1	2	2	2013-04-10 14:22:50	24.375
1	2	782477	2013-04-10 14:22:50	0.7709
3	3	782479	2013-04-10 14:22:50	47.1405196
3	3	4	2013-04-10 14:22:50	150.0
6	5	782480	2013-04-10 14:22:50	65.45
6	5	782481	2013-04-10 14:22:50	34.55
8	6	782482	2013-04-10 14:22:50	6.90778707
8	6	7	2013-04-10 14:22:50	3.0
10	782483	782484	2013-04-10 14:22:50	8.0
10	782483	782483	2013-04-10 14:22:50	2.90838

Kuva 2: Lokimerkintä rahansiirrosta Splunk-työkalun testidatasta.

Kuvassa 2 on esimerkkinä havainnollistettu Splunkin avulla jäsenellyt lokimerkinnät. Kuvasta havaitaan, miten esimerkiksi tässä tapauksessa voidaan saada tietoa kahden henkilön välisestä rahaliikenteestä, ja siitä, mihin aikaan kukin tapahtuma on tapahtunut.

Model	CapacityBytes	DiskFailure	SerialNumber	probable_cause
HGST HMS5C4040ALE640	-9.12E+18	Yes	PL2331LAGPJ89J	CapacityBytes
HGST HMS5C4040ALE640	1.07E+17	Yes	PL2331LAGSU5RJ	CapacityBytes
Hitachi HDS722020ALA330	2.00E+12	Yes	JK2171B9J0G26L	
Hitachi HDS722020ALA330	2.00E+12	Yes	JK11A8B9J7EZJF	
HGST HMS5C4040ALE640	4.00E+12	No	PL2331LAGUHZ6J	
ST31500341AS	1.50E+12	Yes	9VS21JD2	
ST31500541AS	1.50E+12	Yes	6XW05BBP	
HGST HMS5C4040ALE640	4.00E+12	No	PL2331LAGULGGJ	

Kuva 3: Kovalevyn onnistumiset ja epäonnistumiset Splunk-työkalun testidatasta.

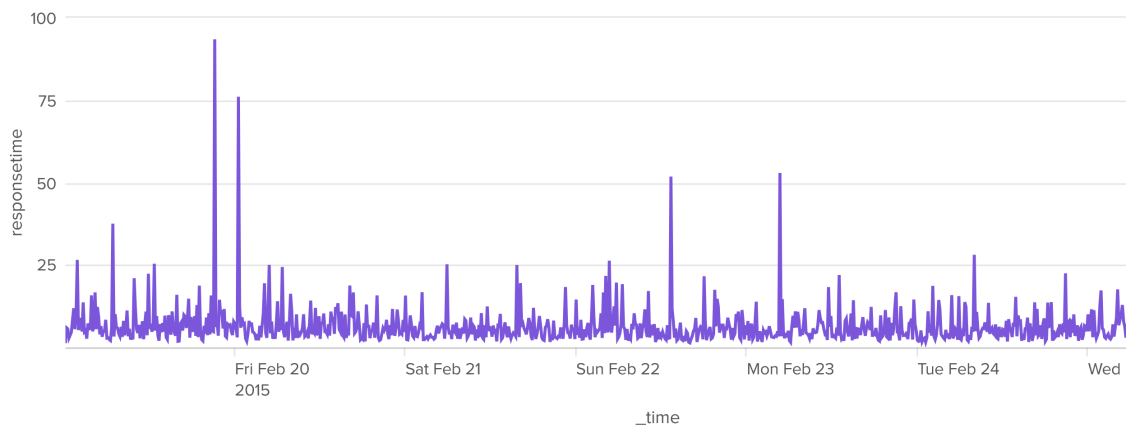
Kuvan 3 lokimerkinnöistä voidaan havaita suoraan, mikä kovalevy on epäonnistunut toiminnassaan, ja mistä se on voinut tapahtua. Tämän havainnollistamiseksi järjestelmien lokidatasta on jäsenellyt ainoastaan tiedot kovalevyn suorituskykyyn liittyvistä lokimerkinnöistä.

3.4 Visualisointi ja poikkeamien tunnistaminen

Shakeel ja muut (2022) tutkivat tutkimuksessaan, että tehokkaat ja interaktiiviset datavisualisoinnit ovat tärkeitä, koska visualisoinnin avulla voidaan saada helposti ymmärrettäviä visuaalisia tietoja monimutkaisista raakadatoista. Tutkimuksessa todettiin, että visualisoinnin tavoitteena on myös auttaa päätöksentekijöitä

hahmottamaan olennainen tieto nopeasti, ja auttaa heitä löytämään piilotettuja yhteyksiä ja uusia oivalluksia. Lopuksi Shakeel ja muut (2022) totesivat, että datavisualisointi ei siis ole vain graafeja ja kaavioita, vaan työväline, jonka avulla voidaan suurista aineistoista löytää ilmiöitä, joita olisi vaikea havaita ilman visualisointia. Tämän takia visualisointi on ratkaiseva tekijä parempien päätöksentekojen kannalta.

