



SSIS televerkkojen väärinkäytösten havainnoijana

Kalevi Karjalainen

Opinnäytetyö
Joulukuu 2013
Tietojenkäsittelyn koulutus-
ohjelma

TIIVISTELMÄ

Tampereen ammattikorkeakoulu
Tietojenkäsittelyn koulutusohjelma

KARJALAINEN KALEVI:
SSIS televerkkojen väärinkäytösten havainnoijana

Opinnäytetyö 58 sivua
Joulukuu 2013

Televerkoissa tapahtuvat väärinkäytökset ovat maailmanlaajuinen ongelma ja aiheuttavat teleoperaattoreille mittavia tulonmenetyksiä. Telekommunikaatioon liittyviä väärinkäytöstapoja on havaittu runsaasti ja niiden löytämiseksi on kehitetty erilaisia lähestymistapoja ja tekniikoita monimutkaisista itse oppivista algoritmeista aina ihmissilmällä tapahtuvaan visuaaliseen tarkasteluun asti.

Opinnäytetyön tarkoituksena oli toteuttaa DNA Oy:lle työkalut DNA:n televerkoissa tapahtuvien väärinkäytösten havaitsemiseksi tukeutuen Microsoftin SQL Server 2008 R2 Integration Services -palvelinohjelmistoon. Työn tavoitteena oli vähentää toimeksiantajan televerkoissa tapahtuvia väärinkäytöksiä ja niistä aiheutuvia kustannuksia erityisesti palvelunumeropalveluiden osalta.

Työssä käytettiin pääasiassa konstruktivistista tutkimusmenetelmää, jossa havaitulle tarpeelle etsittiin ratkaisu tukeutuen ennalta valittuun kehitysalustaan. Valitulle alustalle toteutettiin sääntöpohjainen valvontatyökalu televerkkojen väärinkäytösten havainnointiin ja raportointiin sekä määriteltiin työkalulle tarvittavat säännöt palvelunumeropalveluiden väärinkäytösseurannan mahdollistamiseksi. Työkalun toimintaperiaatteena on hakea väärinkäytöstapauksia sen saamien valvontakäskeyjen mukaisesti ja raportoida löydetty väärinkäytösepäilyt tarvittaville tahoille.

Toteutettu työkalu vastasi asetettuja tavoitteita. Työkalun generoimien valvontaraporttien perusteella tehtiin toimeksiantajan televerkoissa estotoimenpiteitä, joiden seurauksena palvelunumeropalveluiden väärinkäytösepäilyjen kumulatiivinen euromäärä laski kahden kuukauden vertailujaksolla 37 prosenttia ja väärinkäytösepäilyt 34 prosenttia. Työkalun avulla löydettiin myös teknisiä heikkouksia, jotka mahdollistivat tietyissä tapauksissa palveluiden ilmaisen käytön.

SSIS sopii myös arkkitehtuurinsa puolesta erinomaisesti televerkkojen väärinkäytösten havainnointiin monipuolisten työnkulkujen, kattavien tietolähteiden ja suuren suorituskykynsä ansiosta, kunhan huolehditaan käyttäjien riittävästä osaamisesta.

ABSTRACT

Tampereen ammattikorkeakoulu
Tampere University of Applied Sciences
Degree Programme in Business Information Systems

KARJALAINEN KALEVI
SSIS in Fraud Detection

Bachelor's thesis 58 pages
December 2013

Telecommunications fraud is a worldwide problem that deprives operators of enormous sums of money. Many forms of telecommunication fraud have been identified. There are also many different approaches and techniques for detection of telecommunications fraud from self-learning algorithms to visual scrutiny of human eye.

The purpose of this thesis was to design and implement a fraud detection tool for DNA Ltd by using Microsoft SQL Server R2 Integration Services. The goal of this thesis was to decrease the number of frauds and losses in employer networks.

The thesis was based on a constructive research approach where a rule-based tool for fraud detection was implemented in a predefined platform. Appropriate detection fraud rules for Service Numbers were designed during implementation as well. The main idea of the tool is to detect telecommunication frauds based on given rules and to report suspicious cases to necessary parties.

The implemented rule-based tool gave promising results. The cumulative amount of losses in euros decreased by 37 % and the number of suspicious case by 34 % during the comparison period. Technical weaknesses which allowed use services without payment were also found.

The architecture of SSIS is suitable for fraud detection due to versatile workflows, comprehensive data sources and high performance. However, this requires that the users are familiar with the system.

Keywords: Fraud detection, SSIS

SISÄLLYS

1	JOHDANTO.....	6
2	TELEVERKKOJEN VÄÄRINKÄYTÖKSET	7
2.1	Mitä fraud on?.....	7
2.2	Taloudelliset vaikutukset	7
2.3	Väärinkäytösten luokittelu	8
3	TELEVERKKOJEN VÄÄRINKÄYTÖSTEN HAVAINNOIMISMALLI	11
3.1	Tiedon keruu ja yhtenäistäminen	11
3.2	Väärinkäyttöä osoittavien indikaattoreiden tunnistaminen.....	13
3.3	Havainnoimistekniikoiden valinta	14
4	HAVAINNOINTITEKNIIKAT	16
4.1	Sääntöperustaiset tekniikat	16
4.2	Tiedonlouhintaan perustuvat tekniikat.....	17
4.3	Muut havainnoimistekniikat	25
5	SSIS-YMPÄRISTÖN ESITTELY	26
5.1	SSIS-pakettien rakenne	28
5.1.1	Control Flow	29
5.1.2	Data Flow	33
5.1.3	Connection Managers.....	36
5.1.4	Variables	37
5.1.5	Event Handlers	38
5.1.6	Log Providers	40
5.2	SSIS-pakettien muodostaminen.....	41
5.3	SSIS-pakettien konfigurointi ja jakelu.....	42
5.4	SSIS-pakettien suoritus	44
6	TOTEUTUS	46
7	POHDINTA.....	54
	LÄHTEET.....	56

LYHENTEET JA TERMIT

BIDS	Business Intelligence Development Services
CDR	Call Detail Record. Puhelutietue
CRM	Customer Relationship Management. Asiakkuuden hallinta
ETL	Extract, Transform, and Load. Yleisesti tietojärjestelmien välisessä tiedonsiirrossa käytetty malli
FMS	Fraud Management System. Väärinkäytöksiä havainnoiva järjestelmä
IMSI	International Mobile Subscriber Identity. Tilaaajan verkossa yksilöivä tunnus
IMEI	International Mobile Equipment Identity. Mobiili-päätelaitteen verkossa yksilöivä tunnus
SSAS	SQL Server Analysis Services
SSIS	SQL Server Integration Services

1 JOHDANTO

Televerkoissa tapahtuvat väärinkäytökset ovat maailmanlaajuinen ongelma ja niistä aiheutuu operaattoreille vuosittain mittavat tulonmenetykset. Operaattorit ovatkin alkaneet kehittää toimintatapojaan ja tietojärjestelmiään väärinkäytösten pienentämiseksi.

Opinnäytetyön toimeksiantajana on DNA Oy. DNA Oy on suomalainen tietoliikennekonserni, joka tarjoaa yksityishenkilöille ja yrityksille laadukkaita, viimeisintä teknologiaa hyödyntäviä puhe-, data- ja tv-palveluita. DNA:n liikevaihto vuonna 2012 oli 769 miljoonaa euroa ja liikevoitto 56 miljoonaa euroa. Tilaajia DNA:n televerkoissa on noin 2,5 miljoonaa (syyskuu 2013).

Opinnäytetyön tarkoituksena on määrittää ja toteuttaa työkalut toimeksiantajan televerkoissa tapahtuvien väärinkäytösten havaitsemiseksi. Työn tavoitteena on vähentää toimeksiantajan televerkoissa tapahtuvia väärinkäytöksiä ja niistä aiheutuvia kustannuksia. Toimeksiantoa on rajattu opinnäytetyön yhteydessä siten, että opinnäytetyössä kuvataan Microsoftin SQL Server 2008 R2 Integration Services -palvelinohjelmistolla toteutettu sääntöpohjainen työkalu palvelunumeropalveluiden väärinkäytösten havaitsemiseen mahdollisimman aikaisessa vaiheessa.

Opinnäytetyön tekemiseen käytetään pääasiassa konstruktivistista tutkimusmenetelmää, jossa havaitulle ongelmalle/tarpeelle etsitään ja toteutetaan ratkaisu tukeutuen ennalta valittuun kehitysalustaan.

Opinnäytetyö jakautuu seitsemään lukuun. Luvut kaksi, kolme ja neljä pyrkivät antamaan lukijalle tarvittavan yleiskuvan televerkkojen väärinkäytöksiin liittyvistä väärinkäytöstekniikoista ja -tavoista, väärinkäytösten havainnointiin liittyvistä eri vaiheista sekä väärinkäytösten havainnointitekniikoista. Luvussa viisi esitellään Microsoftin SQL Server 2008 Integration Services -ympäristö, johon tukeutuen luvussa kuusi kuvattu työkalu on toteutettu. Lopuksi, luvussa seitsemän esitetään työn tulokset ja tulevat jatkotoimenpiteet, sekä pohditaan alustan soveltuvuutta väärinkäytösten havainnointiin.

2 TELEVERKKOJEN VÄÄRINKÄYTÖKSET

2.1 Mitä fraud on?

Yksinkertaisella tasolla fraudia voidaan kuvata kaikella sillä toiminnalla, jossa hankittua palvelua ei ole aikomustakaan itse maksaa (Gosset & Hyland 1999). Kyse on siis rikollisesta toiminnasta, jota voidaan kansanomaisemmin kuvata termeillä väärinkäyttö tai varkaus. Telekommunikaatioon tai ylipäättään teletoimialaan liittyvällä fraudilla viitataan väärinkäytöksiin, jotka tapahtuvat teleoperaattorin järjestelmiä hyväksikäyttäen. Erityisesti telealalla fraud on muodostunut vakiintuneeksi käsitteeksi puhuttaessa televerkkojen väärinkäytöksistä.

Teletoimialan näkökulmasta väärinkäytöksiin syyllistyvät eniten organisoidut rikollisryhmät, edistyneet hakkerit sekä teleoperaattoreiden oma henkilökunta. On arvioitu, että jopa 73 prosenttia kaikista telealan väärinkäytöksistä aiheutuisi operaattoreiden omista työntekijöistä (Beck Computer Systems 2003, Bihina Bellan, Eloffin ja Olivierin 2005b 2005 mukaan).

Maantieteellisesti väärinkäytöksiä organisoidaan eniten Yhdysvalloista, Intiasta, Iso-Britanniasta, Brasiliasta ja Filippiineiltä. Yleisimmät kohdemaat ovat Latvia, Gambia, Somalia, Sierra Leone ja Guinea (Global Fraud Loss Survey 2013.)

Päämotiivi väärinkäytöksille on luonnollisesti raha. Rahaa voidaan ansaita myymällä esimerkiksi vilpillisin keinoin haltuun otettuja palveluita eteenpäin. Toinen yleinen motiivi väärinkäytöksille on halu saada teleoperaattorin tarjoamia palveluita ilmaiseksi tai halvemmalla hinnalla.

2.2 Taloudelliset vaikutukset

Teleoperaattoreihin kohdistuva fraud on maailmanlaajuinen ongelma. Vaikka fraud on laskenut vuonna 2011 peräti 33 prosenttia vuoden 2008 huippuvuodesta, on arvioitu, että teleoperaattorit menettivät vuonna 2011 väärinkäytösten seurauksena noin 30 miljardia euroa (Communications Fraud Control Association (CFCA) 2011). Edellä maini-

tun raportin mukaan väärinkäytöksistä johtuneet tulon menetykset olivat vuonna 2011 keskimäärin 1,88 prosenttia operaattoreiden kokonaistuloista. Vuoden 2008 huippuvuodesta laskua oli 1,66 prosenttiyksikköä. Pääsyy suhteellisen prosenttiosuuden pienentymiseen selittyy operaattoreiden kokonaistulojen kasvulla, joka ylitti väärinkäytöksistä aiheutuneiden tulon menetysten kasvun. Raportin mukaan myös väärinkäytösten ehkäisemiseen on alettu kiinnittämään entistä enemmän huomiota kehittämällä väärinkäytösten havainnoimiseen liittyviä toimintatapoja ja ohjelmistoja.

2.3 Väärinkäytösten luokittelu

Telekommunikaatioon liittyviä erilaisia väärinkäytöstopoja on havaittu yli 200 kappaletta (Jacobs 2002, Bihina Bellan ym. 2005b mukaan). Usein väärinkäytöksissä käytetään useaa eri menetelmää taloudellisen hyödyn maksimoimiseksi. Yksinkertaistettuna televerkkojen väärinkäytökset voidaan kuitenkin luokitella neljään pääluokkaan: sopimusväärinkäyttöksiin (Contractual Fraud), murtautumisväärinkäyttöksiin (Hacking Fraud), teknisiin väärinkäyttöksiin (Technical Fraud) ja menettelytapaväärinkäyttöksiin (Procedural Fraud) (Gosset & Hyland 1999).

Sopimusväärinkäytökset

Kaikki väärinkäytökset tässä kategoriassa perustuvat olemassa olevien palveluiden normaaliin käyttöön, joista ei ole tarkoitukseen maksaa. Esimerkkeinä tähän kategoriaan liittyvistä väärinkäytöksistä voidaan mainita tilaajaväärinkäytökset (Subscription Fraud) ja lisämaksullisten palvelunumeroiden väärinkäytökset (Premium Rate Fraud).

Gossetin ja Hylandin (1999) mukaan tilaajaväärinkäytökset voidaan jakaa kahteen osaan: sopimussuhteisiin, joita ei ole aikomustakaan maksaa, sekä tapauksiin, jossa tilaaja kesken sopimuskauden päättää olla maksamatta käyttämistään palveluista. Näiden väärinkäytösten havainnointi on vaikeaa. Ero vilpillisen ja ei vilpillisen käyttäytymisen välillä voi olla äärimmäisen pieni ja vaikeasti havaittavissa. Tilaaaja voi esimerkiksi kesken sopimuskauden päättää, ettei maksa käyttämistään palveluista. Tällöin tilaajan oletuskäyttäytymisessä tapahtuu yleensä dramaattisia muutoksia, jotka on mahdollista havaita (Hilas 2012). Mikäli tilaaja on jo sopimusta tehdessään päättänyt, että ei aio maksaa käyttämistään palveluista, ei operaattoreilla ole tilaajasta aikaisempaa

käyttöhistoriaa ja vertailu tehokäyttäjiin voi olla vaikeaa. Tällöin lisäinformaation saaminen asiakkaasta sopimusta tehtäessä on ainoa keino minimoida mahdollisia riskejä. Hyvänä esimerkkinä riskien minimoimisesta on asiakkaan tunnistus ja luottotietojen tarkistus liittymäsopimuksia tehtäessä (Kamtsan, Tirkkonen & Sihvola 2012).

Lisämaksullisissa palveluissa tulot jaetaan tyypillisesti operaattorin ja palveluntarjoajan kesken siten, että operaattori perii palvelumaksut asiakkailta ja tulouttaa ne edelleen palveluntarjoajalle veloittaen tuloutuksesta sovitun prosenttiosuuden itselleen. Lisämaksullisten palveluiden väärinkäytöksissä on tyypillisesti kaksi vaihetta. Ensin perustetaan lisämaksullinen palvelu ja tämän jälkeen pyritään saamaan kyseiseen palveluun mahdollisimman paljon liikennettä siten, että käytetyistä palveluista ei kuitenkaan makseta operaattorille palvelumaksua. Tyypillisesti tämä tapahtuu varastettujen tai kloonattujen liittymien avulla (Gosset & Hyland, 1999). Täysin asiallisia lisämaksullisia palveluita voidaan pyrkiä väärinkäyttämään myös tiedossa olevia operaattorin tuotantojärjestelmien heikkouksia hyväksikäyttäen. Palvelun toteutuksessa saattaa olla puutteita, jonka vuoksi vilpillinen taho pystyy saamaan palvelua ilman vastiketta.

Murtautumisväärinkäytökset

Murtautumisväärinkäytöksissä murtaudutaan suojattuihin järjestelmiin ja valjastetaan ne tämän jälkeen omaan käyttöön myymällä esimerkiksi liikennettä murtaudutun järjestelmän kautta. Tyypillisiä tämän kategorian tapauksia ovat mm. vaihteisiin kohdistuvat murtautumiset (ns. PABX Fraud) sekä verkkohyökkäykset.

Vaihteisiin kohdistuvissa murtautumisissa vilpillinen taho pyrkii vaihteeseen kohdistuvilla toistuvilla soitoilla saamaan pääsyn vaihteen ulosmenolinjalle. Mikäli yritys onnistuu, voidaan vaihteen kautta kierrättää liikennettä esimerkiksi ulkomaille tai lisämaksullisiin palvelunumeroihin maksaen ainoastaan yhteydestä vaihteeseen. Gossetin ja Hylandin (1999) mukaan tapauksiin liittyy usein myös varastettuja tai kloonattuja liittymiä, jolloin liikenteen kierrättäminenkin on vilpilliselle taholle täysin ilmaista.

