

Patrik Salmensaari

TEKSTIMUOTOISEN DATAN HAASTEET LÄHES REAALIAIKAIKAISSA ETL- PUTKESSA

Kandidaatintutkielma
Informaatioteknologian ja viestinnän tiedekunta
Mikko Nurminen
Toukokuu 2025

TIIVISTELMÄ

Patrik Salmensaari: Tekstimuotoisen datan haasteet lähes reaaliaikaisessa ETL-putkessa
Kandidaatintutkielma
Tampereen yliopisto
Tietotekniikka
Toukokuu 2025

Tutkielmassa tarkasteltiin tekstimuotoisen datan käsittelyyn liittyviä teknisiä haasteita lähes reaaliaikaisessa ETL-putkessa (Extract, Transform, Load). ETL-prosessi koostuu kolmesta päävaiheesta: datan poimimisesta, muuntamisesta ja lataamisesta. Prosessin tavoitteena on poimia data lähdejärjestelmästä, muuntaa se analysoitavaan muotoon ja ladata kohdejärjestelmään. Tekstimuotoinen data, kuten sähköpostit ja asiakirjat, sisältää merkittävää liiketoiminnallista arvoa, mutta sen rakenteettomuus, monitulkintaisuus ja kontekstisidonnaisuus tekevät sen käsittelystä haastavaa.

Tutkimuksen tavoitteena oli tunnistaa keskeiset tekniset haasteet, joita lähes reaaliaikainen tekstimuotoinen ETL-prosessi kohtaa. Tutkimus toteutettiin kirjallisuuskatsauksena. Tutkielmassa havaittiin, että haasteet ilmenevät kaikissa prosessin vaiheissa. Poimintavaiheessa korostuvat lähdejärjestelmien kuormitus ja datan oikeellisuuden varmistaminen. Muunnosvaiheessa suurimmat haasteet liittyvät tekstin muuttamiseen rakenteelliseen muotoon lyhyessä ajassa sekä eri datalähteiden yhdistämiseen. Latausvaiheessa resurssikiistat kohdejärjestelmässä voivat heikentää suorituskykyä ja aiheuttaa viiveitä.

Tutkimus osoittaa myös, että tekstimuotoisen datan monimuotoisuus lisää prosessin monimutkaisuutta. Kirjoitusvirheet, monikielisyys ja kontekstin tulkinta tuottavat vaikeuksia. Lisäksi tietosuoja-asetukset (esim. GDPR) tuovat lisähaasteita erityisesti henkilötietojen käsittelyyn.

Tutkielma osoittaa, että vaikka lähes reaaliaikaisen tekstimuotoisen ETL-putken toteuttaminen on teknisesti haastavaa, se tarjoaa organisaatioille mahdollisuuden hyödyntää tekstimuotoista dataa tehokkaasti päätöksenteossa. Prosessi mahdollistaa tekstimuotoisen datan tehokkaan hyödyntämisen päätöksenteossa ja liiketoiminnassa. Suunnittelun ja toteutuksen huolellisuus ovat kuitenkin avainasemassa haasteiden minimoimisessa. Jatkotutkimuksissa voitaisiin keskittyä yksittäisiin prosessin vaiheisiin tai kehittää uusia ratkaisuja tekoälyn ja koneoppimisen avulla.

Avainsanat: ETL, Tekstimuotoinen data, ETL haasteet, tekstimuotoisen datan haasteet, Near real-time ETL challenges, Textual data.

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin Originality Check -ohjelmalla.

TEKOÄLYN KÄYTTÖ OPINNÄYTTEESSÄ

Opinnäytteessäni on käytetty tekoälysovelluksia:

- Ei
- Kyllä

Ilmoitukseni mukaan olen käyttänyt opinnäytteessäni tutkielmaproessin aikana seuraavia tekoälysovelluksia:

Google Gemini, Perplexity AI, Scopus AI

Tekoälysovellusten nimet ja versiot:

Google Gemini Flash 2.0
Perplexity AI Standard
Scopus AI

Käyttötarkoitus:

Tekoälysovelluksia on käytetty tutkielman aikana lähteiden haussa ja suomentamisessa. Alkuvaiheessa työtä tekoälyn avulla on pyritty luomaan kokonaiskuva aiheesta ja siihen liittyvistä tekijöistä. Tekoälyn avulla ei ole kirjoitettu tutkielman tekstiä.

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. ETL-DATAPUTKET	3
2.1 ETL: Extract, Transform, Load	3
2.1.1 Poimiminen	4
2.1.2 Muuntaminen	5
2.1.3 Lataaminen	5
2.2 Erilaisten ETL-putkien erot	6
3. DATAN MONIMUOTOISUUS	8
3.1 Rakenteellinen data	9
3.2 Rakenteeton data	9
3.3 Tekstimuotoinen ja ei-tekstimuotoinen data	11
3.4 Rakenteettomasta datasta rakenteelliseksi	12
3.5 Tekstimuotoisen datan haasteet	12
4. LÄHES REAALIAIKAISEN ETL:N HAASTEET	15
4.1 Poimintahaasteet	15
4.2 Muunnoshaasteet	16
4.3 Lataushaasteet	17
5. POHDINTA	19
5.1 Suunnittelu ja toteutus	19
5.2 Eettiset riskit	20
5.3 Prosessin lopputuote	21
6. YHTEENVETO	22
LÄHTEET	24

LYHENTEET JA MERKINNÄT

API	Application Programming Interface
DSA	Data Staging Area
ELT	Extract Load Transform
<i>ETL</i>	Extract Transform Load
GDPR	Euroopan unionin yleinen tietosuoja-asetus

1. JOHDANTO

Tutkielmassa tarkastellaan lähes reaaliaikaista ETL-dataputkea, joka keskittyy datan poimimiseen lähteistä, muokkaamiseen ja lataamiseen määränpäähän. ETL on lyhenne sanoista Extract, Transform ja Load. (Kimball & Ross, 2013.) Tutkielmassa kolmivaiheista prosessia kutsutaan lyhenteellä ETL, ja sen määränpäästä käytetään nimitystä kohdejärjestelmä. Lähes reaaliaikainen ETL pyrkii hakemaan dataa välittömästi, kun uutta dataa lisätään lähdejärjestelmiin, ottaen kuitenkin huomioon tekniset rajoitukset ja organisaation tarpeet (Vassiliadis & Simitsis, 2008).

Dataputkien käyttäminen on lähes välttämätöntä kaikille dataa hyödyntäville organisaatioille (Raj et al., 2020). Tämä johtuu datan määrän suuresta kasvamisesta. Vuonna 2013 Fan ja Bifet arvioivat tutkimusartikkelissaan kaiken maailmassa olevan datan kasvavan 40 % vuosittain (Fan & Bifet, 2013). Datan määrän kasvamisesta on vaikea arvioida, sillä uusia datalähteitä ilmenee jatkuvasti ja dataa tallennetaan useisiin paikkoihin. Datan arvioidaan kasvavan noin 22 % vuoden 2025 aikana (Statista, 2024). Vuonna 2013 kaiken datan määrä oli noin 9 zettatavua (10^{21}); vuonna 2025 sen odotetaan kasvavan yli 180 zettatavun (Statista, 2024). Datan määrän kasvaessa eksponentiaalisen käyrän mukaisesti, merkityksellisten oivallusten löytämisestä tulee haastavampaa ja merkittävämpää (Mishra & Misra, 2017). Tämä tekee lähes reaaliaikaisista ETL-prosesseista välttämättömiä.

Organisaatiot haluavat yhä nopeammin tietoa helposti luettavassa muodossa, jolloin päätöksenteosta saadaan tehokkaampaa. ETL-putken tarkoituksena on tallentaa data muotoon, jossa sitä on helppo analysoida ja hyödyntää päätöksenteossa. Tekstimuotoinen data on vain murto-osa kaikesta maailman datasta, mutta voi parhaimmillaan sisältää merkittäviä tietoja organisaatioiden kannalta. Tekstimuotoinen data on erityisen haastavaa sen rakenteettomuuden, monitulkintaisuuden ja kontekstisidonnaisuuden vuoksi. Näiden ominaisuuksien vuoksi se vaatii huomattavaa käsittelyä ennen analysointia. Tekstimuotoista dataa ovat esimerkiksi sähköpostit ja pdf-tiedostot. Tutkielman tavoite on selvittää, millaisia teknisiä haasteita organisaatiotasolla liittyy tekstimuotoiseen dataan lähes reaaliaikaisissa ETL-putkissa.

Tutkielman luvussa kaksi (2) käsitellään yleisellä tasolla ETL-putkien toimintaa ja miten erilaiset ETL-putket eroavat toisistaan. Luvussa kolme (3) käsitellään datan

monimuotoisuutta, joka kätkee alleen tekstimuotoisen datan. Luvussa neljä (4) syvennyttään lähes reaaliaikaisen ETL-putken haasteisiin. Luvussa viisi (5) kootaan kaikki aiemmin käsitellyt osakokonaisuudet yhteen ja pohditaan prosessin haasteiden vaikutusta organisaation näkökulmasta. Lopuksi tutkielman kokoaa yhteen kappale kuusi (6) yhteenveto.