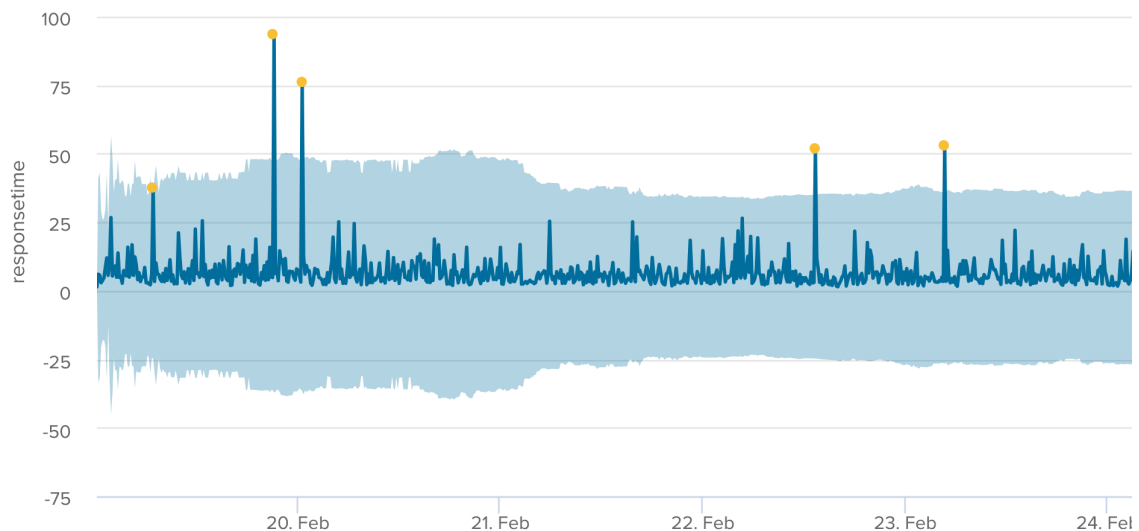
Sitova ja Pecerska (2020) totesivat, että lokeja voidaan esittää visuaalisten jäsentelyjen ja louhintamenetelmien avulla, mikä edistää päätöksenteon prosessia. Lisäksi todettiin, että visualisointitekniikoilla voidaan onnistuneesti luoda synergiaa lokien analytiikan ja datan välille, joten menetelmällä yhdistetään visualisointi, ihminen ja data tiedon hankkimiseksi. Lokien analyttiseen visualisointiin on erilaisia kaavioita, kuten ympyrä- ja viiva-kaavio sekä histogrammi. Aikaisemmassa alaluvussa esitettiin kuvissa 2 ja 3 lokimerkintöjä, jotka on saatu päätöksenteon kannalta hyödylliseen muotoon raakadatasta. Useita tuhansia lokimerkintöjä voidaan esittää visualisoituna, ja niistä voidaan analysoida mahdollisia muutoksia tietyn ajanjakson aikana. Kuvassa 4 on havainnollistettuna visuaalisesti palvelimen vasteaika tietyllä ajanjaksolla.



Kuva 4: Palvelimen vasteaika havainnollistettuna Splunk-työkalun testidatasta.

Kuvan perusteella voidaan päätellä, että visualisointi helpottaa suuren lokimäärän havainnollistamista esimerkiksi, kun tarkastellaan tuotannon järjestelmän suorituskykyä tietyllä ajanjaksolla. Visualisoinnin tukena on poikkeamien tunnistaminen. Poikkeamien tunnistaminen tarkoittaa, että kyseisestä aineistosta löydetään näytteitä, joita ei voida luokitella normaaleiksi. Toisin sanoen näytteet ovat anomaleja (Huang et al., 2025).