Verkkohyökkäyksissä pyritään murtautumaan järjestelmään ylläpito- tai etäliittymän kautta. Murtautumisen onnistuessa, hyökkääjä tyypillisesti konfiguroi järjestelmän kierrättämään omaa liikennettä. Hyvänä esimerkkinä verkkohyökkäyksistä on VoIP-

liikenteen (Voice over IP) SIP-proxyt, joihin murtaudutaan ja kierrätetään tämän jälkeen ulkomaanliikenne proxyn kautta. Jopa yksittäinen hyökkäys saattaa aiheuttaa mittavia taloudellisia menetyksiä. Eräässä tapauksessa yksittäisen hyökkäyksen aikana onnistuttiin ohjaamaan 46 tunnin aikana yli 11000 ulkomaanpuhelua aiheuttaen operaattorille lähes 100 000 euron menetyksen (Tindal 2009).

Tekniset väärinkäytökset

Tämän kategorian väärinkäyttötapaukset perustuvat hyökkäyksiin teknologian heikkouksia vastaan esimerkiksi matkapuhelinverkossa. Tekninen väärinkäyttö vaatii tyypillisesti hyvää teknistä tietotaitoa, vaikkakin internet-aikakaudella tietämys jaetaan usein sellaisessa muodossa eteenpäin, että heikkouden hyväksikäyttö onnistuu myös tahoilta, joilla itsellä ei muuten tietotaitoa ja kykyä asiaan olisi. Esimerkkinä teknisistä väärinkäytöksistä voidaan mainita kloonaus ja tekninen sisäinen väärinkäyttö.

Kloonauksessa mobiilipäätelaitteen tunnistautumisparametrit kopioidaan toiseen päätelaitteeseen siten, että verkko uskoo kyseessä olevan alkuperäisen, verkkoon autentikoituneen päätelaitteen.

Teknisissä sisäisissä väärinkäytöksissä vilpillinen työntekijä saattaa muuttaa operoiminsa järjestelmien sisäisiä tietoja siten, että osa käyttäjistä pystyy käyttämään palveluita alemmilla kustannuksilla.

Menettelytapaväärinkäytökset

Kategoriaan kuuluvissa väärinkäytöksissä käytetään hyväksi tiedossa olevia heikkouksia palvelun toimintatavoissa. Esimerkiksi roaming-tapauksissa roaming-operaattorin lähettämät puhelut saattavat tulla asiakasta laskuttavan operaattorin järjestelmiin hitaimmillaan jopa useiden viikkojen viiveellä. Vilpillinen taho saattaa pystyä tiedossa olevaa heikkoutta hyväksikäyttämällä tekemään suuren laskun operaattorin lukuun.

3 TELEVERKKOJEN VÄÄRINKÄYTÖSTEN HAVAINNOIMISMALLI

Televerkkojen väärinkäytösten havainnointi jaetaan yleisesti kolmeen osaan: tiedon keruuseen ja yhtenäistämiseen, väärinkäytöksistä kertovien indikaattoreiden tunnistamiseen sekä väärinkäytösten tunnistamiseen tarkoituksenmukaisia havainnointitekniikoita hyväksikäyttäen.

3.1 Tiedon keruu ja yhtenäistäminen

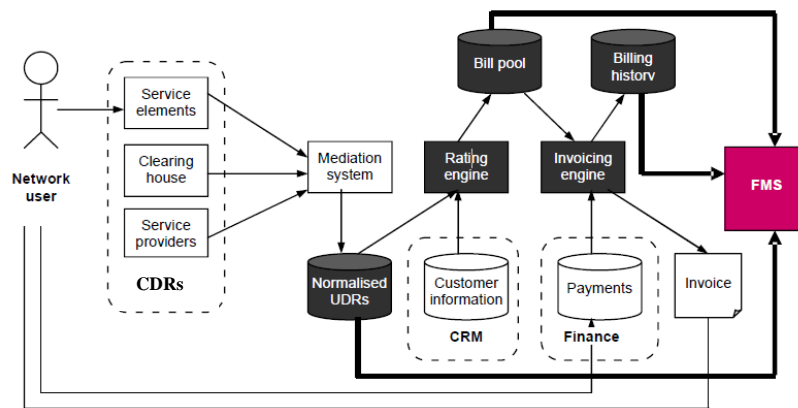
Tarvittavan tiedon keruu erilaisia analyyseja varten on ensimmäinen askel väärinkäytösten havainnointiprosessissa (Bihina Bella ym. 2005b). Tyypillisesti väärinkäytösten havainnoimisen primaareina lähtötietoina käytetään televerkkojen verkkoelementtien - kuten keskusten (Switch) ja yhdyskäytävien (Gateway) - generoimia tapahtumatietueita. Näitä verkkoelementtien muodostamia tapahtumatietueita kutsutaan telealalla yleisesti lyhenteellä CDR (Call Detail Record).

CDR sisältää tarvittavat tiedot laskutuksen ja erilaisten vianselvitysten mahdollistamiseksi. Oleellisia tietoja ovat mm. tiedot käyttäjän identiteetistä (IMEI, IMSI, A-numero), mihin on oltu yhteydessä (B-numero, APN, IP-osoite), minkä tyyppisestä palvelusta tai tapahtumasta on kysymys (esim. tekstiviesti, puhelu palvelunumeroon tai datayhteys) ja kuinka pitkään palvelun käyttö on kestänyt (Hinde 1996.)

Laskutustarkoitusta varten CDR:iä voi tuottaa myös Clearing house roaming-partnereiden roaming-tapahtumista sekä mahdollisesti kolmansien osapuolten palveluoperaattorit (Bihina Bella, Olivier & Eloff 2005a). Tyypillisesti eri verkkoelementtien ja osapuolten tuottamat CDR-tiedot ovat laitevalmistajakohtaisia binääriformaatteja, jotka pitää konvertoida yhtenäiseen muotoon ennen tietojen jatkohyödyntämistä. Telealalla tätä toimintoa kutsutaan medioinniksi.

Seuraavassa kuvassa 1 esitetään teleoperaattoreiden tyypillinen laskutusprosessi ja sen yhteydet väärinkäytösten havainnoimisjärjestelmään (FMS). Kuvaan on liitetty myös käyttötapaus, jossa asiakas käyttää operaattorin tarjoamaa palvelua. Käytetystä palvelusta riippuen verkkoelementti, Clearing house tai palveluoperaattori rekisteröi tapahtuman

ja tekee tapahtumasta binäärimuotoisen CDR:n. Mediointijärjestelmä noutaa CDR:t, tekee tarvittavat oikeellisuustarkistukset ja muuntaa ne laskutusjärjestelmän ymmärtämään formaattiin. Normalisoidut CDR:t lähetetään hinnoitteluun (Rating Engine) joko eräajomuotoisesti tai välittömästi tapahtuman käsittelyn jälkeen. Hinnoittelussa CDR:t yhdistetään asiakkaiden tileihin CRM-järjestelmän avulla ja lasketaan tarvittavat palvelumaksut CDR-kohtaisesti. Hinnoitellut CDR:t varastoidaan (Bill pool) odottamaan seuraavaa laskutusajoa (Invoice engine), jossa muodostetaan kertyneistä tapahtumista laskut lisäten samalla asiakkaille kuukausi ja muut toistuvaismaksut. Kun taloushallinnon järjestelmistä (Finance) saadaan tiedot maksusuorituksista, arkistoidaan suoritettut laskut erilliseen historiakantaan (Billing history). Kuvasta myös nähdään, että FMS voi kerätä lähtötietoa useasta eri tietolähteestä.



Kuva 1. Esimerkki laskutusprosessista ja liitännät väärinkäytösten havainnointiin (Bihina Bella ym. 2005a)

CDR:iin tukeutuminen fraud-valvonnassa ei ole ongelmattonta. Valmistajakohtaisten formaattien lisäksi niiltä puuttuu joustavuus luotettavasti kuvata käytettyä palvelua, jolloin ne eivät ole suoraan yhteensopivia muiden verkkoelementtien tuottamien CDR:ien kanssa. Lisäksi CDR:t generoidaan vasta palvelun käyttämisen jälkeen ja prosessoidaan eräajotyyppisesti, jolloin reaaliaikainen väärinkäytösten seuranta on käytännössä mahdotonta (Hearne 2004, Bihina Bellan ym. 2005a mukaan).

Telekommunikaatioalalla puhutaan myös todella suurista tietomääristä. Esimerkiksi AT&T:lla syntyi yli 300 miljoonaa CDR:ää vuorokaudessa pelkästään kaukopuheluasiakkailta (Cortes & Pregibon 2001, 167). Huikkeista määristä johtuen onkin tärkeää löytää sopivat indikaattorit ja tarkoituksenmukaiset havainnointimenetelmät mahdollisten väärinkäytösten löytämiseksi.

3.2 Väärinkäyttöä osoittavien indikaattoreiden tunnistaminen

Fraud-indikaattorit ovat palveluiden käytöstä saatavia tietoja, jotka saattavat indikoida väärinkäyttöä (Kvarnström, Lundin & Jonsson 2000). Perinteisissä televerkoissa tyypillisiä indikaattoreita voivat olla esimerkiksi puhelun kesto, sykäykset, suuri puhelumäärä yksittäisestä numerosta ja puhelut tiettyihin numerosuuntiin. Kaikki indikaattorit eivät välttämättä ole yksinään käyttökelpoisia, vaan vasta sopivina kombinaatioina muiden indikaattoreiden kanssa. Joskus sopivat indikaattorit tai niiden kombinaatiot saattavat löytyä vasta jälkikäteen, kun palvelusta ja mahdollisesta väärinkäytöksestä on saatu riittävästi analysoitavaa materiaalia (Kvarnström ym. 2000).

Rossetin, Muradin, Neumanin, Idanin ja Pinkan (1999) mukaan CDR:ltä saatavat tiedot eivät aina yksinään riitä todentamaan fraudia. Tapahtuma saattaa olla täysin normaali tietyssä tilanteessa, mutta indikoida väärinkäytöstä jossakin toisessa tapauksessa. Esimerkiksi soittot palvelunumeroihin saattavat olla normaaleja, jos asiakas yleensä soittaa niihin, mutta muussa tapauksessa epäilyttäviä. Tämän vuoksi Rossetin ym. (1999) mukaan tarvitaan tietoja myös itse asiakkaasta ja asiakkaan käyttäytymisestä.

Perusidea käyttäytymisen seurannassa on se, että käyttäjän tai käyttäjien käyttöhistorias-
ta muodostetaan tarkoituksenmukainen profiili tai profiilit, jota verrataan käyttäjän tai tietyn ryhmän tulevaan käyttäytymiseen pyrkien havainnoimaan mahdolliset poikkeamat näiden välillä. Profiili sisältää valittujen indikaattoreiden lisäksi numeerista summatason tietoa erikseen valituista indikaattoreista tietyltä ajanjaksolta. Tyypillisiä summatason tietoja ovat esimerkiksi puheluiden lukumäärät, kestot ja datamäärät.

Vaikka käyttövolyymit - kuten lukumäärät, kestot ja datamäärät - ovat keskeisiä väärinkäytösten havainnoinnissa, tarvitaan niiden lisäksi tietoa myös asiakkaista. Väärinkäytösanalyysien näkökulmasta käyttökelpoisia tietoja ovat Rossetin ym. (1999) mukaan mm. asiakkuuden ikä, etninen tausta, luottotiedot, liittymätyyppi ja laskutushistoria. Esimerkiksi asiakkaan huono luottokelpoisuusarvio saattaa vaatia monitoroimaan asiakasta tiukemmilla kynnysarvoilla. Samoin esimerkiksi uusia asiakkaita pitää analysoida eri mallilla kuin vanhoja asiakkaita.

3.3 Havainnoimistekniikoiden valinta

Väärinkäytösten havainnointiin voidaan käyttää useita erilaisia tekniikoita tai niiden yhdistelmiä. Yleisesti ottaen telealaan liittyvät havainnointitekniikat voidaan jakaa alla olevan kuvan mukaisesti kolmeen osaan: sääntöperustaisiin tekniikoihin (Rule-based), tiedonlouhintaan (Data mining) perustuviin tekniikoihin sekä muihin havainnoimistekniikoihin (Augustin ym. 2012; Hearne 2004).



Kuva 2. Havainnointitekniikoiden kolmijako (Augustin ym. 2012, muokattu)

Sääntöperusteisissa tekniikoissa fraud-tunnisteet määritellään sääntöinä joita verrataan käyttäjätietoihin vasten. Sääntöperustaiset tekniikat ovat työläisiä ylläpitää, mutta erittäin tehokkaita tiedossa olevien väärinkäytösten havainnoimisessa. Tiedonlouhintaan perustuvat tekniikat osaavat itsenäisesti havaita merkittävät muutokset käyttäjien oletuskäyttäytymisessä (Hearne 2004, Bellan ym. 2005b mukaan.) ja soveltuvat erinomaisesti havainnoimaan esimerkiksi uusia, vielä tunnistamattomia väärinkäytöksiä (Gosset & Hyland 1999). Tiedonlouhintatekniikat jaetaan kahteen osaan perustuen tekniikojen tapaan oppia tunnistamaan poikkeamat: ohjattuihin (supervised) ja ohjaamattomiin (unsupervised) tiedonlouhintamethodeihin.

Väärinkäytösten havainnoimista voidaan lähestyä myös oppimisen, opettamisen ja tutkimisen näkökulmasta (Gosset & Hyland 1999).

Oppimiseen perustuva lähestymistapa käytetään tyypillisesti ohjaamattomissa metodeissa, joissa havainnointityökalu opettaa itse itsensä, mikä on käyttäjien odotettua käyttäytymistä. Tämä lähestymistapa on käyttökelpoinen havainnoitaessa muutoksia oletuskäyttäytymisessä ja soveltuu erinomaisesti esimerkiksi tilaaja- ja murtautumisvää-
rinkäytösten havainnointiin (Gosset & Hyland 1999). Oppimiseen perustuvassa lähes-

tymistavassa havainnointityökalu siis oppii, mikä on tyypillistä käyttäytymistä ja osaa reagoida automaattisesti poikkeamiin. Oppimiseen perustuvat työkalut ovatkin erittäin käyttökelpoisia havainnoimaan väärinkäytöksiä, joista ei ennalta ole tietoa (Gosset & Hyland 1999).

Oppimiseen perustuvassa lähestymistavassa on joitakin haittapuolia. Oppimiseen perustuvia työkaluja ei ole mahdollista opettaa etsimään jotakin tiettyä asiaa. Lisäksi, mikäli parametreja ei ole asetettu täsmälleen oikein, voi taitava väärinkäyttäjä oppia sekoittamaan käyttöönsä siten, että hälytyksiä ei muodostu (Gosset & Hyland 1999).

Opettamiseen perustuvaa lähestymistapaa käytetään tyypillisesti ohjatuissa metodeissa ja sääntöperustaisissa havainnointityökaluissa tilaaja- ja murtautumisväärinkäytösten havainnointiin sekä joskus myös teknisten väärinkäytösten havainnoimiseen (Gosset & Hyland 1999). Lähestymistavassa havainnointityökalu opetetaan havaitsemaan väärinkäyttö olemassa olevien väärinkäytöstapausten perusteella. Sääntöperusteisten työkalujen tapauksessa olemassa olevista väärinkäytösesimerkeistä analysoidaan fraud-indikaattorit, jotka tämän jälkeen muutetaan säännöiksi. Ohjatuissa tiedonlouhintameto- deissa opettaminen tapahtuu siten, että työkalulle annetaan esimerkkidataa sekä laillises- ta että laittomasta käytöstä, jonka jälkeen työkalu annetun esimerkkidatan perusteella itse oppii, mikä käyttäytyminen on laillista ja mikä laitonta.

Tutkimiseen perustuvassa lähestymistavassa etsitään heikkouksia esimerkiksi menette- lytavoista ja teknisistä määrittelyistä. Lähestymistapa on käyttökelpoinen teknisessä ja sopimusrikosväärinkäytöksissä (Gosset & Hyland 1999). Tutkimiseen perustuvaan lä- hestymistapaan voidaan lukea myös kaikki ihmisen tekemä muu tarkastelu. Hyvänä esimerkkinä tästä voisi olla visualisointiin perustuvat tekniikat, joissa väärinkäytösten tunnistus tuotetusta graafisesta aineistosta perustuu ihmissilmään.

Mitään yhtä oikeaa havainnointitapaa kaikkien väärinkäytösten tunnistamiseksi ei ole olemassa. Yksi tekniikka voi soveltua erinomaisesti johonkin tiettyyn väärinkäytöstapa- uksen havainnointiin, mutta voi olla täysin soveltumaton johonkin toiseen (Gosset & Hyland 1999). Paras tulos saavutetaan, kun käytetään useaa eri tekniikkaa yhtä aikaa (mm. Taniguchi, Haft, Hollmén & Tresp 1998; Bihina Bella, Eloff & Olivier 2009).

4 HAVAINNOINTITEKNIIKAT

4.1 Sääntöperustaiset tekniikat

Sääntöperusteisissa analyyseissa fraud-tunnisteet esitetään sääntöinä, joita verrataan käyttäjädataan (CDR). Säännöt voivat sisältää useita ehtoja ja kaikkien ehtojen toteutuessa generoidaan tapauksesta hälytys (Rosset ym. 1999).