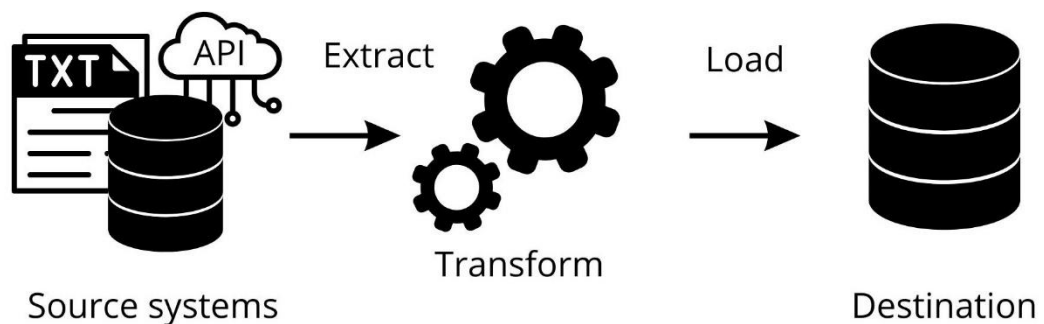
2. ETL-DATAPUTKET

Dataputki tunnetaan yleisesti englanninkielisellä nimellä data pipeline. Tässä tutkielmassa käytetään nimitystä dataputki. Dataputki viittaa prosessiin, jossa raakadata haetaan yhdestä tai useammasta lähteestä, käsitellään ja siirretään kohdejärjestelmään organisaation tarpeiden mukaisesti (Raj et al., 2020). Prosessin tarkoitus on erityisesti tuoda dataa ulkoisista lähteistä organisaation järjestelmiin, vaikka sitä voidaan hyödyntää myös organisaation sisäisen datan siirtämisessä järjestelmästä toiseen (Palmer, 2024).

ETL-putki on yksi dataputken alatyypeistä, sillä dataputki on yleisnimitys monenlaisille datan siirto- ja käsittelyprosesseille. Dataputken tavoite on saada data käyttökelpoiseen muotoon (Raj et al., 2020.) Dataputki kätkee alleen myös muita datan siirtoprosesseja lähteen ja kohteen välillä. Esimerkiksi ELT-prosessin (Extract, Load, Transform), joka toimii samalla periaatteella, mutta muuntaa datan vasta sen lataamisen jälkeen. (Palmer, 2024.)

2.1 ETL: Extract, Transform, Load

Lyhenne ETL tarkoittaa suomeksi datan poimimista (Extract), muuntamista (Transform) ja lataamista (Load) (Kimball & Ross, 2013). Prosessin tavoitteena on muuntaa raakadata organisaation tarpeita vastaavaan muotoon siten, että sitä voidaan hyödyntää esimerkiksi analytiikassa tai päätöksenteossa (Kimball & Ross, 2013). Palmer (2024) kuvaa tavoitteen olevan datan saaminen ympäristöön, jossa organisaatio voi käsitellä ja analysoida dataa haluamallaan tavalla. Kuvassa 1 on esitetty ETL-prosessin vaiheet yksinkertaistetusti.



Kuva 1: Yksinkertaistettu ETL-prosessi (luotu lähteen Palmer, 2024 avulla)

Prosessi sisältää edellä mainitut kolme päävaihetta, jotka on kuvattu kuvassa 1. ETL-prosessia käytetään laajasti organisaatioissa erityisesti silloin, kun halutaan yhdistää useista eri lähteistä tulevaa dataa yhteen järjestelmään. Se mahdollistaa datan tehokkaan käsittelyn ja varmistaa sen laadun ennen analysointia tai tallennusta kohdejärjestelmään (Kimball & Ross, 2013). Seuraavissa alaluvuissa tarkastellaan ETL-prosessin päävaihteita yksityiskohtaisemmin.

2.1.1 Poimiminen

Datan poimiminen tai kerääminen on ETL-putken ensimmäinen vaihe, jossa dataa haetaan lähdejärjestelmistä. Vaiheen tavoitteena on tunnistaa ja hakea vain sellainen data, jota voidaan organisaatiossa hyödyntää (Vassiliadis & Simitsis, 2008). Ylimääräisen datan kerääminen aiheuttaa haasteita, joita käsitellään myöhemmin.

Datan keräämisvaiheessa on tärkeä ottaa huomioon erilaiset datalähteet ja niiden ominaisuudet, näin varmistetaan lähteiden riittävät resurssit. Tämän lisäksi tarvitaan ymmärrys organisaation tavoitteista ja lähdejärjestelmistä, jotta tiedetään, mitä dataa lähteestä halutaan. (Reeve, 2013.) Lähteiden valintaan vaikuttaa datan käyttötarkoitus, päivitystiheys, odotettu datamäärä, datan laatu ja millaisessa muodossa data on (Palmer, 2024). Dataa voidaan hakea lähdejärjestelmistä kolmella eri tavalla: jatkuvasti (reaaliaikainen), ajoittain (lähes reaaliaikainen) tai kertaluonteisesti (perinteinen) (Raj et al., 2020). Valittu ETL-putki vaikuttaa siihen, missä muodossa ja kuinka usein dataa liikutetaan. Dataa voidaan siirtää jatkuvasti järjestelmästä putkeen tai kopioimalla järjestelmän data ja tallentamalla se tiedostoihin (Reeve, 2013).

Datalähteinä voidaan käyttää lähes mitä vain organisaation kannalta tarvittavaa datalähdettä. Lähteet jaetaan organisaation sisäisiin ja ulkoisiin lähteisiin. Sisäiset lähteet ovat organisaation järjestelmiä, tietokantoja tai tiedostoja. Vastaavasti ulkoiset lähteet ulkoisia tietokantoja, API-rajapintoja (application programming interface) tai muita organisaation kannalta tärkeitä datalähteitä. (Reeve, 2013; Vassiliadis & Simitsis, 2008.) Palmer (2024) painottaa, että luotettava ja oikea-aikainen datan kerääminen on ensiarvoisen tärkeää prosessin kannalta. Väärin valitut datalähteet tai heikkolaatuinen data voi aiheuttaa merkittäviä haasteita prosessin onnistumiselle.

2.1.2 Muuntaminen

Datan muuntamisvaihe on keskeinen osa datan käsittelyprosessia, jossa raakadata muunnetaan organisaation tarpeiden mukaan arvokkaaksi resurssiksi (Reeve, 2013). Muuntovaiheen voidaan ajatella olevan jopa prosessin tärkein vaihe (Kimball & Ross, 2013). Käytännössä muuntaminen tarkoittaa datan parantamista erilaisin keinoin. Muuntovaihe voi sisältää esimerkiksi datan jäsentämistä ja rikastamista (Raj et al., 2020). Yksinkertaisimmillaan muuntaminen voi olla ei-toivottujen arvojen poistamista, useiden datalähteiden yhdistämistä ja datan suodatusta, jossa vain oleellinen data otetaan käyttöön. Muuntamista voidaan tehdä useita kertoja eri vaiheissa prosessia, esimerkiksi data voidaan muuntaa heti sen saapuessa ja vielä myöhemmin uudelleen. (Palmer, 2024.)

Datan muuntamisen jälkeen data tallennetaan monissa tapauksissa väliaikaiseen varastoon, jota kutsutaan englanniksi nimellä Data Staging Area (DSA) (Raj et al., 2020). Riippuen käyttäjästä ja dataputken tyypistä DSA-alueen tarkoitukset vaihtelevat. Raj et al. (2020) mukaan tarkoitus on tehdä datalle validointitarkastus, ennen kuin se lisätään kohdejärjestelmään. Vassiliadiksen ja Simitsisin (2008) mukaan itse muunnosvaihe toteutetaan DSA-alueella, jossa dataa voidaan lisäksi käsitellä. DSA ei ole kuitenkaan välttämätön osa prosessia, sillä datan muuntamista voidaan tehdä ilman väliaikaista datavarastoa. Lähtökohtaisesti reaaliaikaisissa prosesseissa ei käytetä DSA-aluetta, sillä se hidastaa prosessin kulkua.

Palmerin (2024) mukaan koko muuntoprosessin tavoite on muuntaa data muotoon, jossa se on johdonmukainen ja tarkka. Datan käsittelyyn liittyy paljon mahdollisuuksia ja sitä voidaan soveltaa käyttötarkoituksen ja tavoitteiden mukaisesti. Datan käsittelyllä varmistetaan, ettei kohdejärjestelmään päädy virheellistä dataa, joka voisi hankaloittaa analysointia.