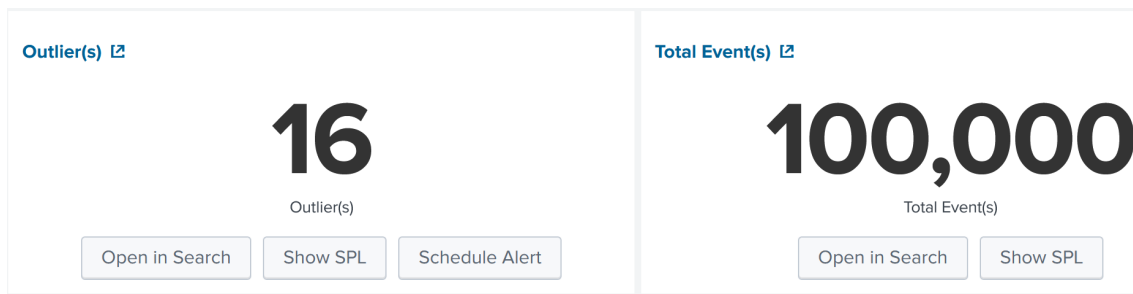
Splunk-työkalun avulla on mahdollista hyödyntää kahta menetelmää poikkeamien tunnistamiselle, jotka ovat IQR (engl. Interquartile Range, Interkvartiiliväli) ja MAD (engl. Median Absolute Deviation, Mediaanin absoluuttinen keskipoikkeama). Swami ja muut (2023) selittivät, miten IQR sopii parhaiten datalle, jossa esiintyy epätasaisuutta ja poikkeavia arvoja, sillä IQR ei ole yhtä herkkä ääripään anomaleille verrattuna MAD:iin. IQR perustuu datan mediaaniin ja kvartiileihin, ja se kuvaa aineiston keskimmäistä 50 %:a ($Q3 - Q1$), jossa $Q3$ ja $Q1$ saadaan laskemalla mediaani pienemmästä ja suuremmasta puolikkaasta aineistoa. IQR-menetelmällä tyypillisesti lasketaan ylä- ja alaraja, jonka ylittävä poikkeava data tulkitaan anomaliaksi. Yleisesti ala- sekä yläraja lasketaan kertoimena. Ylärajalle esimerkiksi $Q3 \cdot 1,5$ ja alarajalle $Q1 \cdot 1,5$. Kuvassa 5 on havainnollistettuna visuaalisesti uudestaan kuvan 4 palvelimen vasteaika, mutta lokidatan poikkeavien arvojen rajausta on tehty IQR-menetelmällä kertoimena 6.



Kuva 5: Palvelimen vasteaika IQR-menetelmällä.

Kuvasta on nähtävissä, kuinka Splunk hälyttää IQR-rajauksen ylimenevistä poikkeavista arvoista. Kuvan 5 poikkeavat arvot on mahdollista havaita ilman poikkeamien tunnistamista, mutta poikkeamien tunnistamisen etuja ovat hälytettävyyden, automaattisuuden ja se, että se edesauttaa nopeaa reagointia ongelmatilanteissa (Hany et al., 2023; Skopik et al., 2023). IQR-menetelmän avulla on mahdollista antaa tarkat rajaukset siitä, milloin arvo voidaan tulkita poikkeavaksi. Kun poikkeava arvo on havaittu, siitä voidaan Splunkin ansiosta tehdä hälytys, jonka voi lähettää suoraan työntekijöiden sähköpostiin, Teams-kanavalle tai suoraan kännykkään. Poikkeamien tunnistamisen avulla voidaan Splunk-työkalun komento-ohjeiden avulla rajata kaikki lokimerkinnät, joissa on havaittu poikkeus,

ja näin voidaan tarkastella ainoastaan poikkeavia näytteitä. Tämä säästää huomattavasti aikaa, sillä poikkeamien tunnistamisen avulla ei tarvitse analysoida jokaista lokimerkintää, jotka eivät ole poikkeavia. Splunk-työkalun avulla voidaan luoda näkymä, josta avulla on nähtävissä helposti anomalioiden määrä, ja kuinka monta lokimerkintää järjestelmään on tullut yhteensä tietyllä ajanjaksolla. Kuvassa 6 on havainnollistettuna summauskomennolla yhteenveto järjestelmän tapahtumista.



Kuva 6: Yhteenveto palvelimen vasteajasta Splunk-työkalun testidatasta.

Tilastollisen IQR-menetelmän lisäksi on erilaisia koneoppimiseen ja kielimalleihin liittyvää anomalioiden tunnistusta. Patrovian ja muut (2023) tutkivat yhteensä 22 eri menetelmää lokien analysointiin. Niissä korostui erityisesti kielimallien ja koneoppimisen menetelmät. Tutkimuksessa todettiin, että ohjattu ja ohjaamaton oppiminen olivat lähes yhtä suosittuja. Ohjaamaton oppiminen oli hyödyllinen poikkeuksien tunnistamisessa. Ohjattu oppiminen vaatii koulutusta ja dataa, ennen kun ohjattua oppimista voi hyödyntää analysoinnissa, kuten lokien luokittelemisessä. Tutkimuksessa oli tutkittu muitakin menetelmiä, kuten syväoppimista, neuroverkkoja, ryvästämistä (engl. clustering) ja luokittelua lokidatan analysointiin. Esimerkiksi erilaisia neuroverkkomenetelmiä on tutkittu, kuten graafiverkko-neuroverkko (engl. graph neural network), toistuva neuroverkko (engl. recurrent neural network) ja konvoluutio neuroverkko (engl. convolutional neural network). Kyseiset neuroverkot lähinnä eroavat hieman siitä, miten poikkeamien havaitsemista halutaan toteuttaa.