Sääntöperustaiset metodit ovat erittäin tehokkaita, mutta työläitä ylläpitää (Kou, Lu, Sirwongwattana & Huang 2004). Havainnoinnin mahdollistamiseksi pitää kaikki kuviteltavissa olevien väärinkäyttöskenaarioiden tunnisteet määritellä ja toteuttaa erikseen. Sääntöjen määrittämisessä tarvitaan lisäksi laajaa palveluiden ja niiden ominaisuuksien tietämystä (Bihina Bella ym. 2005b). Asiantuntijoiden tietämyksen lisäksi myös aikaisemmista väärinkäytöshavainnoista on suuresti hyötyä. Tyypillisesti fraud-säännöt määrittäänsäkin yhdessä asiantuntijoiden tietämyksen ja aikaisemmista väärinkäytöstapauksista saatujen tietojen perusteella (Rosset ym. 1999). Uusia fraud-tekniikoita ilmestyy kuitenkin jatkuvasti, joten sääntöjä pitää päivittää säännöllisin väliajoin kattamaan myös uudenlaiset väärinkäytöstapaukset (Augustin ym. 2012).

Rosset ym. (1999) suosittelee automaattisia, erilaisiin tiedon louhinta-algoritmeihin (Data mining) perustuvia analyyseja, joita ajetaan tallennettua historiadataa vasten entistä tehokkaampien sääntöjen muodostamiseksi. Näiden analyyksien avulla saadaan myös erinomaisesti päivitettyä fraud-asiantuntijoiden tietämystä.

Sääntöperustaisiin metodeihin voidaan laskea myös erilaiset kynnsarvot (thresholds) ja luettelot esimerkiksi luotetuista tai ei-luotetuista asioista (white list/black list).

Kynnsarvoihin perustuvissa analyyseissa CDR:n tietoja (esim. puhelun kesto) verrataan ennalta määriteltuihin kiinteisiin arvoihin. Jos CDR:n tieto ylittää kynnsarvon, tapauksesta generoidaan hälytys. Kynnsarvoihin perustuvat työkalut ovat yksinkertaisia ja tehokkaita äärimmäisissä väärinkäytöstapauksissa (Bihina Bella ym. 2005b). Tyypillisesti kynnsarvoja käytetään siis havainnoimaan suurimmat poikkeamat. Esimerkiksi puhelukohtaiseksi sykäysrajaksi voitaisiin asettaa 1000 sykäystä, jonka ylittyessä tapauksesta generoidaan hälytys.

White/Black – listoja voidaan hyödyntää sellaisenaan tai esimerkiksi osana sääntöperustaisia metodeja. Listojen avulla voidaan määritellä esimerkiksi puhelusuunnat, joihin soittaessa generoidaan hälytys tai jolloin tapaus ohjataan tarkempiin sääntöpohjaisiin analyyseihin. Listoja voidaan käyttää myös päinvastoin (ns. white list). Tällöin listoille määritellään esimerkiksi ne puhelusuunnat, jotka tiedetään ennalta lailliseksi, mutta jotka saattaisivat muuten generoida turhia hälytyksiä muiden analyysien seurauksena.

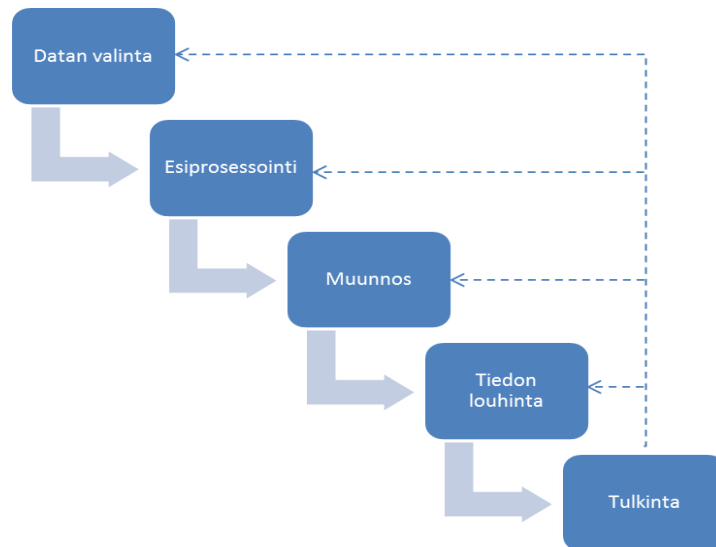
4.2 Tiedonlouhintaan perustuvat tekniikat

Tiedonlouhinta integroi tietokannat, tekoälyn, koneoppimisen, tilastot ja muut tekniikat ja teoriat saman termin alle (Gong 2011). Tiedonlouhintaan perustuvat tekniikat jaetaan kahteen osaan: ohjattuihin ja ohjaamattomiin tiedonlouhintatekniikoihin.

Ohjattujen ja ohjaamattomien tiedonlouhintatekniikoiden välillä on keskeinen ero. Ohjattuja, opettamiseen perustuvia tekniikoita käytetään luokitteluun (classification) ja ennustamiseen (prediction) kun taas ohjaamattomia tekniikoita käytetään silloin, kun ennustettavaa arvoa tai luokittelua ei etukäteen tiedetä (Shmueli, Patel & Bruce 2010, 15).

Tiedonlouhintaprosessin pääkohdat

Tiedonlouhintaprosessi jaetaan yleisesti kuvan 3 mukaisesti viiteen pääkohtaan. Itse tiedonlouhintaprosessin lisäksi oleellisia työvaiheita ovat tiedon esiprosessointi ja tulkinta. Tiedon esiprosessoinnissa lähtötiedoista pyritään poistamaan väärät ja epäjohdonmukaiset arvot sekä täydentämään mahdolliset tyhjät arvot (Zou, Sun, Yu & Liu 2012). Esiprosessoinnin yhteydessä tehdään myös alustavaa datan analysointia sopivien muuttujien löytämiseksi esimerkiksi laskemalla keskiarvot, keskihajonnat, jakaumat ja korrelaatiot normaaleihin tilastollisiin työkaluihin tukeutuen. Tulkinnalla tarkoitetaan tiedonlouhinnasta saatujen lopputulosten arvioimista ja käyttökelpoisuutta ongelman ratkaisemiseksi.



Kuva 3. Tiedonlouhintaprosessin pääkohdat (Kantardzic 2011, muokattu)

Oleellista koko tiedonlouhintaprosessin onnistumisen kannalta on louhintaprosessiin osallistuvien resurssien syvälinen tietämys aihealueesta ja asiaan liittyvistä liiketoimintaprosesseista (Zou ym. 2012).

Ohjatut tiedonlouhintatekniikat

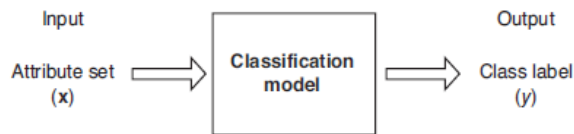
Ohjattuja tiedonlouhintatekniikoita käytetään tiedon luokittelussa ja ennustamisessa (Shmueli ym. 2010, 15). Luokittelussa yritetään ennustaa annetusta aineistosta ennalta määriteltyä luokkaa. Esimerkiksi puhelu palvelunumeroon voi olla laillista tai laitonta. Ennustaminen on samantapaista kuin luokittelu, paitsi että siinä yritetään ennustaa luokan sijasta numeerista muuttujan arvoa, esimerkiksi puhelun kestoa.

Tan, Steinbach ja Kumar (2006, 146) määrittelevät luokittelun seuraavasti:

Luokittelu on oppimistehtävä kohdefunktiolle f , jossa kuvataan jokainen attribuuttijoukko x yhdeksi ennalta määritellyksi luokaksi y .

Attribuutit (x) voivat olla joko diskreettejä tai jatkuvia (continuous) pois lukien luokka-attribuutti (y), jonka pitää aina olla diskreetti. Tämä ominaisuus erottaa luokittelun ennustavista malleista, joissa luokka-attribuutti voi olla jatkuva (Tan ym. 2006, 146). Ennustamisesta käytetään myös nimitystä regressio. Seuraavassa kuvassa 4 on kuvattu yleisellä tasolla luokittelun perusajatus, jossa luokittelumalli muodostaa annetusta syöt-

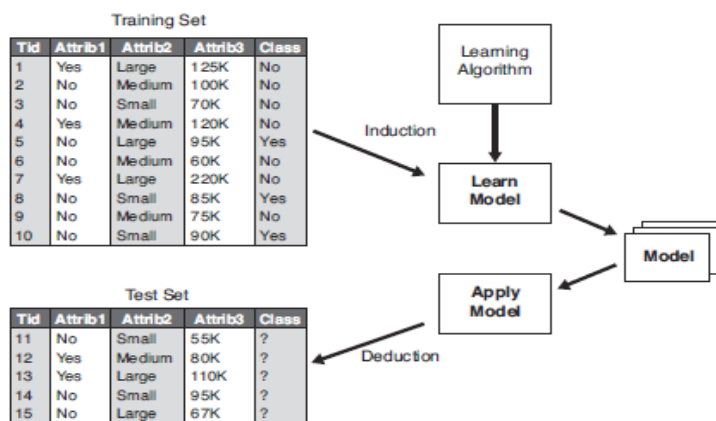
teestä ennalta määritellyn luokan y . Kuvassa kohdefunktiosta f käytetään epävirallista termiä luokittelumalli (Classification model) (Tan ym. 2006, 146).



Kuva 4. Luokittelun perusajatus (Tan ym. 2006, 146)

Luokittelumallien opettaminen tapahtuu erillisen koulutusaineiston avulla. Koulutusaineisto on kokoelma huolella valittuja tietueita, joissa jokainen tietue sisältää valittujen attribuuttien x lisäksi ennalta määritellyn luokan y (Tan ym. 2006). Televerkkojen väärinkäytöstapausten näkökulmasta attribuutit ovat siis fraud-indikaattoreita ja luokka on tieto, onko kyseinen attribuuttijoukko fraudia vai ei. Perusajatuksena siis on, että luokittelumalli pystyy koulutusaineiston perusteella muodostamaan relaation attribuuttien ja ennustettavan luokan välille ja käyttämään tämän jälkeen muodostamaansa mallia tuntemattoman aineiston luokittelussa. Tämän vuoksi oppimisalgoritmin päätavoite on rakentaa hyvän yleistämiskyvykkyyden (geneerisyyden) omaavia malleja (Tan ym. 2006).

Kuvassa 5 esitetään yleisellä tasolla luokittelumallin muodostuminen, jossa ennalta luokitellun koulutusaineiston avulla rakennetaan luokittelumallit, joiden suorituskykyä testataan erillistä testiaineistoa vasten, jossa luokittelu luokittelumallien näkökulmasta on tuntematon. Televerkkojen väärinkäytösten havainnoinnin näkökulmasta koulutus- ja testiaineiston tulee sisältää siis esimerkkejä sekä väärinkäytös- että normaaleista käyttötapauksista.



Kuva 5. Luokittelumallin muodostuminen (Tan ym. 2006, 148)

Luokittelumallin suorituskyvyn evaluointi perustuu testiaineistosta onnistuneesti ja epäonnistuneesti ennustettujen luokkien lukumäärään (Tan ym. 2006, 149). Käytännössä lähes kaikki tarkkuusmittarit johdetaan epäjärjestysmatriisista (Confusion matrix), jossa matriisin rivit ja sarakkeet vastaavat oikeaa ja ennustettua luokkaa ja päinvastoin (Shmueli ym. 2010). Taulukossa 1 havainnollistetaan epäjärjestysmatriisia binääri-luokittelun (0/1) näkökulmasta.

Taulukko 1. Epäjärjestysmatriisi binääri-luokittelun näkökulmasta

		Ennustettu luokka	
		Luokka = 1	Luokka = 0
Todellinen luokka	Luokka = 1	f_{11}	f_{10}
	Luokka = 0	f_{01}	f_{00}

Jokainen merkintä f_{ij} taulukossa 1 osoittaa luokasta i ennustettua luokkaa j . Esimerkiksi f_{01} on alkuperäisestä luokasta 0 virheellisesti luokkaan 1 ennustettujen tietueiden lukumäärä. Epäjärjestysmatriisin avulla voidaan laskea luokittelumallin sekä oikein ($f_{11} + f_{00}$) että väärin ($f_{10} + f_{01}$) ennustettujen luokkien lukumäärä (Tan ym. 2006, 149).

Vaikka epäjärjestysmatriisi tuottaakin kaiken tarvittavan informaation luokittelumallin suorituskyvyn määrittämiseksi, on monesti käyttökelpoisempaa summata edellä mainittu informaatio esimerkiksi eri mallien välisen suorituskyvyn mittaamisen helpottamiseksi. Tämä voidaan tehdä käyttämällä suorituskykymittaria *tarkkuus* (accuracy), joka voidaan määritellä seuraavasti (Tan ym. 2006, 149):

$$Tarkkuus = \frac{\text{Oikein ennustettujen lkm.}}{\text{Tietueiden kokonaismäärä}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

Luokittelumallin suorituskky voidaan vaihtoehtoisesti määritellä myös virheasteena (error rate) seuraavasti:

$$Virheaste = \frac{\text{Väärin ennustettujen lkm.}}{\text{Tietueiden kokonaismäärä}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

Useimmat luokittelualgoritmit hakevat malleja, jotka saavuttavat korkeimman tarkkuuden tai ekvivalentisti alhaisimman virheasteen, kun malleja testataan testiaineistoa vasten (Tan ym. 2006, 149).

Myös koulutusaineiston suorituskky voidaan mitata, mutta arvoja ei sellaisenaan pidä käyttää mallin hyvyyden mittareina johtuen ylioppimisen (overfitting) vaarasta. Ylioppimisella tarkoitetaan luokittelumallin yleistyskyyvyn heikkenemistä, ts. luokittelumalli oppii koulutusaineiston liian yksityiskohtaisesti. Näitä suorituskkymittareita voidaan kuitenkin käyttää testiaineistosta saatujen mallien suorituskkymittareiden vertailuun. Vaikka on oletettavaa, että testiaineiston luokittelun suorituskky on huonompi kuin koulutusaineiston, saattaa suuri ero testi- ja koulutusaineiston välillä indikoida luokittelumallin ylioppimisesta (Shmueli ym. 2010, 97).

Luokittelumalleja voidaan käyttää luokan ennustamisen lisäksi myös selittävänä työvälineenä erottamaan eri luokkien attribuutit toisistaan (Tan ym. 2006, 146). Hyvänä esimerkkinä tästä voisi olla uusien sääntöjen mallinnus sääntöpohjaisiin tekniikoihin.

Yleisimpiä ohjattuja oppimisalgoritmeja ovat päätöspuut (Decision Tree), sääntöpohjaiset luokittelijat (Rule-based Classifiers), neuroverkot (Neural Networks), tukivektorikoneet (Support Vector Machines), bayes-luokittelijat (naïve Bayes Classifiers) ja regressiot (Regression) (Tan ym. 2006, 148-205).

Ohjaamattomat tiedonlouhintatekniikat

Ohjaamattomia tiedonlouhintatekniikoita käytetään tapauksissa, joissa ennustettavaa arvoa tai luokittelua ei etukäteen tiedetä. Tekniikat voidaan jakaa kahteen ryhmään: klusterianalyysiin ja riippuvuussääntöihin (Shmueli ym. 2010).

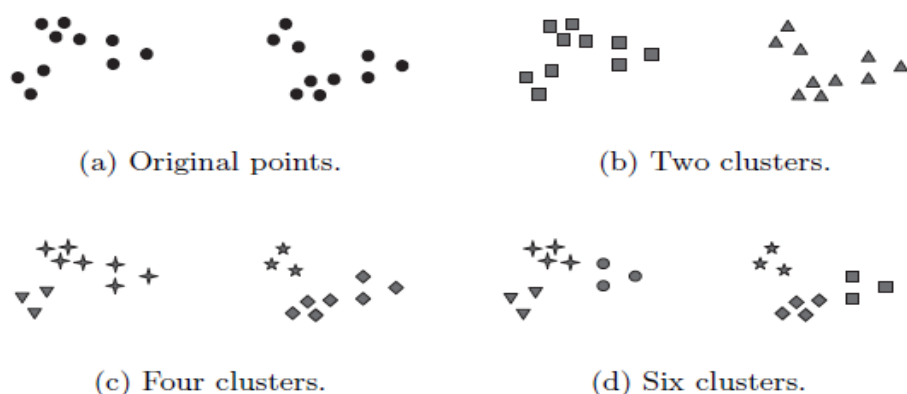
Klusterianalyysit

Luokat, tai käsitteellisesti merkityksellisten objektien ryhmät, joita yhdistävät tietyt ominaisuudet, näyttelevät tärkeää roolia, kun ihmiset analysoivat ja kuvailevat asioita. Ihmisille on luonteenomaista jakaa asioita ja ilmiöitä ryhmiin erilaisten kriteereiden perusteella. Esimerkiksi lapsi oppii jo varhaisessa lapsuudessaan erottamaan lehmän hevosesta tai löytämään eron kasvi- ja eläinkunnan välillä. Vastaavaa ryhmittelyä pyritään tekemään myös klusteroinnissa.