2.1.3 Lataaminen

ETL-prosessin latausvaiheessa muunnettu data ladataan kohdejärjestelmään (Kimball & Ross, 2013). Latausprosessi voi tapahtua usealla eri tavalla. Prosessiin vaikuttaa organisaation tarpeet, ETL-putken tyyppi ja kohdejärjestelmä. Data tallennetaan usein tietovarastoon tai tietokantaan, joka sisältää jäsenneiltyä ja analysoitavaa dataa. Vaihtoehtoisesti se voidaan viedä tietoaltaaseen raakamuodossa tai suoraan analytiikkajärjestelmiin. (Raj et al., 2020.) Datan tallennuspaikka riippuu siitä, miten ja milloin dataa tullaan hyödyntämään (Palmer, 2024).

Dataa voidaan ladata kohdejärjestelmiin erilaisin tavoin. Esimerkiksi reaaliaikaista analytiikkaa vaativissa järjestelmissä dataa voidaan ladata jatkuvasti pieninä erinä, kun taas harvemmin päivitettävissä järjestelmissä voidaan suorittaa suurempia latauksia kerralla tai tietyin aikaväleihin (Palmer, 2024). Valitun ETL-putken tyyppi määrittää datan latausnopeuden (Reeve, 2013). Vassiliadiksen ja Simitsisin (2008) mukaan ETL-prosessi on valmis, kun data on ladattu kohdejärjestelmään.

2.2 Erilaisten ETL-putkien erot

Perinteisessä ETL-prosessissa data kulkee putken läpi lähtökohtaisesti kerran päivässä (Boulahia et al., 2021). Tällöin dataa haetaan lähdejärjestelmästä kertaluonteisesti. Jatkuvamman ja suuremman datamäärän vuoksi ETL-putkilta on vaadittu nopeampaa ja tehokkaampaa suorituskykyä, korkeampaa saatavuutta, pienempää viivettä ja skaalautuvuutta (Boulahia et al., 2021). Koska perinteinen ETL prosessi ei tähän pystynyt, luotiin uusia ratkaisuja, jotka toimivat samalla periaatteella, mutta tehokkaammin ja nopeammin.

Erilaisia ETL-prosesseja on erilaisiin tarkoituksiin, joista suurien datamäärien prosesseja ovat reaaliaikainen (real-time) ja lähes reaaliaikainen (near real-time) ETL-putki. Suurimpana erona erilaisten ETL-putkien välillä on datan hakemisen toistuvuus. Erilaisilla organisaatioilla datan päivitysnopeuteen liittyvät tarpeet vaihtelevat organisaation liiketoiminnallisten vaatimusten mukaan, jonka vuoksi erilaisia ETL-putkia käytetään. (Vassiliadis & Simitsis, 2008.)

Reaaliaikaisessa ETL-prosessissa uusi data siirtyy vaiheesta toiseen välittömästi saapuessaan, kun taas lähes reaaliaikaisessa prosessissa dataa haetaan säännöllisin väliajoin useita kertoja päivässä. Säännöllisyys voi tarkoittaa esimerkiksi minuuttien välein tapahtuvaa datan hakua (Vassiliadis & Simitsis, 2008). Lähes reaaliaikaisella ETL-putkella on korkea datan saatavuus, alhainen viive ja horisontaalinen skaalautuvuus (Sabtu et al., 2017). Jatkuvan datan siirron vuoksi reaaliaikaista prosessia on jouduttu automatisoimaan hyvin pitkälle. Oleellista reaaliaikaisella putkella on virheiden havaitseminen ja niiden käsittely. Näin taataan datan laatu ja luotettavuus, sekä prosessin keskeytymätön luonne. (Vijayalakshmi & Minu, 2021.) Tekstissä käsiteltyjen erilaisten ETL putkien eroja ja tyypillisiä käyttötapauksia on kuvattu taulukossa 1.

Taulukko 1. ETL putkien erot tyypeittäin (Boulahia et al., 2021; Palmer, 2024; Sabtu et al., 2017; Vassiliadis & Simitsis, 2008; Vijayalakshmi & Minu, 2021)

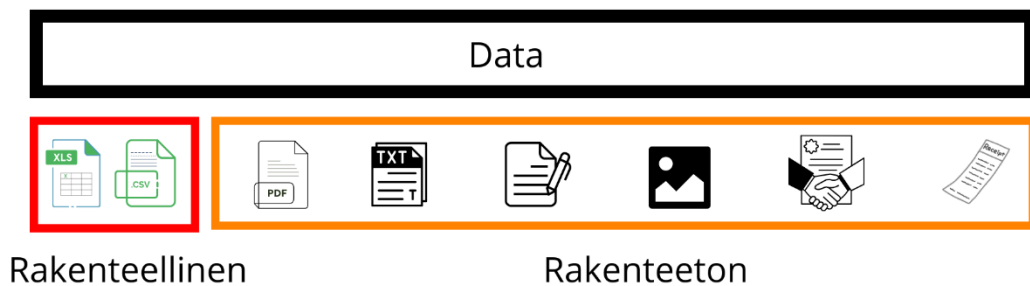
Putken tyyppi	Datan siirtonopeus	Käyttötapaus
Perinteinen ETL (Batch)	Tietyin väliajoin. Usein kerran päivässä.	Kun datan ei tarvitse olla reaaliaikaista, esim. raportointi ja analyysit, jotka eivät vaadi välitöntä päivittämistä.
Lähes reaaliaikainen ETL (Near Real-Time ETL)	Aina kun lähdejärjestelmässä uutta dataa. Esimerkiksi muutaman tunnin välein. Parhaimmillaan jopa minuutin välein.	Kun datan tarvitsee päivittyä useammin kuin kerran päivässä, mutta ei välttämättä reaaliaikaisesti. Esim. varastosaldojen päivitys tai asiakastietojen muutokset.
Reaaliaikainen ETL (Real-Time ETL)	Jatkuvasti, välittömästi	Kun datan tarvitsee olla saatavilla heti muutosten tapahduttua. Esim. reaaliaikainen analytiikka tai petosten torjunta

Erilaiset käyttötapaudet taulukossa 1 kuvaavat erilaisia ETL-prosesseja. Taulukossa on esitetty kuitenkin vain muutamia esimerkkejä ja todellisia käyttötappauksia on huomattavasti enemmän. ETL-prosessin valinta riippuu organisaation tavoitteista, datan luonteesta ja resursseista. Nämä kaikki tekijät tulee ottaa huomioon putken tyyppiä valitessa. ETL-teknologioita kehitetään jatkuvasti datan määrän voimakkaan kasvun vuoksi. Näin ollen myös uusia työkaluja ja lähestymistapoja saattaa ilmetä tulevaisuudessa. ETL-prosessit tulevat todennäköisesti tulevaisuudessa kehittymään yhä reaaliaikaisemmiksi ja automatisoidummiksi.

3. DATAN MONIMUOTOISUUS

Datan monimuotoisuus vaikuttaa merkittävästi siihen, miten se voidaan poimia, muuntaa ja ladata lähes reaaliaikaisessa ETL-prosessissa. Yleisesti data voidaan jakaa kahteen kategoriaan sen tuottajan perusteella, jotka ovat ihmisten tai koneiden tuottama data (Raj et al., 2020). Dataa syntyy esimerkiksi sosiaalisesta mediasta, sensoreista ja tieteellisistä lähteistä (Mishra & Misra, 2017). Ihmisten tuottamaa dataa ovat esimerkiksi sähköpostit, tiedostot, sopimukset ja kuitit, kun taas sensoreiden tuottama data on koneiden tuottamaa dataa (Inmon et al., 2019).

Data voidaan jakaa myös kahteen kategoriaan rakenteen perusteella, jotka ovat rakenteellinen (englanniksi Structured) ja rakenteeton (englanniksi Unstructured). Monissa tilanteissa rakenteetonta dataa on enemmän, joka johtuu datan muodosta. (Inmon et al., 2019) Kuvassa 2 esitetään havainnollisesti, kuinka data jakautuu rakenteelliseen ja rakenteettomaan muotoon. Lisäksi kuvassa havainnollistetaan erilaisten datalähteiden jakautumista näihin kategorioihin.



Kuva 2: Datan jakautuminen rakenteelliseen ja rakenteettomaan dataan (luotu lähteen Inmon et al., 2019 avulla)

Kuvassa 2 käsitellyt kategorioita ja niiden ominaispiirteitä on avattu tarkemmin seuraavissa kappaleissa. Kuva korostaa erityisesti rakenteettoman datan suurta määrää, jonka syitä käsitellään seuraavassa kappaleessa.

3.1 Rakenteellinen data

Rakenteellinen data on selkeästi organisoitua, helposti analysoitavaa ja usein toistuvaa, kuten myyntitiedot (Inmon et al., 2019; Mishra & Misra, 2017). Esimerkiksi joka kerta kun päivittäistavarakauppa tekee myynnin, siitä tallennetaan seuraavaa dataa: tuote, myyntisumma, vero, päivämäärä ja sijainti (Inmon et al., 2019). Esimerkin mukaisessa tilanteessa data on hyvin selkeässä muodossa. Tätä tukee Mishran ja Misran (2017) väite, ”Strukturoitu data on selkeästi järjestettyä dataa, jossa kaikki on hyvin tunnistettavissa”.