4 Tulokset ja keskustelua

Tutkimuksissa on havaittu useita hyviä puolia lokidatan ja visualisoinnin hyödyntämisestä päätöksenteossa. Esimerkiksi He ja muut (2022) totesivat, kuinka tilastolliset menetelmät ovat hyödyllisiä lokidatan analysoinnissa. Toisaalta Hany ja muut (2023), Skopik ja muut (2023) sekä Swami ja muut (2023) huomasivat tutkimuksissaan, että poikkeamien tunnistamisesta on hyötyä lokien analysoinnissa. Poikkeamien tunnistamisen lisäksi Partovian ja muut (2023) totesivat, että koneoppimisen menetelmien käytöstä on havaittu positiivisia tuloksia, kun koneoppimista on hyödynnetty lokien analysoinnissa. Studiawan ja muiden (2019) mukaan lokidata on myös oleellinen osa kyberturvallisuutta ja forensiikkaa. Sitova ja Pecerska (2020) päätyivät tutkimuksessaan tulokseen, että visualisoinnilla on positiivisia vaikutuksia, kun visualisointia hyödynnetään päätöksenteossa. Song ja muut (2021) tutkivat tutkimuksessaan itseparantuvista (engl. self-healing) lokeista, koska lokidatan huononlaatuiset ongelmat ovat yksi merkittävimpiä ongelmia lokidatassa.

Lokidatasta löytyi tutkimuksista myös huonoja puolia. He ja muiden (2022), Sitovan ja Pecerskan (2020), Song ja muiden (2021) sekä Dakic ja muiden (2023) tutkimusten mukaan yleisimmät huonot puolet ovat, että lokeja on liikaa, niitä puuttuu tai lokit ovat huonolaatuisia. Toisaalta He ja muut (2022) sekä Li ja muut (2022) totesivat myös, että lokeja on ajoittain vaikea ymmärtää, tai on vaikea hahmottaa, mikä lokeissa on normaalia ja epänormaali, kun niitä kerätään useista järjestelmistä. Lisäksi Partovian ja muut (2023) tutkivat erilaisia koneoppimisen hyödyntämiseen liittyviä vaikeuksia. Song ja muut (2021) sekä He ja muut (2022) totesivat tutkimuksissaan, että visualisointityökalujen puuttuminen on vaikeuttanut lokien analysointia. Kuitenkin Storm ja Borgman (2020) totesivat tutkimuksessaan, että vaikka visualisointia hyödynnetään, niin päätöksenteko perustuu ajoittain silti intuitioon.

Taulukkoon 4 on yhteenvedona esitetty lokidatan hyvät ja huonot puolet. Vaikka lokidatasta löytyy useita huonoja puolia voidaan silti todeta, että lokien hyödyntämisestä löytyy positiivisia ja merkittäviä hyötyjä, kun lokidataa hyödynnetään päätöksenteossa yhdessä visualisointikeinojen avulla. Seuraavissa alaluvuissa tarkennetaan yhteenvedossa olevia tutkimuksia.

Taulukko 4: Lokidatan hyödyntämisen hyvät ja huonot puolet päätöksenteossa.

Hyvät puolet	Huonot puolet
Tilastolliset menetelmät tukena, kuten sum(), count(), avg(). (He et al., 2022)	Liikaa, puuttuvia tai huonolaatuisia lokeja. (He et al., 2022; Sitova & Pecerka, 2020; Song et al., 2021; Dakic et al., 2023)
Oleellinen kyberturvallisuudessa ja forensiikassa. (Studiawan et al., 2019)	Lokeja on toisinaan vaikea lukea. Lokeista on vaikea havaita, mikä on normaalia ja mikä epänormaalia. Lokien kerääminen useista eri järjestelmistä ja tietokannoista aiheuttaa vaikeuksia. (He et al., 2022; Li et al., 2022)
Lokidata auttaa IT-järjestelmien analysoinnissa. (Li et al., 2020; Song et al., 2021)	Havainto visuaalisten ominaisuuksien puuttumisesta analysoinnissa. (He et al., 2022; Song et al., 2021)
Lokidata on reaaliaikainen. (Li et al., 2020)	Vaikka visualisointia hyödynnetään, päätöksenteko perustuu ajoittain edelleen intuitioon eikä faktaan. (Storm & Borgman, 2020)
Visualisointi auttaa päätöksenteossa. (Sitova & Pecerska, 2020)	Koneoppimisen on vaikea hyödyntää. (Partovian et al., 2023)
Tutkimuksia itse parantavista lokeista. (Song et al., 2021)	
Poikkeamien tunnistaminen auttaa lokien analysoinnissa. (Hany et al., 2023; Skopik et al., 2023; Swami et al., 2023)	
Koneoppimisen menetelmät tukena. (Partovian et al., 2023)	