Klusterianalyysit pyrkivät ryhmittelemään tiedon objektit perustuen ainoastaan käsiteltävästä tiedosta löydettävään informaatioon, joka kuvailee objekteja ja niiden suhteita. Klusteroinnin tavoitteena on, että objektit ryhmän sisällä ovat mahdollisimman samantaisia keskenään ja samalla mahdollisimman erilaisia muiden ryhmien objekteista. Mitä suurempi samankaltaisuus ryhmän sisällä ja mitä suurempi eroavaisuus ryhmien välillä, sitä paremmin määritelty klusteri on. (Tan ym. 2006, 490.)

Klusterianalyyseja käytetään mm. tiedon pelkistämiseen, hypoteesien generoimiseen ja testaamiseen sekä joskus pelkästään aloituspisteenä jollekin muulle toiminnolle, esimerkiksi luokittelulle (Halkidi, Batistakis & Vazirgiannis 2001,109; Tan ym. 2006, 487). Televerkkojen väärinkäyttöshavainnoinnissa klusterianalyyseja voidaan käyttää havainnoimaan tilaajien oletuskäyttötymisen muutoksia ja ne sopivatkin erinomaisesti esimerkiksi tilaaja- ja murtautumisväärinkäytösten havainnointiin (Gosset ja Hyland 1999).

Alan kirjallisuudessa ei ole yhtenäistä linjaa klusterianalyysien luokitteluksi. Karkealla tasolla klusterianalyysit voidaan kuitenkin Halkidi ym. (2001) mukaan jakaa osittaviin (Partitioning), hierarkkisiin (Hierarchical), tiheysperustaisiin (Density-based) ja ristikoperustaisiin (Grid-based) menetelmiin. Jokaisella menetelmällä on omat hyvät ja huonot puolensa. Esimerkiksi tiheysperustaiset analyysit sietävät hyvin ”kohinaa” datassa, mutta tiheyden käsitteen määrittely vaatii vastaavasti tuntemusta käsiteltävästä aineistosta (Tan ym. 2006, 487–496). Klusterianalyysit tuottavat myös erilaisia lopputuloksia käytetystä analyysistä ja annetuista parametreista riippuen. Esimerkiksi seuraavassa kuvassa 6 on esitetty kolme eri tapaa klusteroida sama joukko pisteitä.



Kuva 6. Erilaisia tapoja klusteroida sama joukko pisteitä (Tam ym. 2006)

Ohjatuissa luokitteluissa luokittelumallin lopputulosten evaluointi on kiinteä osa luokittelumallin kehitystä ja niihin on olemassa yleisesti hyväksytyjä mittareita ja proseduurreja, kuten tarkkuus ja ristiintarkistus. Toisin kuin ohjatuissa luokittelumalleissa, klustereiden evaluointi ei ole kovinkaan kehittynyttä tai yleisesti käytettynä osana klusterianalyysia. Evaluointi pitäisi kuitenkin olla osa klusterianalyysia, sillä melkein jokainen algoritmi löytää klustereita annetusta tietojoukosta, vaikka tiedolla ei olisi luonnollista klusterirakennetta (Tan ym. 2006, 532).

Teodoridisin ja Koutroubasin (1999) mukaan on olemassa kolme lähestymistapaa klusterien kelpoisuuden tutkimiseen: ulkoisiin kriteereihin perustuva lähestymistapa, sisäisiin kriteereihin perustuva lähestymistapa ja suhteellisiin kriteereihin perustuva lähestymistapa (Halkidi ym. 2001, 123.)

Ulkoisiin kriteereihin perustuvassa lähestymistavassa klusterialgoritmin tuloksia verrataan ulkoiseen rakenteeseen, jonka rakenne on ennalta määritelty. Sisäisiin kriteereihin perustuvassa lähestymistavassa mitataan klusterialgoritmin tuloksia informaatiosta, joka esiintyy pelkästään käsiteltävässä tiedossa. Klustereiden hyvyttä mitataan kahden mittarin - koheesion ja erottelun - avulla. Koheesio mittaa, kuinka lähellä toisiinsa liittyvät objektit klusterissa ovat. Erottelu määrittää, kuinka erillään oleva tai hyvin eroteltu klusteri on muista klustereista. Suhteellisessa lähestymistavassa eri klustereita verrataan keskenään (Halkidi ym. 2001, 123; Tan ym. 2006, 535).

Riippuvuussäännöt

Riippuvuussääntöjä eli assosiaatiosääntöjä käytetään esimerkiksi kaupan alalla kertoamaan, kuinka kaupasta ostetut tuotteet liittyvät toisiinsa. Riippuvuussääntöjen avulla voitaisiin esimerkiksi todeta, että 90 prosenttia asiakkaista, jotka ostavat vaippoja ja maitoa, ostavat myös vauvanruokaa. Riippuvuussääntöjä kuvataan notaatiolla $\{vaippa, maito\} \rightarrow \{vauvanruoka\}$.

Riippuvuussäännöt ovat siis käyttökelpoisia tutkittaessa mielenkiintoisia piiloyhteyksiä suuresta datajoukosta. Selville saadut yhteydet voidaan esittää riippuvuussääntöjen muodossa tai joukkona usein toistuvia merkintöjä (Tan ym. 2006, 327).

Riippuvuussääntö on seuraamus lausekkeesta $X \rightarrow Y$, missä X ja Y ovat epäjatkuvia alkiojoukkoja. Riippuvuussäännön vahvuutta voidaan mitata termeillä kannatus (Support) ja luottamus (Confidence). Kannatus määrää, kuinka usein sääntö on käyttökelpoinen käsiteltävälle datajoukolle. Luottamus taas määrää, kuinka usein Y :n alkiot esiintyvät tapahtumassa, joka sisältää X :n (Tan ym. 2006, 329–330).

Riippuvuussääntöjen analyysialgoritmeilla on potentiaalia generoida todella suuria määriä hahmoja (pattern), joista moni saattaa olla aiottuun käyttötarkoitukseen vähemmän kiinnostava. Onkin tärkeätä luoda joukko hyväksi havaittuja kriteereitä, joiden avulla voidaan evaluoida riippuvuussääntöjen tuottamien hahmojen laatu. Ensimmäinen kriteerijoukko voidaan luoda tilastollisista argumenteista. Hahmot, jotka koskettavat molemminpuolista joukkoa itsenäisinä alkioina tai kattavat todella harvat tapahtumat, pyritään merkitsemään mielenkiinnottomaksi, koska ne saattavat tuottaa harhaanjohtavia yhteyksiä tiedossa (Tan ym. 2006, 370–371.)

Toiseksi kriteerijoukoksi voidaan luoda subjektiivisia argumentteja. Hahmoa harkitaan subjektiivisesti ei-kiinnostavaksi, jollei se paljasta odottamatonta tietoa datasta tai paljasta käyttökelpoista tietoa, josta voidaan johtaa tuottavaa toimintaa (Tan ym. 2006, 371). Esimerkiksi sääntö $\{leipä\} \rightarrow \{voi\}$ saattaa olla mielenkiinnoton, huolimatta sen korkeasta kannatus- ja luottamusarvoista, koska sääntö on liian ilmeinen.

Sekä objektiivisille että subjektiivisille kriteereille on kehitetty lukuisa joukko erilaisia mittareita (Tan ym. 2006).

Riippuvuussäännöt eivät ole televerkkojen väärinkäytösten havainnoinnissa yleisesti käytettyjä, vaan niitä käytetään enemmän esimerkiksi perinteiseen ostoskorianalyysiin. Telealalla riippuvuussääntöjä voitaisiin käyttää esimerkiksi sääntöjen tuottamiseen sääntöperusteisille havainnointityökaluille.

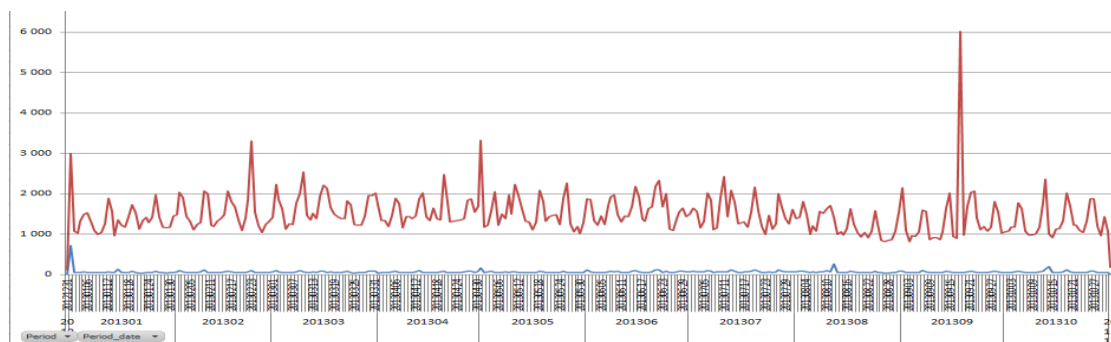
4.3 Muut havainnoimistekniikat

On arvioitu, että jopa 73 prosenttia kaikista telealan väärinkäytöksistä aiheutuisi operaattoreiden omista työntekijöistä (Beck Computer Systems 2003, Bihina Bellan ym. 2005b mukaan). Onkin tärkeitä auditoida säännöllisesti yrityksen prosessit ja järjestelmät mahdollisten heikkouksien löytämiseksi ja kuntoon saattamiseksi (Gosset & Hyland 1999). Auditointi voidaan tehdä joko sisäisesti tai ulkoisen toimijan toimesta. Esimerkiksi on hyvä varmistaa, että järjestelmiin tehdyt muutokset lokitetaan ja dokumentoidaan asianmukaisesti. Väärinkäytösriskkejä voidaan vähentää myös siten, että sama henkilö ei pysty tilaamaan ja tekemään liittymien provisiointia.

Eräs yleisesti käytössä oleva havainnoimistekniikka on tiedon visualisointi. Visualisoinnissa luotetaan ihmisen hahmontunnistuskäytön poikkeuksien havainnoinnissa. Visuaaliset toimintatavat ovat dynaamisia ja voidaan helposti sopeuttaa väärinkäyttäjien jatkuvasti muuttuviin tekniikkoihin (Kou ym, 2004).

Visualisoinnissa monimutkainen tieto muunnetaan selkeiksi hahmoiksi hyväksikäyttäen värejä, sijaintia, kokoa ja muita visuaalisia elementtejä (Sharma & Panigrahi, 2012). Visualisointia voidaan käyttää myös muiden havainnointitekniikoiden tukena esimerkiksi käyttökelpoisten muuttujien etsimisessä ja datan siivouksessa (Shmueli ym. 2010, 97).

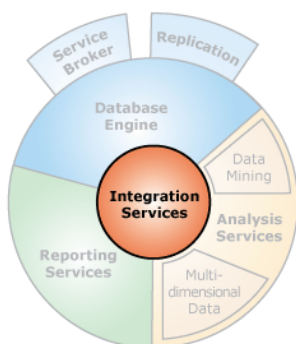
Yksinkertaisimmillaan visuaalinen havainnointi voi perustua esimerkiksi alla olevankaltaiseen diagrammiin. Kuvan 7 diagrammi kuvaa palvelunumeropalvelun päivittäisiä sykäysmäärien keskihajontaa ja niissä mahdollisesti esiintyviä poikkeamia.



Kuva 7. Esimerkki visuaalisesta havainnoinnista

5 SSIS-YMPÄRISTÖN ESITTELY

SSIS on lyhenne sanoista SQL Server Integration Services ja se on osa Microsoftin SQL Server palvelintuotetta. SQL Server 2008 R2 koostuu neljästä osasta: relaatiotietokantapalvelimesta (SQL Server), analyysipalvelimesta (Analysis Services, SSAS), raportointipalvelimesta (Reporting Services, SSRS) sekä integrointipalvelimesta (Integration Services, SSIS). Tuotteita ei voi hankkia irrallisina, vaan ne ovat aina osana SQL Serverin ohjelmistopaketteja. Ohjelmistopaketteja on erilaisia ja kaikki paketit eivät sisällä kaikkia toiminnallisuuksia (SQL Server R2 Books Online). Kuvassa 8 on esitetty SQL Serverin pääkomponentit.



Kuva 8. SQL Server ja sen pääkomponentit (SQL Server 2008 R2 Books Online)

Väärinkäytösten havainnoinnin kannalta erityisen mielenkiintoisia ovat SSAS ja SSIS. SSAS sisältää erillisen Data Mining -osion, joka tukee useita ohjattaviin ja ohjaamattomiin tiedonlouhintateknikoihin perustuvia algoritmeja (SQL Server 2008 R2 Books Online). SSAS:n avulla voidaan toteuttaa esimerkiksi opetettavia luokittelumalleja mm. tilaaja- ja murtautumisväärinkäytösten havainnointiin, klusterianalyysseja esimerkiksi poikkeamien (outliers) havainnointiin sekä erilaisia riippuvuussääntöihin perustuvia analyysseja. SSAS:n arkkitehtuuria ja toimintaa ei tässä opinnäytetyössä käsitellä.

SSIS on yritystason ETL-järjestelmä ja koostuu joukosta työkaluja, sovelluksia, palveluita ja rajapintoja, joiden avulla voidaan toteuttaa erittäin suorituskykyisiä ratkaisuja liike-elämän tarpeisiin (Knight, Veerman, Moss, Davis & Rock 2012, 1–15; Nanda 2008, 2–3). ETL on yleisesti käytössä olevan tiedonsiirtomalli, joka käsittää kolme vaihetta: tiedon poiminnan lähdejärjestelmästä (Extract), muunnoksen kohdejärjestelmän

vaatimaan muotoon (Transform) ja latauksen kohdejärjestelmään (Load) (Hovi, Hervonen & Koistinen 2009, 49).

Televerkkojen väärinkäytösten havainnoinnin näkökulmasta SSIS:n avulla voidaan rakentaa ETL-prosesseja mm. tiedon esiprosessointiin tai SSAS:n tiedonlouhintamallien prosessointiin ja päivittämiseen. Kuten luvussa 6 esitetään, SSIS:n avulla pystytään tekemään myös monipuolista tiedon poimintaa, käsittelyä ja työnkulkua sääntöpohjaisten väärinkäytösanalyysien mahdollistamiseksi.

SSIS:n arkkitehtuurissa työnkulku ja tiedon käsittelyyn liittyvät toiminnot ovat eriytetty toisistaan. SSIS on rakennettu käsittelemään työnkulkua ja tiedon käsittelyä tavalla, joka mahdollistaa erittäin korkean suorituskyvyn. SSIS:ssa on kaksi erillistä ajonaikaismoottoria, joista toinen on optimoitu vastaamaan työnkulusta ja muista ajonaikaispalvelusta ja toinen pelkästään tiedon käsittelyyn liittyvistä asioista. Tiedon käsittelystä vastaava ajonaikaismoottori (Data Flow Engine) käsittelee tietoa pelkästään muistissa, mikä tekee siitä äärimmäisen suorituskykyisen (Nanda 2008, 3). SSIS pitääkin hallussaan tiedonlatauksen epävirallista maailmanennätystä ladattuaan yli kaksi teratavua dataa yhdessä tunnissa (Knight ym. 2012, 5).

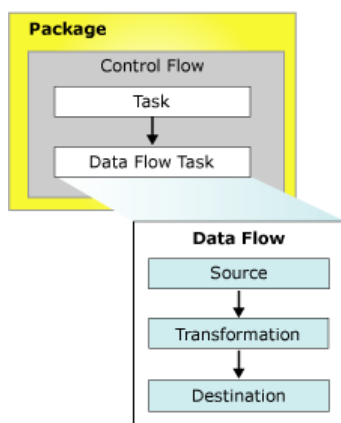
SSIS sisältää valmiiksi monipuoliset yhteydet eri tietolähteisiin, kattavat työnkulkuun liittyvät toiminnallisuudet, erilaiset toistorakenteet, virhekäsittelyt sekä lukuisan joukon valmiita komponentteja tiedon käsittelyyn ja muunnokseen. Lisäksi SSIS tarjoaa valmiit rajapinnat omien laajennusten lisäämiseksi (SQL Server R2 Books Online).

5.1 SSIS-pakettien rakenne

Paketti on SSIS:n tärkein objekti ja samalla aloituspiste halutun toiminteen rakentamiseksi (Haselden 2009, 98). Paketti voi olla yksinkertaisimmillaan esimerkiksi tiedoston siirtoa paikasta toiseen tai hyvinkin laaja ja monimutkainen yrityksen liiketoimintaprosesseja tukeva tehtäväkokonaisuus. Microsoftin mukaan paketti on kokoelma erilaisia yhteyksiä, Control ja Data Flow -elementtejä, tapahtumankäsittelijöitä, muuttujia ja konfiguraatioita, joka voidaan koostaa joko integrointipalvelimen mukana tulevilla graafisella työkalulla tai ohjelmallisesti.

Paketti rakennetaan lisäämällä siihen muita komponentteja, jotka voivat olla myös toisia paketteja. Fyysisesti yksittäinen paketti muodostuu DTSX-päätteisestä XML-tiedostosta. Yksittäisen paketin ”älyn” muodostaa Control Flow -elementti, johon määritellään paketissa suoritettavat tehtävät (Task) ja niiden väliset ehtoliitokset (Preference Constraint). Tiedon siirrosta ja transformaatiosta eri tietolähteiden välillä vastaa Data Flow -elementti yhdessä tietoyhteyksien (Connection Managers) kanssa. Pakettiin voidaan liittää lisäksi erilaisia muuttujia (Variables), tapahtumankäsittelijöitä (Event Handlers) ja lokipalveluita (Log Providers). Pakettien eri komponenttien toiminnallisuutta kuvataan seuraavissa luvuissa.