Esimerkin mukaisessa tilanteessa kaikki data kerätään samalla tavalla. Tällöin dataa on helppo käsitellä ohjelmistoilla ja tallentaa organisaation kannalta oleellisiin järjestelmiin. Rakenteellinen data on kuitenkin vain pieni osa organisaation dataa, sillä suurin osa datasta on rakenteetonta dataa (Inmon et al., 2019). Inmon et al. (2019) esittämän arvion mukaan 2–20 prosenttia datasta on rakenteellista. Mishra ja Misra (2017) vahvistavat tämän toteamalla, että jopa 80 % yritysten kannalta olennaisesta datasta on peräisin rakenteettomista lähteistä. Arvion tarkkuuteen vaikuttaa organisaation toimiala ja se millaista dataa käytetään laskelmaa tehdessä (Inmon et al., 2019). Arvion tarkkuudesta huolimatta rakenteettoman datan määrä on merkittävä.

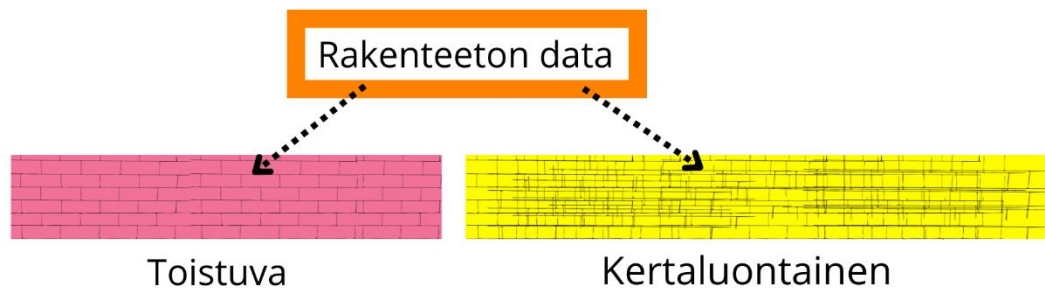
3.2 Rakenteeton data

Erilaiset tieteenalat määrittelevät rakenteellisen ja rakenteettoman datan välisen eron eri tavoin, eikä rajanveto ole aina yksiselitteinen. Tässä tutkielmassa käytetään Inmon et al. (2019) esittämää tietoteknistä näkökulmaa, jonka mukaan teksti katsotaan rakenteettomaksi, jos se ei ole koneluettavassa muodossa. Tämä määritelmä on valittu siksi, että se korostaa juuri teknistä näkökulmaa ja tukee tutkielman tavoitetta analysoida tekstimuotoisen datan muuntamista rakenteelliseen muotoon. Inmon et al. (2019) mukaan rakenteettoman datan muuntamisessa rakenteelliseksi keskeistä on merkityksen ja asiayhteyden tunnistaminen, mikä tekee prosessista sekä haastavan että organisaatioille arvokkaan.

Rakenteeton data ei sisällä selkeää ennalta määriteltyä rakennetta ja on usein järjestämättömässä muodossa (Mishra & Misra, 2017). Datan rakenne vaikuttaa sen tallentamismenetelmiin, jonka vuoksi datan rakenne on merkittävässä roolissa ETL-putkissa (Palmer, 2024). Rakenteetonta dataa ei voida tallentaa relaatiotietokantoihin, joissa data tallennetaan tauluihin, joilla on yhteyksiä toisiinsa. Sen sijaan rakenteeton data tulee tallentaa tietoaaltaaseen tai ei-relaatiotietokantoihin, joissa data voidaan

tallentaa erilaisin menetelmin, kuten dokumentein, avain-arvoparein tai graafein. (MongoDB, 2025.)

Rakenteettomaksi dataksi määritellään muun muassa sähköpostit ja sopimukset sekä video- ja audiosisältö (Inmon et al., 2019; Mishra ja Misra, 2017). Rakenteeton data voidaan jakaa vielä kahteen alakategoriaan sen sisällön tai toistuvuuden osalta. Toistuvuuden osalta dataa on kahdenlaista: toistuvaa (englanniksi Repetitive) ja kertaluonteista (englanniksi Non repetitive) (Brestoff & Inmon, 2015; Inmon et al., 2019). Esimerkiksi sähköpostit ovat kertaluonteisia, sillä ne ovat tyylillisesti yksilöllisiä ja ainutkertaisia. Asiakaspalautteet taas ovat toistuvia, koska niiden sisältö on tyypillisesti samankaltaista ja toistuvaa. Kuvassa 3 esitetään havainnollisesti toistuvan ja kertaluonteisen datan eroa.

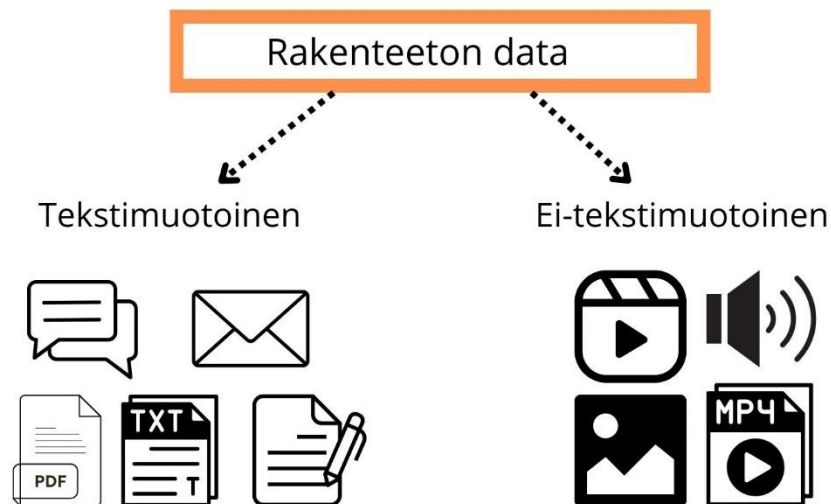


Kuva 3: Rakenteettoman datan jakautuminen toistuvaan ja kertaluonteiseen dataan (luotu lähteen Brestoff & Inmon, 2015 avulla)

Rakenteettoman datan hyötyinä pidetään sen helppoa keräämistä ja tallentamista, sillä data voidaan tallentaa ilman rakennetta tietoaaltaaseen ja käsitellä myöhemmin (Palmer, 2024). Haasteena on kuitenkin datan läpikäymiseen kuluvat resurssit ja aika. Yleisesti kertaluonteiset datamassat sisältävät liiketoiminnan kannalta arvokkaampaa tietoa kuin toistuvat datamassat, jonka vuoksi organisaatiot haluavat tallentaa kyseistä dataa (Inmon et al., 2019). Suurin ero organisaation näkökulmasta on liiketoiminnallinen arvo (Brestoff & Inmon, 2015; Inmon et al., 2019). Toistuvassa datassa vain osalla on liiketoiminnallista arvoa, kertaluonteisessa datassa lähes jokainen tietue on arvokas (Brestoff & Inmon, 2015). Esimerkiksi asiakaspalautelomakkeen avoimet kentät keräävät usein samankaltaisia vastauksia. Tällainen data on helposti analysoitavissa, mutta jokainen uusi palaute ei edistä organisaation toimintaa. Kertaluonteisen datan kohdalla esimerkiksi tarkkakuvaus ongelmatilanteesta voi auttaa organisaatiota korjaamaan tilanteen. Vaikka data on määrällisesti pienempää, se voi sisältää tarkan ja konkreettisen keinon kehittää organisaation toimintaa.

3.3 Tekstimuotoinen ja ei-tekstimuotoinen data

Rakenteeton data voidaan jakaa tekstimuotoiseen ja ei-tekstimuotoiseen dataan. Tekstimuotoista dataa ovat esimerkiksi sähköpostit ja ei-tekstimuotoista dataa valokuvat (Mishra & Misra, 2017). Tekstimuotoinen data sisältää lähtökohtaisesti vain tekstiä (Inmon & Nesavich, 2008). Kuvassa 4 on pyritty havainnollistamaan rakenteettoman datan jakautumista yleisimpiin tekstimuotoisiin ja ei-tekstimuotoisiin datalähteisiin.



Kuva 4: Rakenteettoman datan jakautuminen tekstimuotoiseen ja ei-tekstimuotoiseen dataan (luotu lähteiden Inmon et al., 2019; Inmon & Nesavich, 2008; Mishra & Misra, 2017 avulla)

Tutkielmassa keskitytään erityisesti ETL-putkissa kulkevaan tekstimuotoiseen dataan ja sen haasteisiin. Datan haasteet ovat merkittävässä roolissa datan liiketoiminnallisen arvon maksimoinnissa. Tekstimuotoisen datan rooli organisaatioissa on merkittävä, sillä tekstimuotoinen data sisältää merkittävää arvoa organisaation näkökulmasta (Inmon et al., 2019). Monissa tilanteissa liiketoiminnallisen arvon kasvattaminen tarkoittaa rakenteettoman datan muokkaamista rakenteelliseen muotoon. Tämä ei kuitenkaan ole ainoa mahdollisuus, vaan ratkaisuun vaikuttaa liiketoiminnan tavoitteet.