4.1 Tutkimuskysymys 1: Mitä hyviä ja huonoja puolia lokidatan hyödyntämisellä voi olla päätöksenteossa?

Lokidatan merkittävimmät hyödyt liittyvät sen tarjoamaan reaaliaikaisuuteen ja monipuolisten analyysimenetelmien mahdollisuuksiin. Li ja muut (2022) totesivat, että lokidata mahdollistaa nopeamman puuttumisen poikkeaviin tilanteisiin. Myös Studiawan ja muut (2019) totesivat, että lokidata on oleellinen osa kyberturvallisuutta, jolloin sitä voidaan hyödyntää esimerkiksi oikeudessa rikosten selvittelyssä. Tutkimuksessa (Studiawan et

al., 2019) todetaan, että lokidatan avulla on mahdollista parantaa yrityksen sisäistä kyberturvallisuutta ja reagoida nopeasti mahdollisiin tunkeutumisiin. Lopuksi tutkimuksessa todettiin, että lokidatan ansiosta on mahdollista vertailla erillisten tapahtumien suhteita toisiinsa ja hyödyntää kyberturvallisuuteen liittyvissä hyökkäysten rekonstruoinnissa. He ja muiden (2022) mukaan erilaiset tilastolliset menetelmät tukevat analysointia, sillä tilastollisten menetelmien avulla voidaan tiivistää suuria lokimassoja, jotka helpottavat kriittisten piirteiden tunnistamista.

Kriittisesti tarkasteltuna aikaisemmin todetut hyödyt edellyttävät kuitenkin oletuksia, jotka eivät välttämättä toteudu tosielämän järjestelmissä. Tilastolliset menetelmät nojaavat siihen, että analysoitava lokidata on riittävän laadukasta ja kattavaa. Todellisuudessa lokidatan laatuongelmat ovat yksi keskeisimmistä haasteista analysoinnissa. Dakic ja muut (2023) totesivat, että yleisimmät laatuongelmat lokidatassa ovat puuttuvat, virheelliset, epätarkat ja epärelevantit datat, jotka vievät analysoinnissa paljon aikaa analysoinnissa turhaan. Kuitenkin Songin ja kumppaneiden tutkimuksen (Song et al., 2021) mukaan, ilman strukturoituja prosessimalleja on heidän ehdottama analysointimenetelmä yksi ensimmäisistä, jonka avulla tapahtumalokeja voidaan korjata, joka perustuu ryvästämiseen ja heuristiikkoihin perustuviin menetelmiin. Tutkimuksen mukaan ryvästämisen jälkeen haetaan tapahtumasegmenttejä suoritusten perusteella, jossa on samoja tapahtumia useita kertoja. Sen jälkeen ryvästämisen sisäiset transitiivisten esiintymissuhteiden perusteella tunnistetaan alikulut (engl. Sub-processes, fragments), joiden ansiosta korjatut tapahtumasekvenssit johdetaan alikulujen löydöksiä pohjalta.

Kaiken kaikkiaan lokidatan käytettävyys kärsii usein sen massiivisesta määrästä. He ja muut (2022) totesivat, että eri tiimit voivat kerätä lokeja eri järjestelmistä eri formaateissa, mikä voi vaikeuttaa lokidatan analysointia ja ymmärrettävyyttä, ja se voi lisätä virhetulintoja. Tämä rajoittaa monien tutkimuksissa esiteltyjen analyysimenetelmien käytännön soveltuvuutta. Huomionarvoista on, että Partovian ja muiden (2023) mukaan reaaliaikaisuus ei aina toteudu, sillä osa järjestelmistä hyödyntää eräkäsittelyä, mikä voi viivästyttää päätöksentekoa kriittisissä tilanteissa. Eräkäsittelyn periaate on, että data analysoidaan erissä, eikä reaaliajassa.

Yhteenvetona voidaan todeta, että lokidatalla on useita ja selkeitä hyötyjä, joiden avulla voidaan edistää päätöksentekoa ja ennaltaehkäistä mahdollisia häiriötilanteita. Lokidatan hyödyntämiseen vaaditaan kuitenkin laadukasta lokidataa ja analysoinnin osaamista. Lokidatan laatuongelmiin on pyritty selvittämään mahdollisia korjauskeinoja, kuten Song ja muiden (2021) tutkimuksessa itse parantavista lokeista.

4.2 Tutkimuskysymys 2: Miten lokien visualisointi ja poikkeamien tunnistaminen tukevat päätöksenteon prosessia?

Useat tutkimukset (Hany et al., 2023; Skopik et al., 2023; Swami et al., 2023) korostavat, että keskeinen työkalu lokianalyysin tehokkuuden parantamisessa on poikkeamien tunnistaminen. Poikkeamien tunnistaminen mahdollistaa tarkasteltavan lokidatan määrän huomattavan vähenemisen ja mahdollistaa helpommin kriittisten poikkeavien arvojen erottamisen normaaleista näytteistä. Partovian ja muut (2023) toteavat, että poikkeamat voidaan tunnistaa koneoppimisen ja kielimallien avulla yhä tarkemmin.