Seuraavassa kuvassa (kuva 9) on esitetty yksinkertainen paketti, jossa Control Flow -elementti sisältää kaksi erillistä tehtävää, joista toinen on Data Flow -elementti.

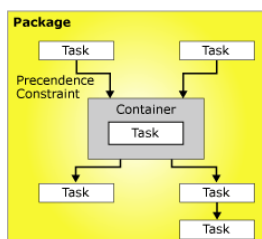


Kuva 9. SSIS-paketti (SQL Server 2008 R2 Books Online)

5.1.1 Control Flow

Paketti sisältää aina Control Flow -elementin sekä tarvittaessa yhden tai useamman Data Flow -elementin. Control Flow sisältää kolmea erityyppistä elementtiä: säiliöitä, jotka tuottavat paketin rakenteen, tehtäviä, jotka tuottavat halutun toiminnallisuuden sekä ehtoliitoksia, jotka yhdistävät säiliöt ja tehtävät tarkoituksenmukaiseen järjestykseen.

Seuraava kuvassa (kuva 10) on esitetty Control Flow -elementti, joka sisältää yhden säiliön ja kuusi tehtävää. Viisi tehtävää on määritelty pakettitasolle ja yksi säiliötasolle. Nuolet kuvassa kertovat suoritusjärjestyksen.



Kuva 10. Control Flow (SQL Server 2008 R2 Books Online)

Säiliöt (Containers)

Säiliöt ovat objekteja, jotka sisältävät muita säiliöitä tai Task-tehtäväkomponentteja muodostaen hierarkkisen äiti-lapsi -relaation. Säiliöt antavat paketille rakenteen ja tuottavat erilaisia palveluita tehtäväkomponenteille transaktiokäsittelyjen sekä muuttujien, lokitusten, suoritusten ja tapahtumankäsittelijöiden näkyvyysmäärittelyjen muodossa (Haselden 2009, 99).

Säiliöitä käytetään paketeissa seuraaviin tarkoituksiin (SQL Server R2 Books Online):

- kokoelmien läpikäyntiin, esimerkiksi lukemaan tietyn hakemiston kaikki tiedostot
- toistotehtäviin kunnes määritelty ehto täyttyy
- ryhmittelemään tehtäväkomponentteja ja säiliöitä, joiden suorituksen pitää onnistua tai epäonnistua kokonaisuutena

Säiliöitä voidaan suorittaa samalla tavoin kuin tehtäväkomponentteja. Säiliö eroaa kuitenkin tehtäväkomponentin suorituksesta siten, että kun säiliö suoritetaan, suoritetaan

myös kaikki säiliöön liitettyt objektit. Säiliöillä on myös tiettyjä ominaisuuksia, joilla pystytään kontrolloimaan niiden suunnittelun- ja ajonaikaista käyttäytymistä. Niille voidaan määritellä esimerkiksi pysäytyspisteitä (Breakpoints) tai ne voidaan tarvittaessa kytkeä pois päältä tarvitsematta poistaa niitä itse paketista. Säiliöiden avulla voidaan lisäksi vaikuttaa transaktioihin. Jos säiliö konfiguroidaan tukemaan transaktiota, aloitetaan transaktio siitä hetkestä, kun säiliön suoritus alkaa ja lopetetaan siihen hetkeen, kun säiliön suoritus päättyy. Se hyväksytäänkö vai hylätäänkö transaktio, riippuu säiliön suorituksen onnistumisesta. Taulukossa 2 on kuvattu eri säiliötyypit sekä niiden tarkoitus.

Taulukko 2. Säiliötyypit ja niiden tarkoitus

Säiliö	Kuvaus
Foreach Loop Container	Suorittaa tehtäväkomponentteja ja säiliöitä toistuvasti käyttäen saamaansa arvojoukkoa, esimerkiksi tiedostolistaa.
For Loop Container	Suorittaa tehtäväkomponentteja ja säiliöitä toistuvasti annettuun ehtoon perustuen.
Sequence Container	Ryhmittelee tehtäväkomponentit ja säiliöt omaksi osajoukoksi.
Task Host Container	Tuottaa palvelut yksittäiselle tehtäväkomponentille. Jokaisella tehtäväkomponentilla on oletuksena säiliö, vaikka sitä ei erikseen näytetä.

Tehtäväkomponentit (Tasks)

Task-tehtäväkomponentit ovat yksittäisiä tehtäviä, jotka suoritetaan paketin Control Flow -elementissä (SQL Server 2008 R2 Books Online). Tehtävät tuottavat toiminnallisuuden paketille hieman samalla tavalla kuin metodit tuottavat jonkin tietyn toiminnallisuuden ohjelmointikielessä. Ohjelmoinnin sijaan SSIS:ssä valitaan haluttu toiminnallisuus raahaamalla tarvittava tehtäväkomponentti Control Flow -elementtiin ja konfiguroimalla se tarkoituksenmukaisella tavalla (Haselden 2009, 91). Paketti sisältää aina vähintään yhden tehtäväkomponentin. Mikäli paketti sisältää useampia tehtäviä, pitää ne yhdistää ja järjestää keskenään käyttäen ehtoliitoksia (Constraint Precedences). SSIS:n mukana tulee lukuisia valmiita tehtäväkomponentteja eri käyttötarkoituksiin. Tarvittaes-

sa tehtäväkomponentteja voidaan tehdä myös itse millä tahansa CLR-yhteensopivalla ohjelmointikielellä (SQL Server 2008 R2 Books Online). Tehtäväkomponentit voidaan Microsoftin mukaan ryhmitellä taulukon 3 osa-kokonaisuuksiin:

Taulukko 3. Control Flown tehtäväkomponenttien jaottelu

Tehtäväkomponentin tyyppi	Tarkoitus
Data Flow Task	Tehtäväkomponentti datan poimintaan, transformointiin ja lataukseen eri tietolähteiden välillä.
Data Preparation Tasks	Tehtäväkomponentteja käytetään tiedostojen ja hakemistojen kopioimiseen, tiedostojen ja tiedon lataukseen, XML-dokumenttien käsittelyyn ja datan siivoukseen.
Workflow Tasks	Tehtäväkomponentteja käytetään kommunikoimaan muiden prosessien kanssa esimerkiksi ajamaan SSIS paketteja, ohjelmia tai komentojonoja, lähettämään ja vastaanottamaan viestejä SSIS-pakettien välillä sekä lähettämään sähköpostia.
SQL Server Tasks	Tietokantaobjektien ja datan käsittelyyn liittyvät tehtäväkomponentit.
Scripting Tasks	Tehtäväkomponenteilla laajennetaan pakettien toiminnallisuutta. Komponentit voivat sisältää skriptien lisäksi mitä tahansa NET-ohjelmointikieltä.
Analysis Services Tasks	Tehtäväkomponenteilla ylläpidetään SSAS:n objekteja esimerkiksi kuutioita ja tiedonlouhintaan perustuvia ohjattuja ja ohjaamattomia malleja.
Maintenance Tasks	Tehtäväkomponenteilla suoritetaan erilaisia järjestelmävalvojatason tehtäviä kuten tietokantojen varmuuskopiointeja ja uudelleenindeksointeja.
Backward Compatibility Tasks	SSIS:n aikaisempiin versioihin liittyviä tehtäväkomponentteja taaksepäin yhteensopivuuden varmistamiseksi.

Tärkein tehtäväkomponentti on Data Flow Task, joka poimii erillisessä datavuossa valitusta tietolähteestä tietoa, tekee tiedolle tarvittavat transformaatiot ja lataa muunnetun tiedon haluttuun tietolähteeseen (SQL Server R2 Books Online; Haselden 2009, 93).

Ehtoliitokset (Precedence Constraints)

Ehtoliitokset voidaan ajatella eräänlaisiksi SSIS:n liikennevaloiksi. Niiden avulla määrätään, mitkä komponentit suoritetaan ja missä järjestyksessä. Suoritettavat komponentit voivat olla säiliöitä, tehtäväkomponentteja tai tapahtumankäsittelijöitä (SQL Server R2 Books Online).

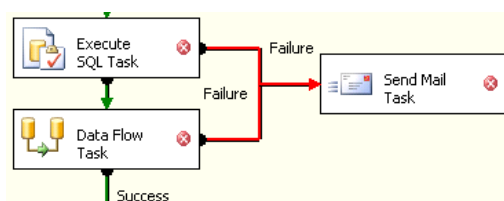
Ehtoliitoksessa kaksi suoritettavaa komponenttia linkitetään toisiinsa siten, että toinen komponenteista määritellään ns. etusijalla olevaksi komponentiksi ja toinen ns. rajoittavaksi komponentiksi. Etusijalla oleva komponentti suoritetaan aina ennen rajoittavaa komponenttia. Rajoittavan komponentin suoritus riippuu ehtoliitokseen määritellystä ehdosta. Ehto voi olla etusijalla ajettavan komponentin ajon lopputulos, SSIS:n lauseke tai lausekkeen ja etusijalla ajettavan komponentin ajon lopputuloksen yhdistelmä (SQL Server R2 Books Online).

SSIS-lausekkeet voivat sisältää funktioita, operaattoreita ja muuttujia. Etusijalla olevan komponentin ajon lopputulos on taas jokin seuraavassa taulukossa 4 kuvatusista arvoista.

Taulukko 4. Ehtoliitoksen etusijalla olevan komponentin paluuarvot

Arvo	Selitys
Completion	Etusijalla oleva komponentti on suoritettu
Success	Etusijalla oleva komponentti on suoritettu onnistuneesti
Failure	Etusijalla olevan komponentin suoritus on epäonnistunut

Seuraavassa kuvassa (kuva 11) esitetään kolme ehtoliitoksilla toisiinsa linkitettyä tehtäväkomponenttia. Suoritettavaan tehtäväkomponenttiin voidaan liittää useampia ehtoliitoksia, esimerkiksi virhetilanteiden käsittelyyn. Komponenttien keskinäistä suoritussyajjstystä kuvataan nuolella.

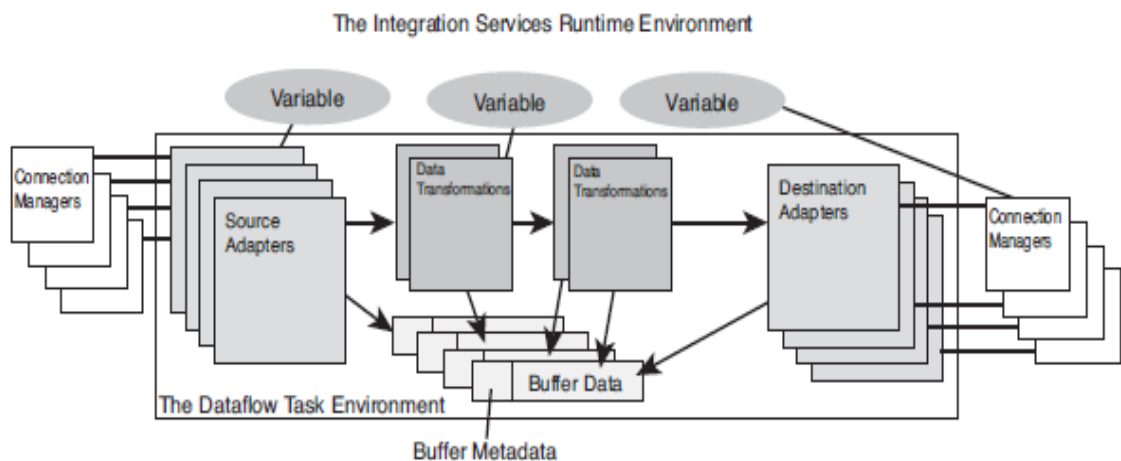


Kuva 11. Ehtoliitokset, virneenhallinta ja suoritusjärjestys

5.1.2 Data Flow

Data Flow Task on SSIS:n tärkein komponentti ja sen avulla tehdään käytännöllisesti katsoen lähes kaikki tiedon prosessointi (Haselden 2009, 93). Data Flow on myös äärimmäisen tehokas johtuen komponentin tavasta käsitellä tietoa. Data Flow käyttää erityistä puskuriorientoitunutta arkkitehtuuria, jossa tietoa kuljetetaan ja käsitellään muistissa läpi koko käsittelyketjun (Nanda 2008, 4).

Data Flow Task sisältää kolmenlaisia komponentteja: luettavia tietolähteitä (Source Adapters), transformaatioita sekä kirjoitettavia tietolähteitä (Destination Adapters). Komponentit yhdistetään Data Flowssa toisiinsa poluilla (Paths). Polut muistuttavat ulkonäöllisesti ehtoliitoksia, mutta niiden tarkoitus on täysin eri. Ehtoliitokset kontrolloivat tehtävien suoritusta kun taas polut kertovat, mistä mihin tietoa kuljetetaan (SQL Server R2 Books Online). Yksinkertaistaen voidaan todeta, että Data Flow:n tarkoituksena on poimia erilaisista tietolähteistä tarvittava tieto, muokata sitä tarkoituksenmukaisilla tavoilla ja tallentaa se tämän jälkeen valittuihin tietolähteisiin. Seuraavassa kuvassa on esitetty Data Flow Task ja siihen liittyvät komponentit.



Kuva 12. Data Flow ja siihen liittyvät komponentit (Haselden 2009, 94)

Lähdeadapterit (Source Adapters)

Lähdeadaptereiden tarkoituksena on hankkia tietolähteiden data muiden Data Flow -komponenttien käytettäväksi. Lähdeadapterilla on tyypillisesti yksi säännönmukainen ulostulopuskuri (output), johon luettava tietolähde on lisännyt tietolähteestä saamiaan sarakkeita sille annettujen määritysten mukaisesti. Lähdeadapterilla on tyypillisesti myös virheulostulo (error output) mahdollisten virhetilanteiden varalta. Virheulostulo on kahta virheinformaatiota sisältävää saraketta lukuun ottamatta identtinen säännönmukaisen ulostulon kanssa.

Kohdeadapterit (Sink Adapters)

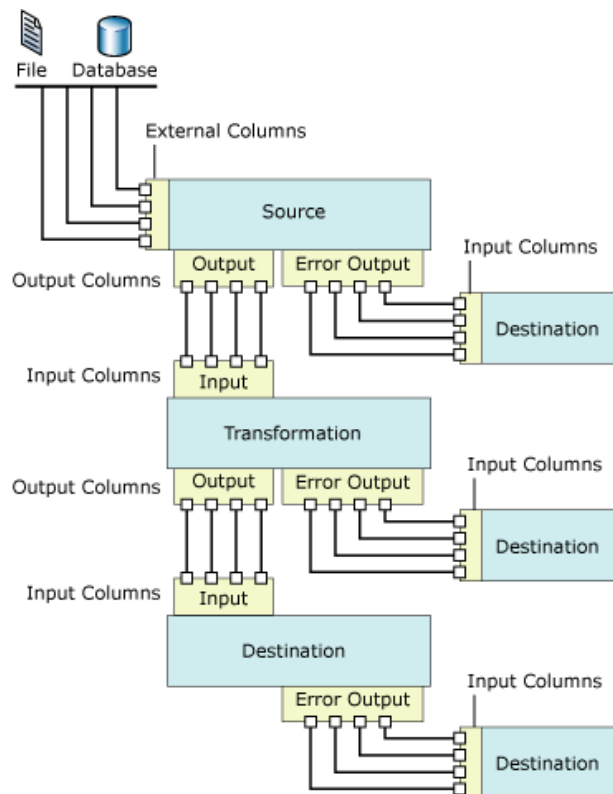
Kohdeadapteri on lähdeadapterin vastakohta ja sen tarkoituksena on kirjoittaa Data Flow -komponenteilta saatua tietoa määriteltyyn tietolähteeseen. Kohdeadapterilla on vähintään yksi sisääntulopuskuri (input), joka sisältää siihen kytketyn Data Flow -komponentin antamat sarakkeet. Lisäksi kohdeadapterilla on tyypillisesti myös virheulostulo mahdollisten virhetilanteiden varalta. Kohde- ja lähdeadapterit eivät itse ota yhteyttä tietolähteisiin, vaan yhteyden muodostukseen käytetään erillisiä Connection Managers -komponentteja. Connection Managers -komponenteista on kerrottu enemmän alaluvussa 5.1.3.

Transformaatiokomponentit

Transformaatiokomponentti muuntaa käsiteltävää tietoa. Muunnos voi olla esimerkiksi tiedon päivittäminen tai yhdistäminen, uusien sarakkeiden lisääminen tai vaikkapa summaan logiikkaan perustuva luokittelu. SSIS:n mukana tulee lukuisia valmiita transformaatiokomponentteja ja tarvittaessa komponentteja voidaan tehdä myös itse. Transformaatiokomponenteilla on aina vähintään yksi ulos- ja sisääntulopuskuri sekä tarvittaessa myös virheulostulopuskuri mahdollisten virhetilanteiden varalle. Tyypillinen tilanne virheulostulopuskurin käytölle on esimerkiksi Lookup-transformaatio, jossa etsitään halutulla sarakeella tai sarakkeiden yhdistelmällä tietoa toisesta taulusta pyrkien täydentämään lähtötietoa tämän perusteella. Mikäli arvoa ei saada, voidaan nämä rivit ohjata virheulostulopuskuriin tarkoituksenmukaisia käsittelyjä varten (SQL Server R2 Books Online).