3.4 Rakenteettomasta datasta rakenteelliseksi

Vaikka rakenteeton data sisältää merkittävää tietoa, jopa 98 % organisaatioiden päätöksistä perustuu rakenteelliseen dataan. Tämä johtuu rakenteellisen datan helpommasta käsittelystä, säilytyksestä ja analysoinnista, erityisesti suurten datamäärien kohdalla. (Inmon et al., 2019.) Tämän vuoksi rakenteettoman datan muuntaminen rakenteelliseen muotoon on olennaista, jotta organisaatiot voivat hyödyntää kaikkea relevanttia tietoa päätöksenteossaan. On kuitenkin tärkeää huomioida, että kaikki rakenteeton data ei ole organisaation kannalta oleellista. Tekstimuotoinen data saattaa sisältää paljon epäoleellisia asioita, jotka tulisi suodattaa pois. Ylimääräisen datan säilyttäminen tuottaa kustannuksia, ja voi johtaa päätöksentekoa harhaan (Inmon & Nesavich, 2008).

Rakenteettoman datan muuntaminen rakenteelliseen muotoon on huomattavasti helpompaa toistuvalla datalla, koska toistuvan datan luonne on ennakoitavissa (Inmon et al., 2019). Toistuvan rakenteettoman datan analysointia voidaan automatisoida ennalta määritettyjen kaavojen ja koneoppimismenetelmien avulla. Esimerkiksi tekstistä voidaan poistaa ylimääräiset merkit säännöllisillä lausekkeilla tai luokitella asiakaspalautteiden tunne koneoppimismalleilla. Kertaluonteisella datalla tällainen automaatio on haastavampaa johtuen datan vaihtelevuudesta ja epäselvästä muodosta.

Oleellista rakenteettoman datan muuntamisessa rakenteelliseksi on merkityksen ja asiayhteyden löytäminen (Inmon et al., 2019). Mikäli tekstile ei tiedetä kontekstia, on sitä todella haastavaa käyttää analysoinnissa tai päätöksenteossa. Kontekstin löytämisen ajatellaankin olevan muuntamisprosessin merkittävin, mutta myös haastavin vaihe (Inmon et al., 2019).

3.5 Tekstimuotoisen datan haasteet

Tekstimuotoisen datan käsittelyyn liittyy useita erilaisia haasteita, joiden korjaamatta jättäminen voi aiheuttaa merkittävää haittaa organisaatiolle. Tekstiä syntyy erilaisissa ympäristöissä eri henkilöiden toimesta, mikä johtaa merkittäviin vaihteluihin tekstin tyylissä, rakenteessa ja muodossa. Tekstin kappaleiden pituudet ja lauserakenteet voivat vaihdella, oikeinkirjoitus ei ole aina taattua sekä teksti voi olla monitulkinnaista ja monimutkaista (Brestoff & Inmon, 2015). Esimerkiksi sähköpostin ja virallisen asiakirjan tyyli sekä rakenne ovat täysin erilaisia, joka tekee tekstin käsittelystä haastavampaa.

Kansainvälisessä toimintaympäristössä organisaatiot kohtaavat usein monikielistä dataa, joka sisältää erilaisia murteita ja kulttuurisia konteksteja. Tekstin monimuotoisuus, kuten slangisanat, virallinen kieli ja lyhenteiden käyttö tekee kääntämisestä haastavaa.

Tämä näkyy epäluottamuksena automaattikäännöksiä kohtaan, koska nykyiset järjestelmät eivät pysty huomioimaan sanojen kaksoismerkityksiä tai lyhenteitä. (Brestoff & Inmon, 2015). Koneen tulkitessa tekstiä myös erilaiset tekstin sisäiset haasteet voivat vääristää tekstiä ja sen kontekstia (Inmon et al., 2019). Esimerkiksi ”palo auto” ja ”paloauto” tarkoittavat eri asiaa.

Lisäksi eri kielet voivat sisältää sanojen kaksoismerkityksiä tai synonyymeja. On myös mahdollista, että eri ihmiset voivat kutsua samaa asiaa eri nimellä. Kaikki edellä mainitut haasteet voivat aiheuttaa tekstin kontekstin vääristymistä. (Brestoff & Inmon, 2015.) Tekstimuotoista dataa käsitellessä teksti on kuitenkin yhdenmukaistettava, joka voi tuottaa haasteita tekstin analysoimisessa (Inmon & Nesavich, 2008). Mikäli synonyymi korvataan yleisemmällä vastaavalla sanalla, korvaamisen jälkeen alkuperäistä tekstiä ei enää säilytetä (Inmon & Nesavich, 2008). Tämä saattaa aiheuttaa lisähaasteita myöhemmin, sillä synonyymien korvaaminen luo epä johdonmukaisuutta, mikä lisää virhetulkintojen mahdollisuutta.

Tekstin sisäisten haasteiden lisäksi kaksi merkittävintä haastetta ovat tekstidatan määrä ja kontekstin liittäminen tekstiin (Brestoff & Inmon, 2015). Tekstin kontekstin merkitystä pyritään havainnoimaan esimerkillä. Lause, ”pankki oli täynnä” voi tarkoittaa kahta eri asiaa. Pankki oli täynnä rahaa, tai pankki oli täynnä ihmisiä. Mikäli tekstin konteksti on väärä tai sitä ei ole tiedossa, on analyysia hyvin vaikea tehdä. Usein tekstin kontekstin ymmärtämistä pidetään itsestäänselvyytenä, vaikka se on yksi haastavimmista ja välttämättömmistä prosessin vaiheista (Brestoff & Inmon, 2015). Kontekstin ymmärtämisen jälkeen todellinen haaste piilee tekstin ulkopuolisessa kontekstissa (Brestoff & Inmon, 2015). Tekstin ulkopuolinen konteksti tarkoittaa kaikkia niitä tekijöitä, jotka eivät ole suoraan läsnä tekstissä, mutta jotka vaikuttavat tekstin merkitykseen ja tulkintaan. Tällaisia tekijöitä ovat esimerkiksi tekstin kirjoittaja ja kohderyhmä. Tämän kontekstin ymmärtäminen on olennaista, jotta voidaan saada kokonaisvaltainen kuva tekstin sisällöstä ja sen tavoitteista. (Brestoff & Inmon, 2015.)

Uudempana haasteena tekstin käsittelyssä voidaan pitää tietoturvaa ja henkilötietoja. Mikäli tekstidata sisältää henkilötietoja, kuten nimiä tai osoitteita, tulee ottaa huomioon yleinen tietosuojasäädös GDPR (Euroopan unioni, 2016). Monissa tapauksissa esimerkiksi sähköpostit sisältävät henkilötietoja. Kyseinen Euroopan unionin säädös aiheuttaa lisähaasteita, sillä säätöön mukaan henkilötietojen käsittelyyn tarvitaan henkilön lupa. Tämän lisäksi henkilöllä on oikeus saada tietää, millaista tietoa hänestä säilytetään ja miten tietoa käsitellään. Kaiken lisäksi henkilö voi pyytää tietojen poistamista milloin tahansa. (Euroopan unioni, 2016.) Automatisoidun tekstinkäsittelyprosessin voi olla vaikeaa tunnistaa henkilötietoja tekstin seasta, mikä

tulee ottaa huomioon ETL-putkia suunnitellessa. Mikäli teksti sisältää henkilötietoja, tulee sen käsittelyssä olla varovainen.

GDPR-asetusten lisäksi dataa käsitellessä tulee toimia huolellisesti ja varovaisesti. Data tulee säilyttää turvallisessa paikassa, jotta ulkopuolisilla ei ole mahdollisuutta päästä käsiksi dataan. Inmon ja Nesavich (2008) korostavat datan turvallisen käytön ja säilytyksen merkitystä sekä sitä, että kaikki data ei saa olla avoimesti käytettävissä organisaation henkilöillä. Pääsyä tulee rajoittaa ainoastaan niille henkilöille, joiden tehtäviin data kuuluu, erityisesti henkilötietoja käsiteltävissä organisaatioissa.

4. LÄHES REAALIAIKAISEN ETL:N HAASTEET

Maailmassa ollaan siirtymässä yhä enemmän kohti jatkuvaa datan keräämistä. Siirtyminen perinteisistä ETL-prosesseista reaaliaikaisempiin on tuonut merkittävää edistystä, sillä reaaliaikaiset prosessit voivat käsitellä suurempia datamääriä. Datan kasvaessa myös uusia haasteita ilmenee. Jatkuvasti tuotettua dataa ovat esimerkiksi liikennekamerat, jotka ottavat uusia kuvia jopa sekuntien välein. Tämän tyyppiset datalähteet sisältävät suuren määrän dataa, joka päivitetään uudella datalla, usein muutamien sekuntien välein. Jatkuva datan muuttuminen ja suuri datamäärä aiheuttavat ongelmia, esimerkiksi datan tallentamiseen ja hyödyntämiseen liittyen (Vassiliadis & Simitsis, 2008; Brestoff & Inmon, 2015).