Kriittisen arvioinnin näkökulmasta poikkeamien tunnistamiseen liittyy kuitenkin rajoitteita ja haasteita. Ensinnäkin Partovian ja muiden (2023) mukaan kaikki menetelmät eivät kykene ennustamaan poikkeamia tai selittämään niiden perimmäistä syytä, mikä rajoittaa niiden käytettävyyttä päätöksenteossa. Mallit ovat herkkiä datan vaihtelulle ja erityisesti huonolaatuisille lokidatoille. Lisäksi kielimallien käyttö edellyttää koneoppimisen menetelmien osaamista sekä aikaa ja resursseja. Ne saattavat kuitenkin olla rajoittavia tekijöitä etenkin yrityksille, joille saattaa tulla huomattavia kustannuksia koneoppimisen menetelmien käyttöönotosta.

Visualisoinnin osalta tutkimukset nostavat esiin sekä huomattavia hyötyjä että rajoitteita. Sitova ja Pecerska (2020) huomasivat, että suurten lokimäärien tulkinnessa visualisointi auttaa huomattavasti. He ja muut (2022) totesivat, että Microsoftissa lokidatan analysoinnissa toivotaan enemmän visualisointityökaluja, jotka tukevat lokidatan analysointia. Siitä huolimatta visualisointi voi tuoda esiin piileviä anomaliaita, tai se voi helpottaa kokonais kuvan hahmottamista. Näitä olisi vaikea havaita pelkästään yksittäisiä lokimerkin-
töjä tutkimalla.

Toisaalta Storm ja Borgman (2020) nostavat esiin tärkeän näkökulman: vaikka visualisointi ja datan analysoinnin työkaluja olisi käytössä, päätöksenteko ei siitä huolimatta

välttämättä perustu tilastoihin, vaan intuitioon ja aikaisempaan kokemukseen. Tutkimuksen mukaan, jotkut työntekijät saattavat tukeutua aikaisempaan kokemukseen, koska “näin on aina tehty”, mikä rajoittaa paremman päätöksenteon toteuttamista lokidatan analysoinnissa. Visualisoinnin haasteina voidaan nähdä myös se, että huonolaatuinen lokidata voi antaa väärän kuvan visualisoinnissa, jolloin lokidataa on vaikea analysoida, ja se saattaa aiheuttaa huonoja päätöksentekoa. Storm ja Borgman (2020) tuovat myös esiin, että vastuu työkalujen opettamisesta ja niiden käytön varmistamisessa on esihenkilöillä. Heidän tulisi katsoa, että työkaluja osataan käyttää oikein ja että päätöksenteko perustetaan tilastoihin.

Kokonaisuutena visualisointi ja poikkeamien tunnistaminen ovat tärkeitä tekijöitä päätöksenteon kannalta, mutta vain silloin, kun yrityksillä on riittävä osaaminen, luotettava data ja oikeat työkalut. Kaikessa muussa tapauksessa, jos osaamista tai resursseja ei ole, voi kaikki ylimääräiset työkalut tuoda enemmän haasteita, kun hyötyjä.

4.3 Tutkielman luotettavuus

Tutkielma toteutettiin noudattamalla systemaattisen kirjallisuuskatsauksen periaatteita. Se tukee tutkimuksen luotettavuutta ja toistettavuutta. Kaikki tieteelliset ja vertaisarvioidut lähteet on löydetty tietokannan avulla tai Scopus AI:n avulla. Erilaisia tietokantoja on käytetty yhteensä 5 kappaletta. Tutkielman kirjoitusprosessin aikana on saatu jatkuvasti vertaispalautetta muilta tutkielman tekijöiltä ja myös ulkopuolisilta. Tutkielmaan on pyydetty kriittistä näkökulmaa ja palautetta tutkielman ohjaajan lisäksi läheisiltä ja tutuilta, jotka erikoistuvat tai ovat erikoistuneet omaan tieteenalaansa esimerkiksi teknilliseen fyysiikkaan, hallintotieteisiin ja suomen kieleen. Luotettavuutta heikentää se, että kaikki lähteet on hankittu yksin, mikä voi lisätä subjektiivisuuden riskiä ja vaikuttaa tulkintoihin esimerkiksi sopivien lähteiden valinnassa ja niiden käsittelyssä. Rajoittavana tekijänä on myös se, että tutkimuksia on hankittu ainoastaan väliltä 2019–2025, minkä tarkoituksena on tarkastella ainoastaan tutkimuksen nykyhetkeä ja tutkimuksia siitä, mitä aiheesta tiedetään tällä hetkellä. Rajoittavana tekijänä on myös, että kaikki muut, kuin englanninkieliset tutkimukset ovat poissuljettu.

5 Yhteenveto

Tässä tutkielmassa käsiteltiin, mitä mahdollisuuksia ja haasteita lokidatan hyödyntämisellä päätöksenteossa voi olla, sekä miten visualisointi ja poikkeamien tunnistaminen tukevat tätä prosessia. Tutkielmassa käytiin läpi mistä lokidataa generoituu, mitä vaikutuksia big datalla on ja miten yritykset hukkuvat informaatiotulvaan nykyaikana. Tutkimuksessa todettiin erilaisia näkökulmia siihen, miten IT-järjestelmät sekä niistä syntyvät datat ovat keskeisessä roolissa yritysten päätöksenteossa ja menestyksessä. Tutkielmassa onnistuttiin vastaamaan tutkimuskysymyksiin kriittisesti systemaattisella kirjallisuuskatsauksella. Tutkielmassa pohdittiin tutkimuskysymyksen hyviä ja huonoja puolia.