Komponentit yhdistetään toisiinsa Data Flow:ssa siten, että yhdistettävän komponentin ulostulopuskuri liitetään toisen yhdistettävän komponentin sisääntulopuskuriin. Tämän jälkeen toisen komponentin sisääntulopuskurilla on käytössään kaikki toisen komponentin ulostulopuskurin sarakkeet.

Kuvassa 13 esitetään Data Flow, johon on määritelty sekä lähde- että kohdekomponentit sekä yksi transformaatioelementti. Lähdekomponentin ulostulo on yhdistetty transformaatiokomponentin sisääntuloon, samoin kuin transformaatiokomponentin ulostulo on yhdistetty kohdekomponentin sisääntuloon. Lisäksi jokaiselle komponentille on määritelty myös virheulostulo.



Kuva 13. Data Flow -komponenttien liitynnät (SQL Server 2008 R2 Books Online)

5.1.3 Connection Managers

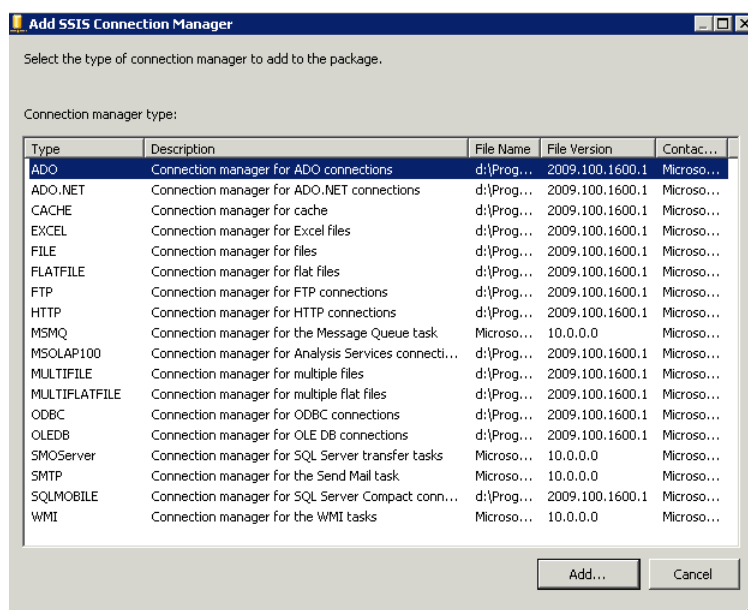
Connection Managerit ovat SSIS:n objekteja, jotka tuottavat paketin muille komponenteille linkin paketin ulkopuoliseen tietolähteeseen, esimerkiksi tietokantaan. Connection Managerit ylläpitävät yhteyttä luomalla yhteyden määritellyyn tietolähteeseen ja palauttavat yhteyden sitä pyytäneelle komponentille.

Connection Managerin palauttama paluuarvo tai objektityyppi riippuu kytketystä tietolähteestä. Esimerkiksi File Connection Manager palauttaa yksinkertaisen merkkijonon kun taas OLEDB Connection Manager palauttaa osoittimen OLEDB istunnon objektiin. Kutsuvan komponentin pitääkin tietää, kuinka käyttää saamaansa vastetta (Haselden 2009, 222).

Connection Managerit voidaan kategorisoida kolmeen luokkaan:

- tietokantayhteyksiin kuten ODBC, OLEDB ja ADO.NET
- verkkoyhteyksiin kuten FTP, HTTP ja Web Services
- tiedostojärjestelmäresursseihin

Seuraavassa kuvassa (kuva 14) voidaan nähdä SSIS:n mukaan tulevat Connection Managerit.

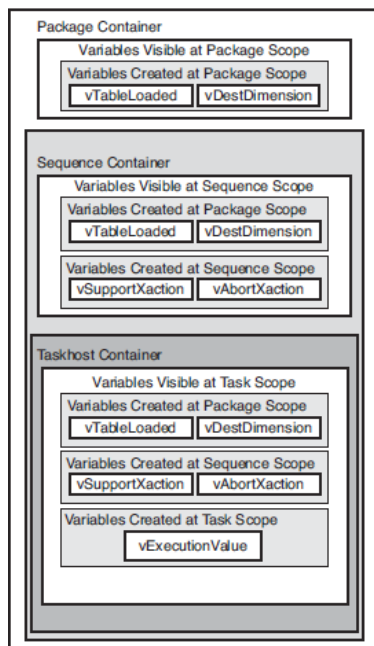


Kuva 14. SSIS:n mukana tulevat Connection Managerit

5.1.4 Variables

Kuten missä tahansa ohjelmoinnissa, kovakoodausta tulee välttää myös SSIS:n pakettien ja niiden objektien määrittelyssä. SSIS:ssa voidaan muiden ohjelmointikielten tavoin käyttää muuttujia. Muuttujat näyttelevätkin tärkeätä roolia SSIS:ssa, sillä niiden avulla paketin objektit pystyvät kommunikoimaan keskenään. Itse määriteltyjen muuttujien lisäksi SSIS tarjoaa paketin ja sen objektien käyttöön systeemimuuttujia, jotka luodaan ajonaikana tuottamaan ajonaikaista tietoa, jota muuten ei olisi saatavilla.

Muuttujilla on näkyvyysalue joka riippuu tasosta, jolle muuttuja määritellään. Korkein taso on paketti. Mikäli muuttuja määritellään pakettitasolle, näkyy muuttuja paketin kaikille objekteille. Mikäli muuttuja määritellään esimerkiksi säiliöön, näkyy muuttuja säiliön kaikille objekteille, mutta ei säiliön ulkopuolisille objekteille. Kuva 15 havainnollistaa muuttujien näkyvyysalueita.

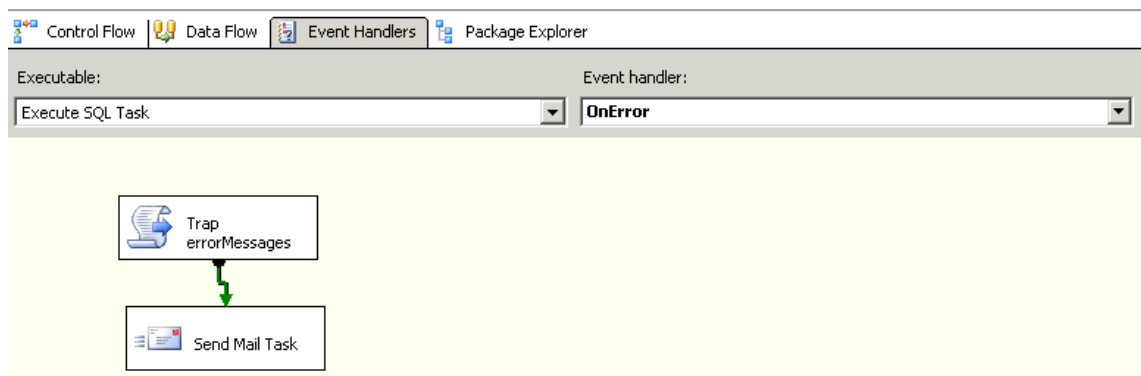


Kuva 15. Muuttujien näkyvyysalueet (Haselden 2009, 142)

5.1.5 Event Handlers

Kun pakettia suoritetaan, paketti ja pakettiin liitetyt tehtävä- ja muut komponentit laukeavat tapahtumia. Esimerkiksi OnError-tapahtuma laukeaa aina, kun virhe tapahtuu. Pakettiin voidaan määritellä tapahtumankäsittelijöitä (Event Handlers), jotka reagoivat näihin tapahtumiin halutulla tavalla, kun paketti laukaisee tapahtuman. Tapahtumankäsittelijät ovat paketissa olevia säiliöitä, jotka sisältävät kukin oman Control Flow:n. Säiliöihin voidaan määritellä tehtäväkomponentteja ja muita säiliöitä samalla tavalla kuin paketin normaalissa Control Flow:ssa (Knight ym. 2012, 557-558).

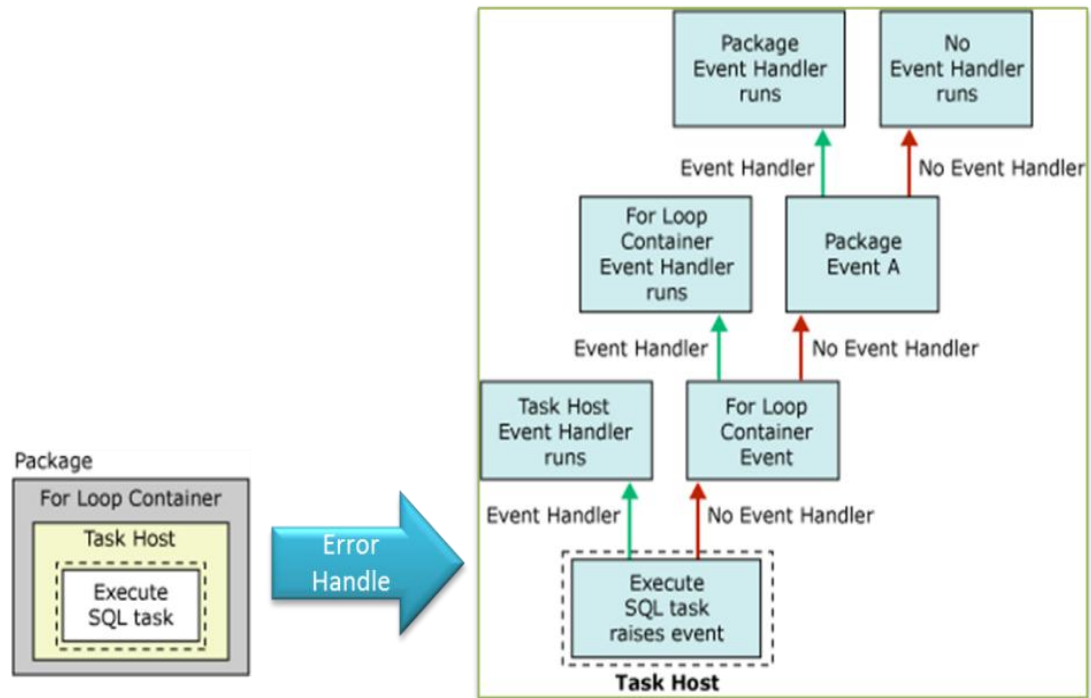
Seuraavassa kuvassa (kuva 16) on liitetty tapahtumankäsittelijä paketin Execute SQL Task -tehtäväkomponentin OnError-tapahtumaan. Kyseinen tapahtumankäsittelijä suoritetaan, jos Execute SQL Task -tehtäväkomponentin suorituksessa tapahtuu virhe. Tapahtumankäsittelijään on määritelty kaksi tehtäväkomponenttia ja niiden väliin ehtoliitos.



Kuva 16. OnError-tapahtumankäsittelijään liitetty työnkulku

Jos tapahtumalle ei ole määritelty tapahtumankäsittelijää, tapahtuma laukeaa ylemmällä tasolla paketin hierarkiassa jatkaen kulkuaan ylöspäin, kunnes löydetään tapahtumaan kiinnitetty tapahtumankäsittelijä tai tapahtuma saavuttaa paketin ylätasoa (SQL Server R2 Books Online).

Seuraavassa kuvassa (kuva 17) yksinkertaiseen pakettiin on määritelty For Loop -säiliö ja säiliöön Execute SQL Task -tehtäväkomponentti. Jos virhe tapahtuu Execute SQL Task -tehtäväkomponenttia suoritettaessa, kutsutaan OnError-tapahtumankäsittelijää alhaalta ylöspäin kuvan mukaisesti, kunnes tapahtumakäsittelijä löydetään tai tapahtuma saavuttaa paketin ylitason.

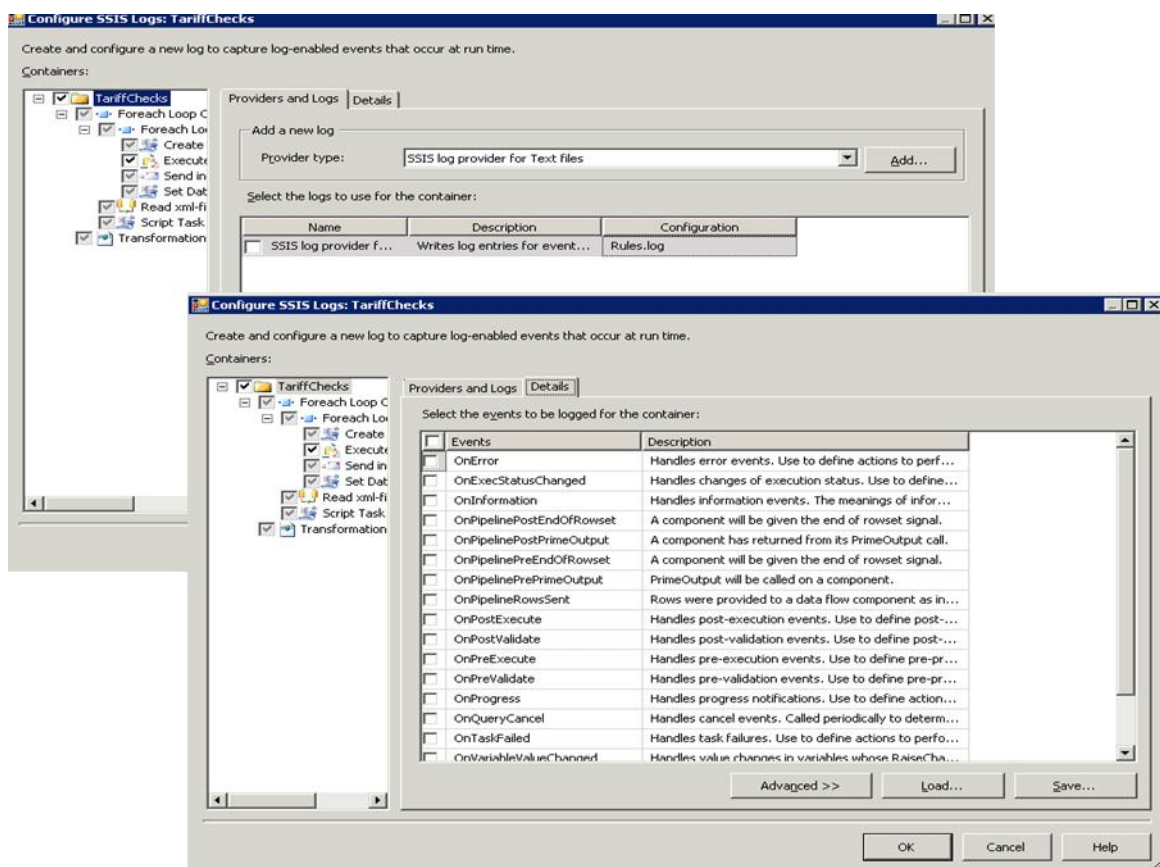


Kuva 17. Tapahtuman käsittelyjärjestys SSIS:ssa (SQL Server R2 Books Online)

5.1.6 Log Providers

SSIS:n mukana tulee viisi erilaista lokikomponenttia (Log Provider), jotka vastaanottavat SSIS:n muodostamia lokikirjauksia ja kirjoittavat ne valitusta lokikomponentista riippuen joko tiedostoon, tietokantaan tai Windowsin tapahtumalokiin. Lokikirjauksia voidaan kirjoittaa samanaikaisesti useaan eri paikkaan ja lokikirjauksille voidaan asettaa suodattimia eliminoimaan ns. ”turhia” viestejä. Sekä suodatus että lokikirjausten kohde voidaan asettaa paketin sisällä tarvittaessa komponenttikohtaisesti, joskin lokikomponentit pitää aina määritellä pakettitasolla (Haselden 2009, 295-311).