Suuren datamäärän lisäksi väärin valittu ETL-putken tyyppi voi aiheuttaa haasteita. Lähes jatkuvasti päivittyvää dataa tulisi käsitellä reaaliaikaisella ETL-putkella. Lähes reaaliaikaisen ETL-putken käyttäminen prosessissa voi aiheuttaa suorituskyvyn heikkenemistä (Vassiliadis & Simitsis, 2008). Lähes reaaliaikaisella ETL-putkella ei siis voida käsitellä lähdejärjestelmiä, jotka tuottavat dataa jatkuvasti. Liian suuri datamäärä tai väärin valittu ETL-putki ovat yleisiä haasteita ETL-prosessille. Tämän lisäksi jokainen ETL-prosessin vaihe sisältää omat haasteensa, joita käsitellään seuraavaksi. Tutkielmassa keskitytään teknisiin haasteisiin, jonka vuoksi kustannuksiin liittyvät haasteet on jätetty tutkielman ulkopuolelle.

4.1 Poimintahaasteet

Organisaation omia lähdejärjestelmiä voivat olla esimerkiksi tuotantojärjestelmä, joka sisältää taloushallinnon ja myynnin dataa, sekä asiakastietojärjestelmä, joka tallentaa asiakassuhteiden hallintaan liittyvää dataa. Näiden järjestelmien päätarkoitus ei ole tuottaa ja kerätä dataa, vaan helpottaa organisaation toimintaa (Vassiliadis & Simitsis, 2008). Tästä huolimatta mikä tahansa ongelma näissä järjestelmissä voisi vaikuttaa organisaation liiketoimintaan ja lähes reaaliaikaiseen ETL-prosessiin. Monissa tilanteissa organisaation omat datalähteet ovat päädatalähteitä, jonka vuoksi niiden jatkuva-aikainen toiminta on erittäin oleellinen osa ETL-prosessin datan poimintavaihetta. (Vassiliadis & Simitsis, 2008.)

Organisaation lähdejärjestelmiin liittyy myös kuormitushaasteita. Mikäli datan lähdejärjestelmää kuormitetaan datan hakemisella liikaa, voi se hidastaa organisaation kannalta oleellisten järjestelmien toimintaa. Tämä voi vaarantaa niiden ensisijaisen

tehtävän. (Sabtu et al., 2017; Vassiliadis & Simitsis, 2008.) Datan kerääminen ja järjestelmien yhtäaikainen käyttö ovat siis merkittävä riski organisaatiotasolla. Tämän vuoksi datan poimiminen on toteutettava tehokkaasta, ylikuormittamatta lähdeä (Vassiliadis & Simitsis, 2008).

Lähdejärjestelmien dataa käsiteltäessä on välttämätöntä varmistaa datan oikeellisuus. Prosessin tulisi poimia vain sellainen data, joka on sitoutunut johonkin, esimerkiksi kontekstiin. Erityisesti ulkoisten datalähteiden kohdalla datan oikeellisuuteen tulee kiinnittää huomiota. Ulkoiset datalähteet voivat sisältää virheellistä tai puutteellista dataa, mikä voi johtaa harhaan ja vaikuttaa analyysien luotettavuuteen. Kimball'in ja Ross'in (2013) mukaan datan laadun ja yhdenmukaisuuden varmistaminen poimintaprosessin aikana on ratkaisevan tärkeää, mutta haastavaa lähes reaaliaikaisissa skenaarioissa. Kyseisen oikeellisuuden ja kontekstuaalisuuden laiminlyönti voi johtaa epätarkkuuksiin ja epäyhdenmukaisuuksiin tietovarastossa (Kimball & Ross, 2013).

Ulkoisia datalähteitä käytettäessä on varmistettava niiden yhteensopivuus. Eri lähdejärjestelmien ominaisuudet voivat edellyttää erilaisia toimintoja datan hakemisessa. Vaikka datan yhdistäminen tapahtuu muunnosvaiheessa, ei dataa voida poimia, mikäli järjestelmät eivät ole yhteensopivia. Jos järjestelmistä puuttuu yhdistämiseen vaadittavia ominaisuuksia, kuten yhdistävä data sarake, voi yhdistämisessä ilmetä ongelmia, jotka tulee ratkaista manuaalisella työllä. Useat datalähteet voivat sisältää samaa dataa, kuten asiakas numero, jolloin tarvittava data tulee poimia vain kertaalleen. (Vassiliadis & Simitsis, 2008.)

4.2 Muunnoshaasteet

Datan muunnosvaiheen oleellisin tehtävä on yhdistää tai integroida eri lähdejärjestelmien data yhdeksi dataksi. Erityisesti tekstimuotoista dataa käsitellessä datan käsittely korostuu. Yhdistämisprosessi toteutetaan usein liitosoperaatioilla, joissa yhdistetään kahden tai useamman datalähteen dataa. Operaation haasteena on sen raskas suorittaminen, joten suorituskyvyn optimointi on välttämätöntä. Tätä varten käytetään suorituskykyä tehostavia tekniikoita, kuten indeksointia ja datan jakamista osiin. (Bornea et al., 2011.)

Muunnosprosessin sykli voi olla hyvinkin tiheä, jolloin liitosten nopeus ja tehokkuus korostuvat. Liitosten suorituskyky määrittää pitkälti koko ETL-prosessin läpimenoajan ja vaikuttaa suoraan kohdejärjestelmän päivitystiheyteen. Käsiteltävän datan määrä vaihtelee ETL-ajojen välillä, mikä vaikeuttaa tehokkaan liitostoiminnon suunnittelua (Bornea et al., 2011). Liitosalgoritmien on kyettävä mukautumaan datamäärien

muutoksiin, kuten saapumisnopeuden ja datan jakautumisen vaihteluihin (Bornea et al., 2011). Liitosten on tapahduttava määrättyissä aikarajoissa, jolloin liitosalgoritmin on priorisoitava käsittelyä sen mukaisesti (Bornea et al., 2011). Mikäli liitosta ei ehditä toteuttaa tietyssä ajassa, täytyy se hylätä ja suorittaa myöhemmin uudelleen.

Suurten datamassojen muuntaminen on haastavaa, minkä vuoksi muunnostehtävät on mahdollista jakaa pienempiin osiin. Poimittu data voidaan siirtää väliaikaiseen tallennusalueeseen (DSA), jossa se muunnetaan ja puhdistetaan ennen lopullista tallennusta kohdejärjestelmään. Tämä toimintatapa helpottaa virheiden tunnistamista ja korjaamista, sillä ongelmatilanteissa riittää korjata vain kyseisen osatehtävän käsittelemä data. (Vassiliadis & Simitsis, 2008.) Lähes reaaliaikaisessa prosessissa väliaikaistallennusaluetta käytetään kuitenkin harvemmin. Prosessilla ei nimittäin ole aikaa ylimääräisille vaiheille, sillä data halutaan nopeasti kohdejärjestelmään.

Datan lähes jatkuva muuttuminen kohdejärjestelmässä monimutkaistaa datan muunnosvaihetta. Toisin kuin perinteisissä ETL-prosesseissa, joissa liitokset kohdistuvat staattisiin tauluihin, lähes reaaliaikaisessa ympäristössä jatkuva päivitysvirta on yhdistettävä olemassa olevaan, levyllä tallennettuun dataan. Tämä vaatii uudenlaisia lähestymistapoja. (Bornea et al., 2011.) Lisäksi liitosten muistinhallinta on kriittistä. Algoritmin on optimoitava muistin käyttö lähes reaaliaikaisesti saapuvan datan puskurointiin ja tallennettujen yhteyksien väliaikaiseen varastointiin, mahdollisesti datavirran ominaisuuksien mukaan dynaamisesti (Bornea et al., 2011).

Monissa datalähteissä on käytössä niin kutsuttua metadataa, joka on viitetietoja esimerkiksi mittausdatasta tai myyntiluvuista. Kyseistä dataa säilytetään, koska sen avulla voidaan luoda konteksti datalle, kuten asettaa päivämäärä. Ongelmana on datan vähäinen muuttuminen, suhteessa oleellisen datan muuttumiseen. (Machado et al., 2019; Sabtu et al., 2017.) Metadatan jatkuva hakeminen voi johtaa järjestelmän lisäkuormittumiseen (Sabtu et al., 2017). Mikäli metadata haetaan jokaisella kerralla, tekee tämä koko muunnosprosessista tehottoman. Tällöin metadatataulut kasvavat ajan myötä ja voivat lopulta olla suurempia kuin varsinaiset lähdetiedot. (Machado et al., 2019.)