Tutkielman aikana todettiin useita hyviä puolia lokidatalle, visualisoinnille ja poikkeamien tunnistamiselle. Lokidata helpottaa päätöksentekoa useilla tavoilla. Esimerkiksi lokidatan reaaliaikaisuuden ansiosta, tietoa voidaan kerätä ja siihen voidaan reagoida heti sen ilmetessä. Lokidatan analysointiin löytyy useita menetelmiä, joiden avulla voidaan helpottaa lokidatan analysointia. Visualisoinnista todettiin, että visualisoinnin puuttuminen on hankaloittanut lokien analysointia. Ilman visualisointia lokeja täytyy tutkia manuaalisesti yksi kerrallaan, vaikka siinä hyödynnettäisiin poikkeamien tunnistamista. Datasta voidaan tehdä helpommin tulkittavaa ja läpinäkyvää, ja se voi helpottaa päätöksentekoa hyödyntämällä visuaalisia analytiikan menetelmiä tapahtumalokien esittämiseen.

Lokidatasta on tutkittu myös useita huonoja puolia, kuten yleisimmin lokidatan laatuongelmia. Lokidatassa voi ilmetä virheellistä tietoa tai lokeista voi jäädä oleellisia tietoja pois. Lokeja voi syntyä valtava määrä, mikä hankaloittaa niiden analysointia. Valtavan lokidatan määrän vuoksi yritykset joutuvat seulomaan, mikä tieto on oleellista ja mikä ei. Poikkeuksien tunnistamisessa ilmeni myös ongelmia, kuten resurssi- ja kustannusongelmat. Tekoälyn kouluttaminen vaatii aikaa, osaamista ja resursseja, eikä ole yksinkertaisesti yhtä parasta menetelmää jokaiseen tapaukseen. Visualisoinnista todettiin positiivisia asioita, mutta siitä huolimatta yrityksissä saatetaan silti tukeutua yrityksen omiin aikaisempiin menetelmiin, eikä visualisoinnin tuloksiin.

Tutkielman aikana osoittautui selkeäksi, että lokidatan tutkimuksessa on tietoaukkoja. Monessa eri tutkimuksessa todettiin tutkimuksen puutteita lokien laadun korjaamisessa

esimerkiksi kielimallien tai koneoppimisen menetelmien avulla. Kielimallien ja koneoppimisen hyödyntäminen lokidatan analysoinnissa on myös osoittautunut olevan varhaisessa vaiheessa, ja vasta lähivuosina on pyritty ratkaisemaan ensimmäisiä askeleita lokidatan analysoinnissa kielimallien avulla. Yritysten kannalta löytyy huomattavan vähän tutkimuksia siitä, miten lokidataa voidaan hyödyntää mahdollisimman järkevästi. Yrityksillä on erilaisia haasteita lokidatan analysoinnissa, sillä esimerkiksi tekoälyn kouluttaminen vie resursseja, aikaa ja rahaa, eikä esimerkiksi Partovian ja muiden (2023) tutkimuksessa todettu, mikä 22 eri analysointimenetelmistä olisi kaikista paras, sillä menetelmä riippuu käyttökohteesta, mikä voi tuoda yrityksille uusia ongelmia pohdittavaksi.

Lähdeluettelo

Alshawawreh, A. R. E., Liébana-Cabanillas, F., & Blanco-Encomienda, F. J. (2024). Impact of big data analytics on telecom companies' competitive advantage. *Technology in Society*, vol. 76, pp. 102459, <https://doi.org/10.1016/j.techsoc.2024.102459>.

Carrera-Rivera, A., Ochoa, W., Larrinaga, F., & Lasa, G. (2022). How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, vol. 9, pp. 101895, <https://doi.org/10.1016/j.mex.2022.101895>.

Dakic, D., Stefanovic, D., Vuckovic, T., Zizakov, M., & Stevanov, B. (2023). Event Log Data Quality Issues and Solutions. *Mathematics*, vol. 11, no. 13, pp. 2858, <https://doi.org/10.3390/math11132858>.

Hany Fawzy, A., Wassif, K., & Moussa, H. (2023). Framework for automatic detection of anomalies in DevOps. *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 3, pp. 8–19, <https://doi.org/10.1016/j.jksuci.2023.02.010>.

He, S., Zhang, X., He, P., Xu, Y., Li, L., Kang, Y., Ma, M., Wei, Y., Dang, Y., Rajmohan, S., & Lin, Q. (2022). An empirical study of log analysis at Microsoft. *ESEC/FSE 2022 - Proceedings of the 30th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 1465–1476, <https://doi.org/10.1145/3540250.3558963>.