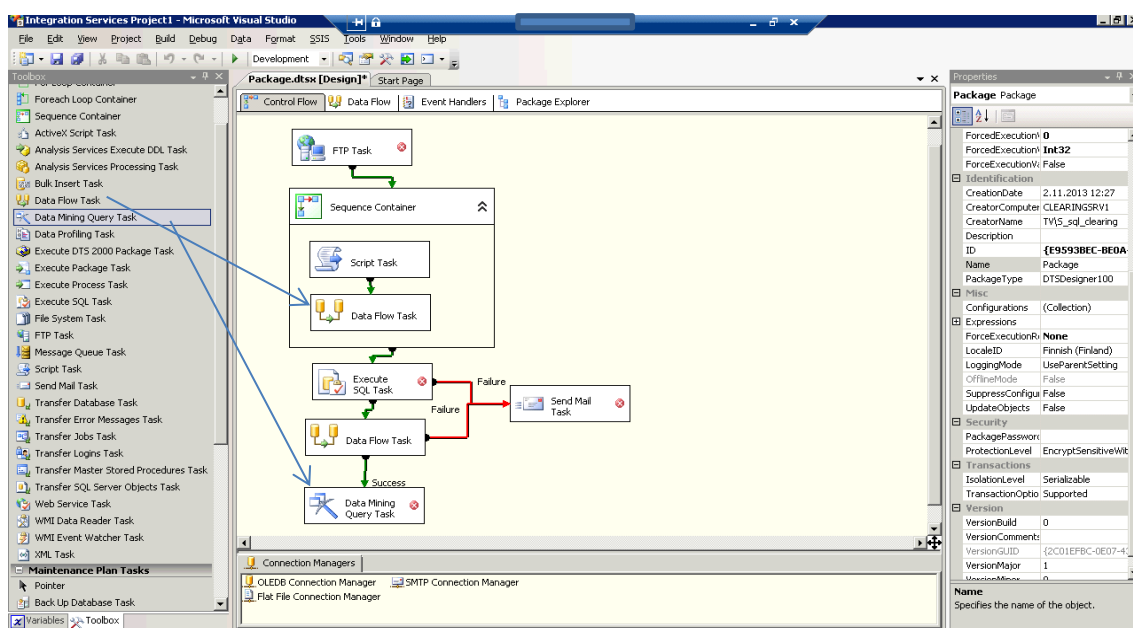
SSIS:n lokikomponentit tarvitsevat tarkoituksenmukaisen Connection Managerin pystyäkseen kirjoittamaan lokikirjauksia tietolähteeseen. Seuraavassa kuvassa 18 esitetään tekstitiedostoon lokikirjauksia kirjoittavan lokikomponentin konfigurointidialogit yksittäiselle tehtäväkomponentille määriteltynä. Connection Manager on määritelty kuvassa kohtaan Configuration. Details-välilehti sisältää tapahtumat, joista lokikirjauksia voidaan valita eli suodattaa.



Kuva 18. Log Providerin konfigurointi

5.2 SSIS-pakettien muodostaminen

SSIS Designer on graafinen työkalu SSIS-pakettien luomiseen ja ylläpitoon. SSIS Designer ei ole itsenäinen ohjelma, vaan yksi Business Intelligence Development Studio projektiksi, joka tulee SSIS:n mukana. Business Intelligence Development Studio (BIDS) on Microsoftin Visual Studio -ohjelmistotuotteen laajennus SQL Server Intelligence -sovellusten kehitykseen (SQL Server R2 Books Online).



Kuva 19. SSIS Designer

SSIS Designer sisältää kaikki paketin rakentamiseen tarvittavat komponentit, jotka ovat selitetty aluvuossa 5.1. Käyttöliittymässä Control Flow, Data Flow ja tapahtumankäsittelijät on eriytetty omille välilehdilleen. Pakettien rakentaminen tapahtuu ”Drag-and-Drop” -tyyppisesti raahaamalla välilehdille komponentteja ja luomalla niiden välille haluttu suoritusjärjestys ehtoliitoksien avulla. SSIS Designer sisältää myös työkalut pakettien debuggaukseen, pakettien jakeluun ja konfigurointiin.

SSIS:n mukana tulee myös Import and Export Wizard -niminen työkalu, jolla voidaan tehdä yksinkertaisia tiedon lataus-, muunnos- ja tallennustoimenpiteitä eri tietolähteiden välillä ilman ohjelmointia tai SSIS Designerin käyttöä. Työkalulla tehdyn tiedonsiirron voi halutessaan tallentaa myös SSIS-paketiksi (Knight ym. 2012, 19).

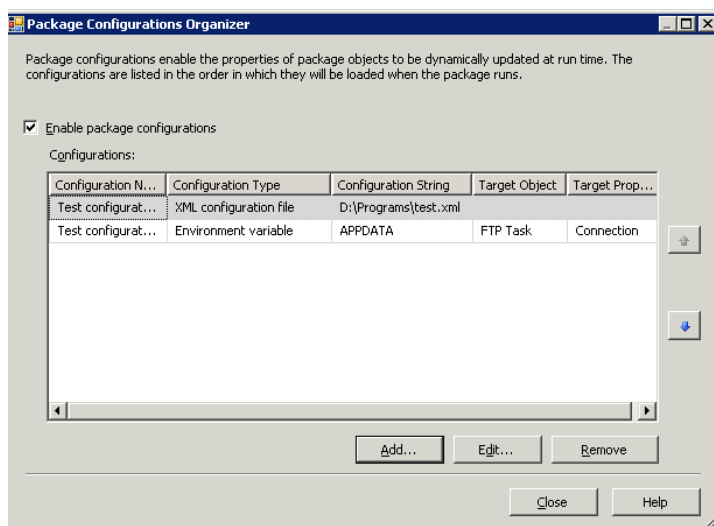
5.3 SSIS-pakettien konfigurointi ja jakelu

Konfigurointi

Valmiiden SSIS-pakettien komponenttien arvoja voidaan muuttaa suorituksen aikana käyttäen erilaisia konfiguraatioita. SSIS tukee useaa eri tapaa tallentaa paketin konfiguraatioita. Konfiguraatiot voidaan tallentaa joko XML-tiedostoon, SQL Serverin tietokantaan, rekisteriin, ympäristömuuttujiin tai toisen paketin muuttujiin. Jokaisella konfiguraatiolla on ominaisuus/arvo -yhdistelmä, jonka kautta haluttu arvo asetetaan (SQL Server R2 Books Online).

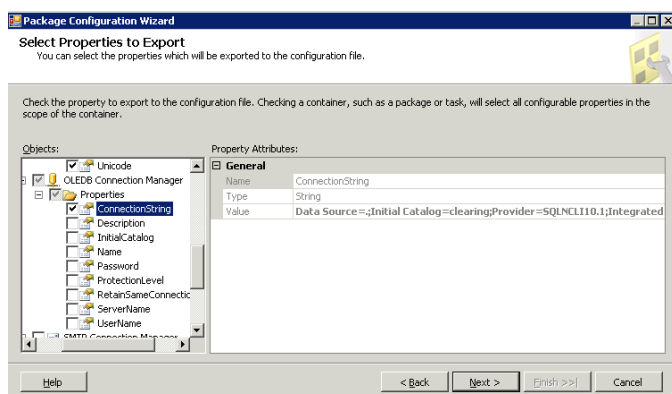
Kuten todettu, kaikilla paketin konfiguraatioilla on kaksi keskeistä elementtiä: paketin polku eli osoite, missä objekti ja sen ominaisuus paketissa sijaitsevat sekä kyseisen ominaisuuden arvo. Koska paketti on hierarkkinen kokoelma erilaisia objekteja, on paketin polkua käyttämällä mahdollista viitata yksikäsitteisesti melkein mihin tahansa ominaisuuteen paketissa (Haselden 2009, 525-528). Esimerkiksi paketitason muuttujan x polku paketissa voisi olla seuraava: \Package.Variables[x].value.

Konfiguraatioita voidaan luoda SSIS Designerin mukana tulevan Package Configurations Organizer -työkalun avulla. Pakettiin voidaan luoda useaan eri paikkaan tallennettuja konfiguraatioita, jotka ladataan Organizeriin määritellyssä järjestyksessä. Seuraavassa kuvassa (kuva 20) on määritelty paketille kaksi konfiguraatiota, joista ensimmäinen on tallennettu XML-tiedostoon ja toinen ympäristömuuttujaan.



Kuva 20. Packet Configuration Organizer -työkalulla määritellyt konfiguraatiot

Packet Configuration Organizer -työkalun voidaan asettaa paketin objektien ominaisuuksien arvoja varsin helppokäyttöisesti, kuten kuvasta 21 on havaittavissa.

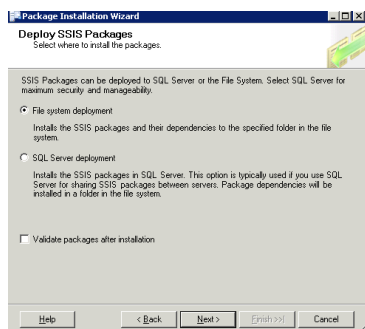


Kuva 21. Paketin objektin ominaisuuden arvon asettaminen

Pakettien jakelu

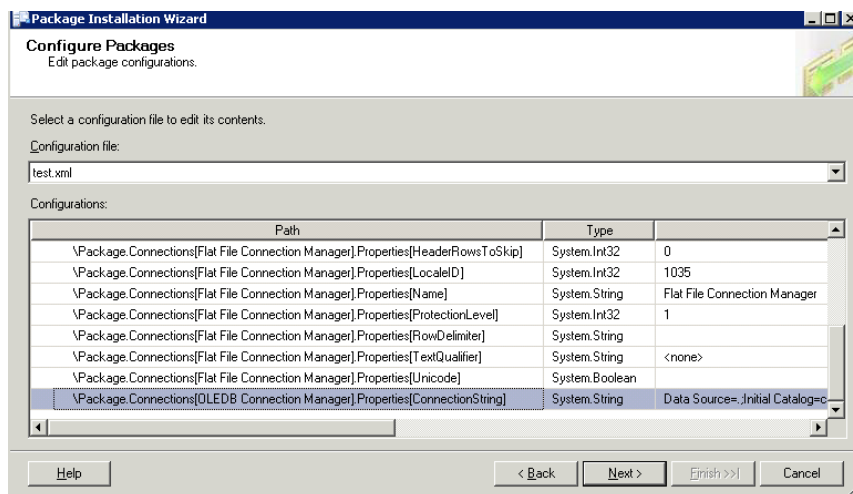
Tyypillisesti pakettien kehitys ja testaus tehdään erillisessä kehitysympäristössä, josta valmiit ja toimivaksi todetut paketit asennetaan tuotantoympäristöön. Pakettien asennus voidaan tehdä joko SQL Serveriin tai tiedostojärjestelmään. Jos paketit asennetaan SQL Serverille, pitää erikseen määritellä konfiguraatiolle ja muille pakettiin mahdollisesti liittyville tiedostoille tarkoituksenmukainen hakemisto. Mikäli paketit asennetaan tiedostojärjestelmään, voidaan siihen liittyvät tiedostot asentaa samaan hakemistoon itse paketin kanssa (SQL Server R2 Books Online).

Pakettien asennus voidaan tehdä manuaalisesti esimerkiksi kopioimalla paketti ja siihen liittyvät tiedostot tarkoituksenmukaiseen hakemistoon tai käyttämällä SSIS:n mukana tulevaa Deployment Utility -työkalua. Kun työkalu käynnistetään, esiin tulee kuvassa 22 esitetty dialogi:



Kuva 22. Deployment Utility -työkalu

Deployment Utility -työkalun avulla voidaan paketti ja siihen liittyvät tiedostot jakaa joko tiedostojärjestelmään tai SQL Serverin tietokantaan. Mikäli pakettiin on määritelty konfiguraatioita, työkalu antaa jakelun yhteydessä määrittää konfiguraatioihin halutut arvot. Seuraavassa kuvassa paketin konfiguraatiot on tallennettu XML-tiedostoon. Paketin jakelun yhteydessä työkalu mahdollistaa paketin ominaisuuksien arvojen muuttamisen.



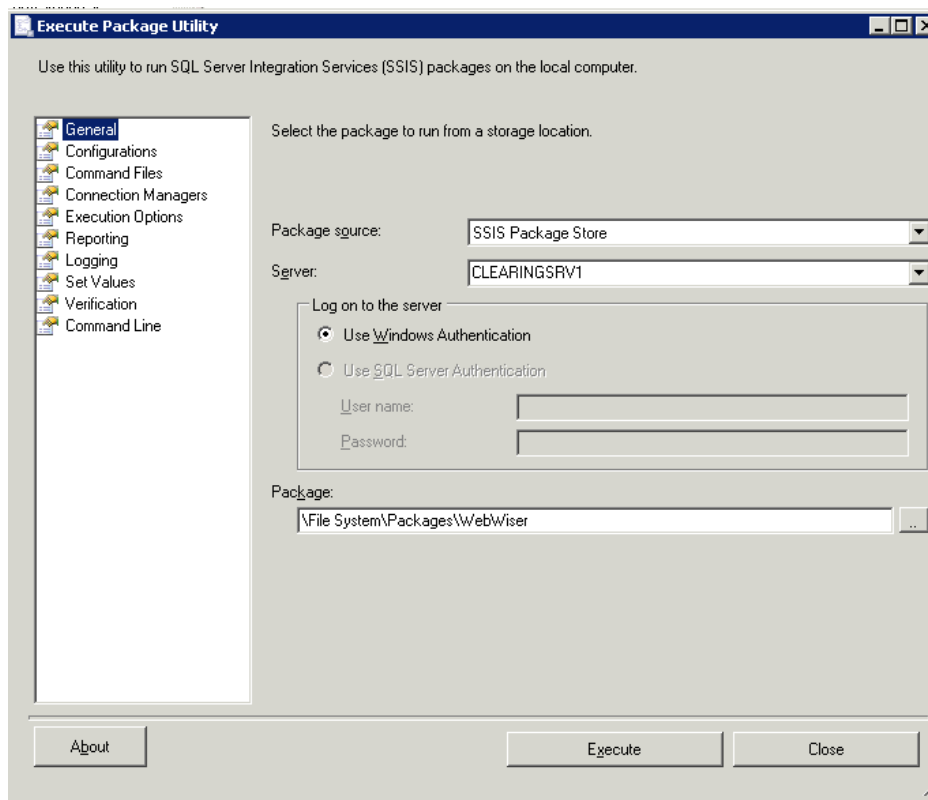
Kuva 23. Deployment Utility ja konfiguraation muuttaminen

5.4 SSIS-pakettien suoritus

Paketti voidaan suorittaa usealla eri tavalla. Pakettia kehitettäessä luonnollisin tapa on käyttää graafista käyttöliittymää eli SSIS Designeria. SSIS:n mukana tulee myös kaksi pakettien ajamiseen soveltuvaa erillistä ohjelmaa, merkkipohjainen DTEXEC ja graafinen DTEXECUI. Paketteja voidaan ajaa myös SQL Server Agent job -työkalulla, joka mahdollistaa pakettien erittäin monipuolisen ajastamisen (SQL Server R2 Books Online).

Merkkipohjainen DTEXEC-ohjelma soveltuu erinomaisesti esimerkiksi erilaisten skriptien tai tallennettujen proseduurien (Stored Proc) yhteydessä tapahtuviin pakettien käynnistykseen. DTEXECUI-ohjelmalla paketit voidaan käynnistää käyttäjäystävällisemmin. Ohjelma soveltuu erityisesti tilanteisiin, joissa yksittäinen paketti pitää ajaa kerta-luontoisesti. Kuvassa 24 on esitetty DTEXECUI-ohjelman käyttöliittymä. Paketin käynnistuksen yhteydessä voidaan antaa paketille tarvittavat parametrit, kuten mahdollinen

konfiguraatiodiedosto tai yksittäisen paketin objektin ominaisuus, kuten muuttujan arvo. Tämä mahdollistaa yksittäisen paketin ajamisen erilaisilla ehdoilla.



Kuva 24. DTEXECUI-ohjelma

6 TOTEUTUS

Toteutettu työ on tässä luvussa kuvattu yleisellä tasolla johtuen tiedon arkaluonteisuudesta ja liikesalaisuuksista.

Toimeksianto

Toimeksiantona on toteuttaa sääntöpohjainen työkalu televerkkojen väärinkäytösten havainnointiin. Työkalu saa lähtötiedoikseen valvontakäskyjä ja raportoi annettujen käskyjen mukaiset poikkeamat erikseen määritellyille tahoille.

Työkalu toteutetaan SQL Server 2008 R2 -tuotteella ja siinä erityisesti SSIS:ään tukeutuen.

Ensimmäisessä vaiheessa työkalun avulla pyritään havaitsemaan sopimus- ja menettelytapaväärinkäytöksiä palvelunumeropalveluissa. Palvelunumeropalveluilla tarkoitetaan kaikkia niitä lisämaksullisia palveluita, joihin voidaan soittaa kiinteällä tai mobiililla päätelaitteella. Lisämaksullisia palveluita ovat mm. asiointipalvelut, aikuisviihde- ja ajanvietepalvelut sekä pienmaksamisen mahdollistavat maksupalvelupalvelut.

Palvelunumeroiden palvelutaksat Suomessa perustuvat televerkoissa lähetettäviin sykäyksiin (Pulse). Yhden sykäyksen hinta on operaattorista riippumatta aina 0,0673 euroa. Yksittäisen palvelunumeron hinta voi olla enimmillään 60 euroa eli 891 sykäystä.

Palvelunumeropalveluiden tuloutus tapahtuu siten, että palveluntarjoajan liittymäoperaattori tulouttaa palvelumaksut sovitun jakoperusteen mukaan palveluntarjoajalle ja veloittaa ne palvelua käyttäneeltä asiakkaaltaan tai palvelua käyttäneen asiakkaan liittymäoperaattorilta, joka veloittaa palvelumaksut edelleen asiakkailtaan osana normaalia asiakaslaskutustaan.

Palvelunumeropalveluiden käytöllä on merkittävä osuus suurten asiakaslaskujen synnyssä lisäten näin riskiä sopimusväärinkäytöksille. Palvelunumeroita voidaan pyrkiä myös väärinkäyttämään mahdollisia palvelun tai tekniikan heikkouksia

hyväksikäyttäen. Sekä palvelunumeropalveluiden sopimus- ja menettelytapaväärinkäytöksillä on mahdollista aiheuttaa lyhyessä ajassa operaattorille mittavia taloudellisia menetyksiä, joten väärinkäytösten havaitseminen mahdollisimman aikaisessa vaiheessa on ensiarvoisen tärkeää.