4.3 Lataushaasteet

Datan latausvaihe kohtaa haasteita, jotka liittyvät tietovaraston suorituskykyyn ja datan monimuotoisuuteen (Sabtu et al., 2017). Monissa tilanteissa samalla kun dataa ladataan kohdejärjestelmään, käytetään kohdejärjestelmää sen pääsääntöiseen tarkoitukseen. Kyseinen järjestelmän käyttäminen samanaikaisesti useaan toimintaan aiheuttaa

kilpailun järjestelmän muistista ja resursseista, joka voi joissakin tapauksissa johtaa mahdollisiin lukitusongelmiin (Vassiliadis & Simitsis, 2008). Sabtu et al. (2017) mukaan tilanne heikentää suorituskyykyä, hidastaa prosessia ja lisää datan epäyhdenmukaisuutta. Tilannetta voidaan helpottaa kohdejärjestelmän suunnittelulla ja laitteiston parantamisella (Vassiliadis & Simitsis, 2008). Esimerkiksi tietojen lataamista kohdejärjestelmään voidaan ajoittaa hetkille, jolloin kohdejärjestelmän käyttöaste on matalampi (Vassiliadis & Simitsis, 2008).

Riippuen latausprosessin suunnittelusta dataa voidaan tallentaa joko joukkolatauksella, jossa suuri määrä dataa lisätään kerralla, tai rivisarjan lisäämisellä, jossa data lisätään rivi kerrallaan. Joukkolataus on suorituskyyvyltään parempi, mutta haasteena on erottaa uusi data jo ladatusta datasta, jolloin kohdejärjestelmään saattaa päätyä jo siellä olevaa dataa. (Vassiliadis & Simitsis, 2008.)

Lähes reaaliaikaisen ETL-prosessin suurimmat haasteet liittyvät lähdejärjestelmien kuormituksen hallintaan, datan oikeellisuuden varmistamiseen, muunnosvaiheen tehokkuuteen ja latausvaiheen resurssien kilpailuun. Näiden haasteiden ratkaiseminen vaatii huolellista suunnittelua ja optimointia, jotta prosessi voi toimia lähes reaaliaikaisessa ympäristössä ilman merkittäviä viiveitä tai suorituskyykyongelmia.

5. POHDINTA

Lähes reaaliaikaisen ETL-prosessin hyödyt ovat kiistattomat prosessin onnistuessa. Prosessi sisältää kuitenkin merkittävän määrän erilaisia haasteita, kuten aiemmin käsitellyt tekstimuotoisen datan haasteet sekä poiminta-, muunnos- ja latausvaiheen haasteet. On tunnistettavissa, että prosessin rakentaminen ja ylläpito vaativat huomattavia organisaatiotason panostuksia, jonka vuoksi prosessin käyttöönottoa tulee pohtia monilta eri näkökulmilta. Seuraavissa luvuissa prosessin haasteita on pohdittu suunnittelun ja toteutuksen, eettisyyden ja prosessin lopputuotteen näkökulmasta.

5.1 Suunnittelu ja toteutus

Prosessin haasteiden ilmenemiseen voidaan vaikuttaa ETL-prosessia suunniteltaessa ja toteutettaessa. Suurin osa ongelmista on ennaltaehkäistävässä tai niihin voidaan varautua mukauttamalla prosessia. Laadukkaasta suunnittelusta ja toteutuksesta huolimatta haasteet eivät välttämättä ilmene jokaisessa prosessissa tai jokaisella suorituskerralla. Esimerkiksi datan määrän vaihtelut voivat aiheuttaa kuormitushaasteita vain silloin, kun käsiteltävä datamäärä on poikkeuksellisen suuri.

Jotta prosessi olisi mahdollisimman toimiva, suunnittelussa tarvitaan syvällistä ymmärrystä yleisistä haasteista. Näistä huolimatta uusia haasteita voi ilmetä käytön aikana, joten suunnittelu- ja kehitystyön tulisi olla jatkuvasti osana organisaation toimintaa. Luottaminen yleisiin lähes reaaliaikaisen ETL-prosessin haasteisiin ei täysin riitä, sillä prosessia hyödynnetään usein erityisesti numeerisen datan käsittelyyn.

Numeerisen ja tekstimuotoisen datan analysointi ei ole täysin vertailukelpoista, vaikka monet yleiset haasteet, kuten suuret datamäärät, lähdejärjestelmien kuormitus ja resurssien kilpailu, koskevat molempia. Numeerista dataa voidaan käsitellä ennalta määritetyillä toiminnoilla, kun taas tekstimuotoinen data vaatii tarkempaa ennakkotietoa sisällöstä ja rakenteesta. Toistuvaa ja kontekstiltään selkeää tekstiä, kuten asiakaspalautejärjestelmästä saatua dataa voidaan käsitellä sääntöpohjaisesti. Sen sijaan vapaampimuotoisempi sähköpostipalaute vaatii kehittyneempiä tekniikoita, sen sisällön ja kertaluonteisuuden vuoksi.

Tämän vuoksi organisaation on tärkeä tunnistaa, millaista dataa prosessissa tullaan käsittelemään. Jos käytettävä data on ennalta määriteltävää ja toistuvaa, ETL-prosessi voi tarjota tehokkaan ja vakaan ratkaisun. Sen sijaan tilanteissa, joissa data on kertaluonteista, rakenteeltaan vaihtelevaa tai kontekstiltään epäselvää, prosessin

suunnittelu ja toteutus muuttuvat huomattavasti haastavammiksi. Tällöin tarvitaan kehittyneempiä työkaluja ja huomattavaa panostusta laadunvarmistukseen. Organisaation onkin syytä pohtia, onko ETL-prosessi tällaisessa tapauksessa tarkoituksenmukainen ratkaisu. Mikäli datan käsittelyyn liittyy suurta epävarmuutta eikä prosessia voida suunnitella riittävällä tarkkuudella, voi olla järkevämpää harkita vaihtoehtoisia lähestymistapoja. Huonosti suunniteltu tai väärään käyttötarkoitukseen sovellettu prosessi ei ainoastaan kuormita järjestelmiä, vaan heikentää koko analytiikan luotettavuutta ja hyödyllisyyttä. Tällöin voi olla tarpeellista pohtia, onko lähes reaaliaikainen tekstimuotoisen datan käsittely ETL-putken avulla paras vaihtoehto suorittaa datan analysointia.

5.2 Eettiset riskit

Teknisten haasteiden lisäksi prosessissa tulisi ottaa huomioon eettiset näkökulmat. Vaikka tekstin konteksti ja toistuvuus olisi organisaation tiedossa, voi tekstin sisältö silti yllättää. Yllätyksellinen teksti yhdistettynä lähes reaaliaikaiseen aika vaatimukseen tuottaa monimutkaisen tilanteen, jossa teksti tulee käsitellä nopeasti sen sisällöstä huolimatta.

Organisaatioiden tulisi valmistautua esimerkiksi siihen, että monikielinen teksti sisältää yllättäviä henkilötietoja, joita automaattiset käännökset eivät tunnista. Mikäli dataa hyödynnetään, voi se pahimmassa tapauksessa altistaa organisaation tietosuojarikkeille. Kun kyse on henkilötiedoista, virhetulkinnat voivat aiheuttaa vakavia oikeudellisia ja eettisiä seurauksia. Organisaation on arvioitava, onko prosessi riskien arvoinen ja kuka kantaa vastuun, jos prosessi epäonnistuu tai aiheuttaa haittaa ulkopuolisille. Vastuu ja tieto prosessin riskeistä ei voi jäädä vain tekniselle henkilöstölle. Organisaation johdon tulee olla tietoisia riskeistä ja varmistaa, että prosessi noudattaa laillisia ja eettisiä periaatteita.

Eettisestä näkökulmasta organisaation on syytä miettiä, voiko se kerätä ja hyödyntää tekstiä, jonka sisältö voi olla epäselviä. Organisaatio ei voi unohtaa prosessin vaikutusta yksilöön. Esimerkiksi jos yksilön henkilötietoja käytetään osana analyysia ilman yksilön suostumusta. Tällöin kyse ei ole enää teknisestä haasteesta, vaan myös yksilön oikeuksista ja luottamuksen loukkaamisesta.

5.3 Prosessin lopputuote

Prosessin yhtenä ongelmana voidaan pitää sen keskeneräisyyttä. Vaikka Vassiliadiksen ja Simitsisin (2008) mukaan ETL-prosessi on valmis, kun data on ladattu kohdejärjestelmään todellisuudessa tilanne ei kuitenkaan ole näin. Brestoff'in ja Inmonin (2015) huomauttavat, että datan tallentaminen ei viimeistele prosessia, vaan hyöty syntyy vasta kun datasta tuotetaan analyyseja päätöksentekoon. Ilman viimeistä vaihetta prosessin lopputulos voi jäädä vajaaksi organisaation kehittämisen näkökulmasta.

Tämä herättää kysymyksen siitä, onko prosessi tarpeellinen ja resurssitehokas nykyisessä muodossaan. Datan hyödyntäminen on mahdollista prosessin avulla, mutta organisaation kannalta oleellisin vaihe datan jalostus tiedoksi ja päätöksiksi jää uupumaan. Prosessin suunnittelu ja kehitys vaativat kuitenkin merkittäviä resursseja. Tämän takia prosessin tarkoitus voi jäädä irralliseksi ja epäselväksi. Organisaatioiden tulisi pohtia olisiko heidän mahdollista integroida prosessiin automaattista analytiikkaa tai suorita datan hyödyntämismenetelmiä. Lisäksi organisaation on arvioitava vastaako lähes reaaliaikainen datan käsittely organisaation tarpeita. Prosessi on hyödyllinen vain, jos organisaation vaatimukset datan saamiseen rakenteelliseen muotoon ovat lähes reaaliaikaiset. Tekstin hyödyntämisen on oltava niin välttämätöntä, ettei sen manuaalinen käsittely ole realistista.