Huang, J., Quan, W., & Li, X. (2025). Visual anomaly detection algorithms: Development and Frontier review. *Journal of Visual Communication and Image Representation*, vol. 112, pp. 104585, <https://doi.org/10.1016/j.jvcir.2025.104585>.

Kitchenham, B. (2007). Guidelines for performing Systematic Literature Reviews in software engineering. EBSE Technical Report EBSE-2007-01. https://www.researchgate.net/publication/258968007_Kitchenham_B_Guidelines_for_performing_Systematic_Literature_Reviews_in_software_engineering_EBSE_Technical_Report_EBSE-2007-01.

Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering - A systematic literature review. In *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, <https://doi.org/10.1016/j.infsof.2008.09.009>.

Ma, J., Liu, Y., Wan, H., & Sun, G. (2023). Automatic Parsing and Utilization of System Log Features in Log Analysis: A Survey. In *Applied Sciences (Switzerland)*, vol. 13, no. 8, pp. 4930, <https://doi.org/10.3390/app13084930>.

Li, X., Yang, S., Huang, Y., Peng, J., & Zhou, M. (2022). Research on Real-time Log Data Processing And Monitoring Scheme of Printing Equipment Based on Flink Framework. *ACM International Conference Proceeding Series*, pp. 1096–1100, <https://doi.org/10.1145/3573428.3573625>.

Partovian, S., Bucaioni, A., Flammini, F., & Thornadtsson, J. (2023). Analysis of log files to enable smart-troubleshooting in Industry 4.0: a systematic mapping study in *IEEE Access*, vol. 12, pp. 147640–147658, <https://doi.org/10.1109/ACCESS.2023.3342365>.

Patil, Y., Solpaure, S. S., Umare, S., & Bhosale, S. (2025). LogInsight: A Tool for Log Analysis and Threat Detection. 2nd International Conference on Electronics, Computing, Communication and Control Technology, ICECCC, pp. 1–6, <https://doi.org/10.1109/ICECCC65144.2025.11063920>.

Shakeel, H. M., Iram, S., Al-Aqrabi, H., Alsboui, T., & Hill, R. (2022). A Comprehensive State-of-the-Art Survey on Data Visualization Tools: Research Developments, Challenges and Future Domain Specific Visualization Framework. *IEEE Access*, vol. 10, pp. 96581–96601, <https://doi.org/10.1109/ACCESS.2022.3205115>.

Shen, C. C., Yeh, C. C., & Lin, C. N. (2022). Using the perspective of business information technology technicians to explore how information technology affects business competitive advantage. *Technological Forecasting and Social Change*, vol. 184, pp. 121973, <https://doi.org/10.1016/j.techfore.2022.121973>.

Sitova, I., & Pecerska, J. (2020). Process Data Analysis Using Visual Analytics and Process Mining Techniques. 2020 61st International Scientific Conference on Information Technology and Management Science of Riga Technical University, ITMS 2020 – Proceedings, pp. 1–6, <https://doi.org/10.1109/ITMS51158.2020.9259296>.

Skopik, F., Wurzenberger, M., Hold, G., Landauer, M., & Kuhn, W. (2023). Behavior-Based Anomaly Detection in Log Data of Physical Access Control Systems. *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 4, pp. 3158–3175, <https://doi.org/10.1109/TDSC.2022.3197265>.

Song, W., Jacobsen, H. A., & Zhang, P. (2021). Self-Healing Event Logs. *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2750–2763, <https://doi.org/10.1109/TKDE.2019.2956520>.

Storm, M., & Borgman, H. P. (2020). Understanding challenges and success factors in creating a data-driven culture. Proceedings of the 53rd Hawaii International Conference on System Sciences, pp. 5399-5408, <https://doi.org/10.24251/HICSS.2020.663>.

Studiawan, H., Sohel, F., & Payne, C. (2019). A survey on forensic investigation of operating system logs. In *Digital Investigation*, Elsevier Ltd, vol. 29, pp. 1–20, <https://doi.org/10.1016/j.diin.2019.02.005>.

Swami, R., Dave, M., & Ranga, V. (2023). IQR-based approach for DDoS detection and mitigation in SDN. *Defence Technology*, vol. 25, pp. 76–87, <https://doi.org/10.1016/j.dt.2022.10.006>.

Zhang, J., Wolfram, D., & Ma, F. (2023). The impact of big data on research methods in information science. *Data and Information Management*, vol. 7, no. 2, pp. 100038, <https://doi.org/10.1016/j.dim.2023.100038>.

Zhao, X., Guo, K., Huang, M., Qiu, S., & Lu, L. (2025). ELFA-Log: Cross-System Log Anomaly Detection via Enhanced Pseudo-Labeling and Feature Alignment. *Computers*, vol. 14, no. 7, pp. 272, <https://doi.org/10.3390/computers14070272>.