Mikäli organisaation tarpeet ovat reaaliaikaisempia kuin mihin lähes reaaliaikainen prosessi pystyy, tulee organisaation hyödyntää reaaliaikaista putkea. Käytännössä tällainen tarve on tekstimuotoisen datan osalta harvinainen. Useimmissa tapauksissa perinteinen ETL-putki riittää käsittelemään tekstimuotoista dataa, sillä datamäärä ovat usein hallittavissa. Perinteinen ETL-prosessi on myös selvästi helpompi suunnitella ja ylläpitää, sillä se tarjoaa paremmat mahdollisuudet virheiden korjaamiseen, ennen kuin data siirtyy lopullisesti kohdejärjestelmään.

6. YHTEENVETO

Datan määrän eksponentiaalinen kasvu on tehnyt lähes reaaliaikaisista ETL-prosesseista välttämättömiä organisaatioille, jotka pyrkivät hyödyntämään dataa tehokkaasti päätöksenteossa. Tekstimuotoinen data, kuten sähköpostit ja asiakirjat, sisältää merkittäviä liiketoiminnallisia oivalluksia, mutta sen käsittelyyn liittyy huomattavia teknisiä haasteita. Tämä tutkielma keskittyi tekstimuotoisen datan haasteisiin lähes reaaliaikaisessa ETL-prosessissa.

Tutkimuksessa havaittiin, että tekstimuotoisen datan käsittelyyn liittyvät haasteet ilmenevät prosessin kaikissa vaiheissa: poiminnassa, muuntamisessa ja lataamisessa. Poimintavaiheessa korostuvat lähdejärjestelmien kuormitushaasteet ja datan oikeellisuuden varmistaminen. Muunnosvaiheessa suurimmat haasteet liittyvät datan liittämiseen eri lähteistä sekä rakenteettoman tekstin muuttamiseen rakenteelliseen muotoon lyhyessä ajassa. Latausvaiheessa resurssikiistat kohdejärjestelmässä voivat heikentää prosessin suorituskykyä ja aiheuttaa viiveitä. Näiden lisäksi tekstimuotoisen datan monimuotoisuus tekee siitä erityisen haastavan käsitellä.

Tutkielmassa esiteltiin ETL-prosessin peruseräatteen (luku 2), jonka jälkeen syvennyttiin tekstimuotoisen datan erityispiirteisiin (luku 3) sekä lähes reaaliaikaisten prosessien haasteisiin (luku 4). Lopuksi tutkimustulokset koottiin yhteen prosessin suunnittelun, toteutuksen sekä lopputuotteen näkökulmasta ja pohdittiin prosessin tarpeellisuutta (luku 5). Tulokset osoittavat, että vaikka lähes reaaliaikaisen tekstimuotoisen ETL-putken toteuttaminen on teknisesti haastavaa, se tarjoaa organisaatioille mahdollisuuden hyödyntää tekstimuotoista dataa tehokkaasti päätöksenteossa.

Kyseinen ETL-prosessi on kuitenkin monilla organisaatioilla käytössä sen haasteista huolimatta. Suurimpana syynä voidaan pitää suurta datamäärää, jonka vuoksi tekstimuotoisen datan käsittely manuaalisesti ei ole mahdollista (Brestoff & Inmon, 2015). Prosessi on kuitenkin mahdollista toteuttaa onnistuneesti ilman haasteiden ilmenemistä. Onnistuneessa tilanteessa prosessin vahvuutena pidetään sen kykyä muuttaa lähes millainen tekstityyppi tahansa tietokantamuotoon (Brestoff & Inmon, 2015).

Tutkielma on ajankohtainen ja tuo esille suurimmat haasteet organisaatioiden tekstimuotoisen datan käsittelyn osalta lähes reaaliaikaisessa ETL-prosessissa. Jatkotutkimuksissa voitaisiin keskittyä tarkemmin yksittäiseen prosessin vaiheeseen,

kehittää ratkaisuja ongelmiin tai vertailla muiden ETL-prosessien haasteita lähes reaaliaikaiseen prosessiin. Näiden lisäksi ajankohtaisena tutkimusaiheena voitaisiin käsitellä tekoälyä ja koneoppimista. Kyseisten toimintojen avulla prosessia voitaisiin mahdollisesti automatisoida ja tehostaa. Nämä voisivat tarjota uusia ratkaisuja haasteisiin, mahdollistaen nopeamman ja tarkemman prosessin toteutuksen.

LÄHTEET

Brestoff, N. E. & Inmon, W. H. (2015) *Preventing litigation :an early warning system to get big value out of big data*. First edition. New York, New York (222 East 46th Street, New York, NY 10017): Business Expert Press. Part IV.

Bornea, M. A. et al. (2011) 'Semi-Streamed Index Join for near-real time execution of ETL transformations', in *2011 IEEE 27th International Conference on Data Engineering*. [<https://doi.org/10.1109/ICDE.2011.5767906>]. 2011 IEEE. pp. 159–170.

Boulahia, C. et al. (2020) 'Towards Semantic ETL for integration of textual scientific documents in a Big Data environment: a theoretical approach', in *2020 6th IEEE Congress on Information Science and Technology (CiSt)*. [<https://doi.org/10.1109/CiSt49399.2021.9357280>]. 2020 IEEE. pp. 133–138.

Euroopan unioni. (2016). Euroopan unionin ja Neuvoston Asetus (EU) 2016/679, Yleinen tietosuojaa-asetus. Saatavilla: <https://eur-lex.europa.eu/legal-content/FI/TXT/?qid=1528874672298&uri=CELEX%3A02016R0679-20160504#toctd1>

Fan, W. & Bifet, A. (2013) Mining big data: current status, and forecast to the future. *SIGKDD explorations*. [<http://dx.doi.org/10.1145/2481244.2481246>] 14 (2), 1–5.

Inmon, W. H. et al. (2019) *Data architecture: a primer for the data scientist*. Second edition. London, England: Academic Press.

Inmon, W. H. & Nesavich, A. (2008) *Tapping into unstructured data: integrating unstructured data and textual analytics into business intelligence*. 1st edition. Upper Saddle River, N.J: Prentice Hall.

Kimball, R. & Ross, M. (2013) *The data warehouse toolkit: the definitive guide to dimensional modeling*. Third edition. Indianapolis, Ind: Wiley.

Machado, G. V. et al. (2019) DOD-ETL: distributed on-demand ETL for near real-time business intelligence. *Journal of internet services and applications*. [<https://doi.org/10.1186/s13174-019-0121-z>] 10 (1), 1–15.

Mishra, S. & Misra, A. (2017) 'Structured and Unstructured Big Data Analytics', in *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*. [<https://doi.org/10.1109/CTCEEC.2017.8454999>]. 2017 IEEE. pp. 740–746.

MongoDB. (2025). *Relational vs. Non-Relational Databases*. MongoDB. Saatavilla: <https://www.mongodb.com/resources/compare/relational-vs-non-relational-databases> Viitattu: 27.1.2025.

Palmer, M. (2024) *Understanding ETL*. First edition. Sebastopol, CA: O'Reilly Media, Inc.

Raj, A. et al. (2020) 'Modelling Data Pipelines', in *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. [<https://doi.org/10.1109/SEAA51224.2020.00014>]. 2020 IEEE. pp. 13–20.

Reeve, A. (2013) *Managing Data in Motion: Data Integration Best Practice Techniques and Technologies*. 1st edition. Chantilly: Elsevier Science & Technology.

Sabtu, A. et al. (2017) 'The challenges of Extract, Transform and Loading (ETL) system implementation for near real-time environment', in *2017 International Conference on Research and Innovation in Information Systems (ICRIIS)*. [<https://doi.org/10.1109/ICRIIS.2017.8002467>]. 2017 IEEE. pp. 1–5.

Statista (2024) 'Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2023, with forecasts from 2024 to 2028', *Statista*. Saatavilla: <https://www.statista.com/statistics/871513/worldwide-data-created/>. Viitattu: 7.3.2025.

Vassiliadis, P. & Simitsis, A. (2008). Chapter 2. Teoksessa: Kozielski, S. & Wrembel, R. (toim.) *New Trends in Data Warehousing and Data Analysis*. 1. Aufl. Vol. 3 (s. 19-49). [<https://doi.org/10.1007/978-0-387-87431-9>]. New York, NY: Springer-Verlag.

Vijayalakshmi, M. & Minu, R. I. (2021) Novel Solution for Real Time Challenges of ETL in Big Data. *Turkish journal of computer and mathematics education*. 12 (10), 3661–3674.