Lähtötietojen keruu ja yhtenäistäminen sekä väärinkäyttöä osoittavien indikaattoreiden valinta

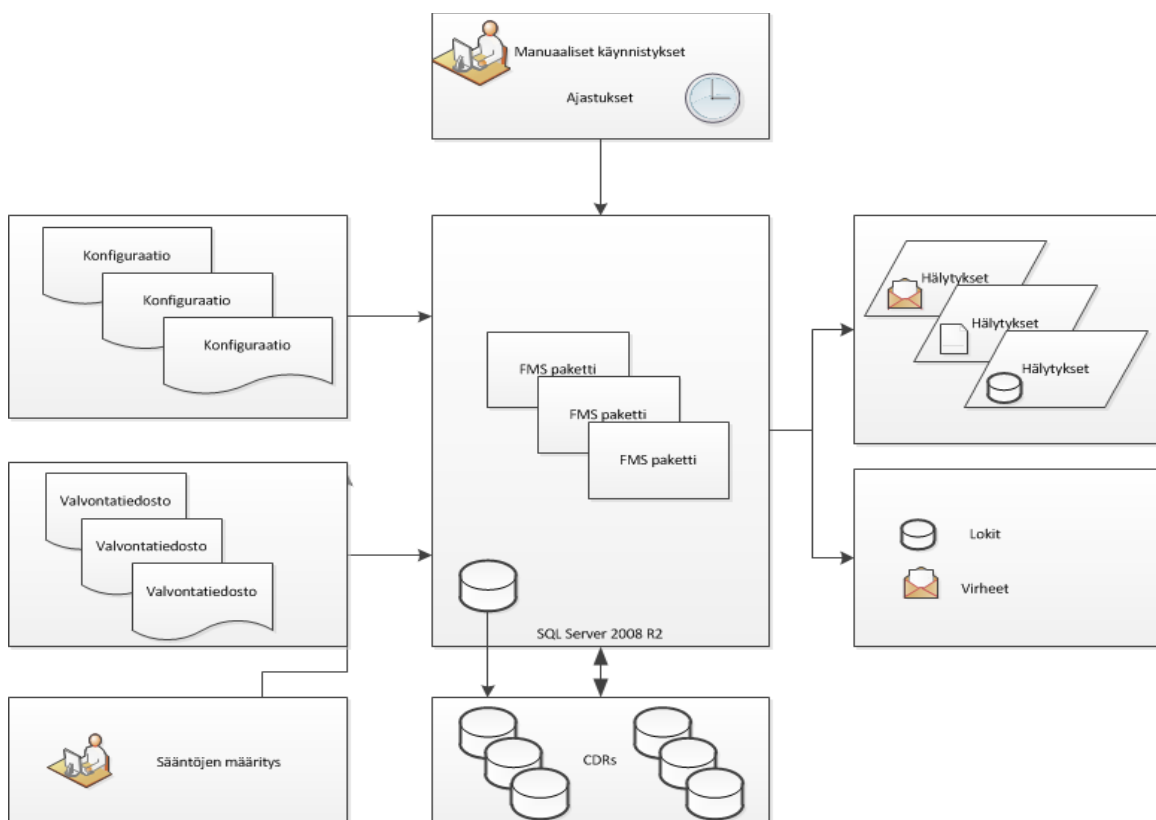
Väärinkäytösten seurannan lähtötietoina käytetään kiinteän- ja mobiiliverkon puhelutietokantoja, joihin medioinnin eri verkkoelementeistä kerätyt tapahtumatiedot tallennetaan yhtenäisessä, tekniikkariippumattomassa muodossa erilaisten tapahtumatietoa hyödyntävien järjestelmien lähtötiedoiksi. Tekniikkariippumattomuuden lisäksi kiistattomina etuina tietokantojen käytössä on suuri suorituskky sekä mahdollisuus toteuttaa valvontakäskyt eli säännöt SQL-kielisinä.

Palvelunumeropalveluiden väärinkäyttöä osoittavien indikaattoreiden valinta tapahtuu pääasiassa aikaisemmista väärinkäytöksistä saatujen kokemusten ja fraud-asiantuntijoiden tietämyksen perusteella. Käyttökelpoisia indikaattoreita ovat mm. puhelukappaleet, kestot ja sykäykset. Puhelutietokantojen tapahtumatietueita täydennetään tarvittavilla tiedoilla väärinkäytösanalyysien luotettavuuden parantamiseksi. Palvelunumeropalveluiden väärinkäytösseurannan kannalta täydennettäviä tietoja ovat mm. liittymä- ja asiakastyypit.

Palvelunumeroihin soitetuista puheluista muodostetaan lisäksi erillinen, yhden vuorokauden tiedot sisältävä osajoukko valvontakäskyjen suorituskyyvyn parantamiseksi. Tiedot liikennetapaukset myös summataan ja niille lasketaan erilaisia tilastollisia tunnuslukuja esimerkiksi keston, sykäyksiin ja kappaleisiin perustuen.

Yleiskuvaus sääntöpohjaisesta televerkkojen havainnointityökalusta

Fraud-asiantuntijat määrittelevät fraud-tunnisteet valvontatiedostoihin valvontatyökalun lähtötiedoksi. Valvontatyökalu hakee mahdollisia väärinkäytöstopauksia määriteltyjen aikataulujen, konfiguraatioiden ja annettujen valvontakäskyjen mukaisesti valvontatiedostojen valvontapakeeteissa kuvatuista tietolähteistä raportoiden mahdolliset hälytykset valvontapakeeteissa määritetyille tahoille. Valvontaprosessi kokonaisuudessaan lokite-
taan mahdollisten virhetilanteiden varalta sekä teletunnistetietojen käsittelystä johtuvista lainsäädännöllisistä velvoitteista johtuen (Sähköisen viestinnän tietosuojalaki 2004). Myös mahdolliset virheet valvontapaketien fraud-tunnisteiden määrittelyssä kirjataan ja raportoidaan eteenpäin. Seuraavassa kuvassa (kuva 25) esitetään yleiskuva valvontatyökalusta.



Kuva 25. Yleiskuva valvontatyökalusta

Valvontapaketit

Valvontapaketit määritellään XML-muotoisiin valvontatiedostoihin. Valvontakäskyjen eli sääntöjen lisäksi valvontapaketteihin määritellään mm. valvontakäskyjen kohdetietokannan tietokantayhteyttä kuvaava vakio, väärinkäytöstapauksesta hälytyksen saavien tahojen sähköpostiosoitteet sekä hälytystiedostojen muodostamiseen tarvittavia muita tietoja.

Valvontapaketien rakenne on kuvan 26 skeeman mukainen.

```

<?xml version="1.0"?>
<xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified" xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="Packets">
    <xs:complexType>
      <xs:sequence>
        <xs:element minOccurs="0" name="Packet">
          <xs:complexType>
            <xs:sequence>
              <xs:element minOccurs="0" name="name" type="xs:string" />
              <xs:element minOccurs="0" name="description" type="xs:string" />
              <xs:element minOccurs="0" name="repository" type="xs:string" />
              <xs:element minOccurs="0" name="result_name" type="xs:string" />
              <xs:element minOccurs="0" name="result_path" type="xs:string" />
              <xs:element minOccurs="0" name="xslt" type="xs:string" />
              <xs:element minOccurs="0" name="condition" type="xs:string" />
              <xs:element minOccurs="0" name="distribution_list" type="xs:string" />
              <xs:element minOccurs="0" name="distribution_type" type="xs:string" />
              <xs:element minOccurs="0" name="valid_from" type="xs:string" />
              <xs:element minOccurs="0" name="valid_until" type="xs:string" />
              <xs:element minOccurs="0" name="created_by" type="xs:string" />
              <xs:element minOccurs="0" name="created" type="xs:string" />
            </xs:sequence>
          </xs:complexType>
        </xs:element>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>

```

Kuva 26. Valvontapaketien skeema

Kuten kuvan 26 skeeman perusteella voidaan havaita, XML-muotoinen valvontatietue voi sisältää 1-n kappaletta valvontapaketteja. Valvontapaketien tulee täyttää annetun skeeman vaatimukset.

Valvontapaketin elementtien tietosisältö ja tarkoitus on kuvattu taulukossa 5.

Taulukko 5. Valvontapaketin elementtien tietosisältö ja tarkoitus

Elementti	Selite
name	Paketin vapaamuotoinen nimi
description	Kuvaus paketin tarkoituksesta ja toiminnallisuudesta
repository	Tietokantayhteyttä kuvaava vakio
result_name	XML-tiedosto, johon mahdollinen vaste kirjoitetaan
result_path	Hakemistopolku, johon mahdollinen vaste kirjoitetaan
xslt	Viittaus xslt-tiedostoon, mikäli XML-muotoiseen vasteeseen halutaan kohdistaa muotoilua, laskentaa tai muuta muunnosta
condition	Valvottavan väärinkäytöstopauksen säännöt SQL-lauseena
distribution_list	Sähköpostiosoitteet, joihin mahdollinen hälytys generoidaan, puolipisteellä eroteltuna
distribution_type	Varattu tulevaan käyttöön
valid_from	Ehdon voimassaolon alkuaika
valid_until	Ehdon voimassaolon loppuaika
created_by	Paketin luoneen käyttäjän tunnus
created	Aika, jolloin paketti luotu

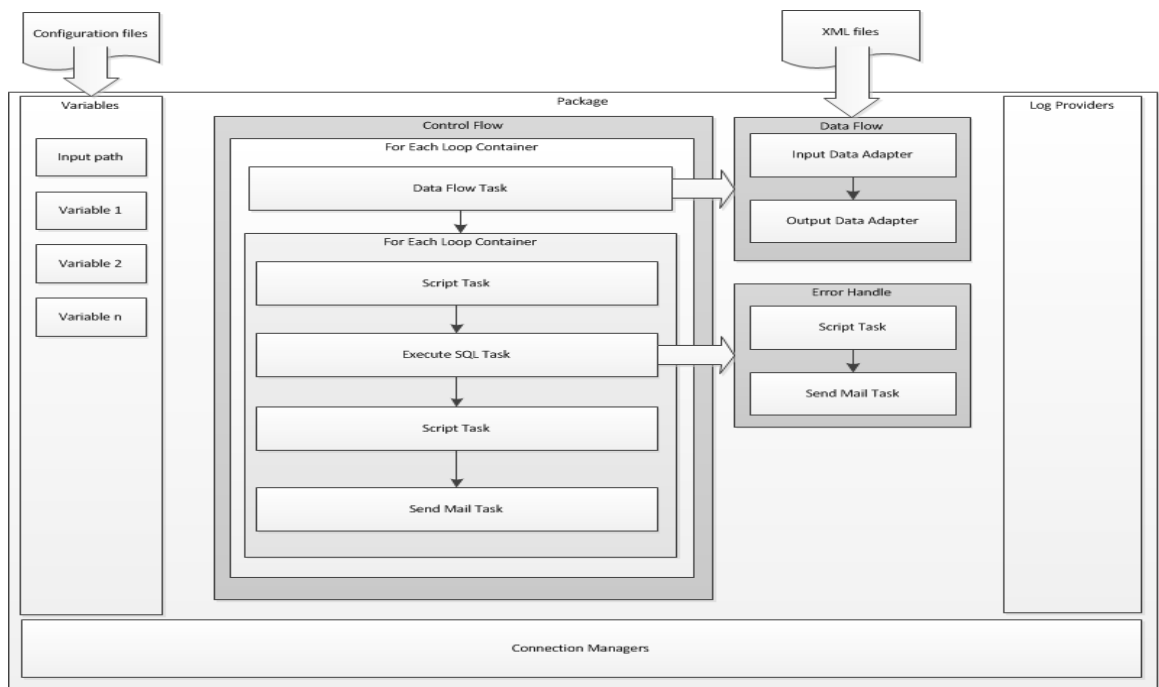
Kuvassa 27 esitetään yksittäinen valvontapaketti viisi pakettia sisältävästä valvontatiedostosta.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<Packets desc="Fraud and tariff checks for ServiceNumbers">
  <Packet>
    <name>Fraud/Tariff_report:xxxxxxx</name>
    <description>Test for demonstration purposes</description>
    <repository>GTTYPE05</repository>
    <result_name>Servicenumber_xxx</result_name>
    <result_path>\\Fraud\Output\ServiceNumber\</result_path>
    <xslt>none</xslt>
    <condition>
      <![CDATA[
        --Note! LAST_DAY must be exists!!
        SELECT
          EXID,ORIGCID,RCN,RECTYPE,RCODE,IROUTE,OROUTE,NVL(OBACOMP,O) AS OBACOMP,NVL(OBBCOMP,O) AS OBBCOMP,SD
        FROM LAST_DAY
        WHERE
          REGEXP_LIKE (ASUB, '^0209[89]|^0[67]0')
          AND IROUTE NOT LIKE 'SSF%'
          --WHITELIST
          AND ASUB NOT IN ('0x00xxxxx','0x00xxx','0x00xxxxx','0x0xxxxx')
      ]]>
    </condition>
    <distribution_list>foo@dna.fi;fake@dna.fi</distribution_list>
    <distribution_type>attachment</distribution_type>
    <valid_from>1.8.2013</valid_from>
    <valid_until></valid_until>
    <created_by>FMS</created_by>
    <created>1.8.2013</created>
  </Packet>
  <Packet>...</Packet>
  <Packet>...</Packet>
  <Packet>...</Packet>
  <Packet>...</Packet>
</Packets>
```

Kuva 27. Yksittäinen valvontapaketti ja sen tietosisältö

Valvontatyökalu

Valvontatyökalu on integrointipalvelimelle toteutettu SSIS-paketti, joka käynnistyy joko manuaalisesti tai ajastettuna. Paketti saa käynnistysparametrinaan konfiguraatio-tiedoston, jonka oleellisimpana parametrina on valvontatiedostojen sijainnista kertova hakemistopolku. Valvontatyökalu käy läpi rekursiivisesti parametrinaan saadun hakemistorakenteen etsien hakemistoista XML-päätteisiä valvontatiedostoja. Löydetyt valvontapaketit validoidaan ja luetaan järjestelmän muistiin tiedosto kerrallaan ja käydään tämän jälkeen läpi valvontapaketti kerrallaan. Jokaisen läpikäytävän valvontapaketin tietosisällön perusteella asetetaan ohjelmallisesti tarvittava tietokantayhteys ja suoritetaan tämän jälkeen valvontapaketissa määritelty SQL-muotoinen valvontakäske. Jos valvontakäsken suorituksesta saadaan vaste, kirjoitetaan vaste XML-tiedostoksi valvontapaketissa annettuun hakemistoon ja lähetetään hälytys valvontapaketissa määriteltyihin sähköpostiosoitteisiin. Mikäli valvontapakettiin on määritelty viittaus käytettävään XSLT-tiedostoon, tehdään vasteelle myös XSLT-transformaatio vasteen kirjoituksen yhteydessä. Mahdollisia virheitä valvontakäskejen määrittelyssä valvotaan erillisellä tapahtumankäsittelijällä, joka raportoi virheet eteenpäin määritetyille tahoille. Lisäksi kaikki valvontatyökalun tapahtumat valvontakäskeyneen lokitetaan erilliseen tietokantaan. Kuvassa 28 on esitetty edellä mainitun toiminnallisuuden toteuttavan SSIS-paketin arkkitehtuuri.



Kuva 28 Valvontatyökalun SSIS-paketin arkkitehtuuri

SSIS-arkkitehtuurin näkökulmasta parametrina saadun hakemistorakenteen ja XML-tiedostojen sisältämien valvontapakettien läpikäynti tapahtuu kahden toistorakenteita tukevan säiliön avulla (For Each Loop Container). SSIS-paketin hierarkiassa ylempänä oleva säiliö vastaa saadun hakemistorakenteen läpikäynnistä ohjaten löydettyjä XML-tiedostoja Data Flow Task -tehtäväkomponentille tiedosto kerrallaan. Hierarkiassa alempana oleva säiliö vastaa yksittäisten XML-tiedostojen sisältämien valvontapakettien käsittelyistä.

Käsiteltävänä oleva XML-tiedosto luetaan järjestelmän muistiin Data Flow Task -tehtäväkomponentilla käyttäen hyväksi XML-tiedostojen lukuun soveltuvaa lähdeadapteria ja tietoa muistiin tallentavaa Recordset-tyypistä kohdeadapteria.

Script Task -tehtäväkomponentteja käytetään ajonaikaisesti muodostettavien tietokantayhteyksien (Connection Manager) asettamiseen, valvontapakettien sääntöjen perusteella generoitavien vastetiedostojen kirjoittamiseen ja muotoiluun sekä tapahtumankäsittelijän virhepinon rakentamiseen. Tehtäväkomponenttien ohjelmointi tapahtuu VB.NET -ohjelmointikielellä ADO.NET -palveluita hyväksikäyttäen. Kuvassa 29 valvontatyökalu on löytänyt kuusi valvontatyökalun ehtojen mukaista poikkeamaa ja kirjoittanut niistä valvontakäskeyjen mukaisen XML-tiedoston.

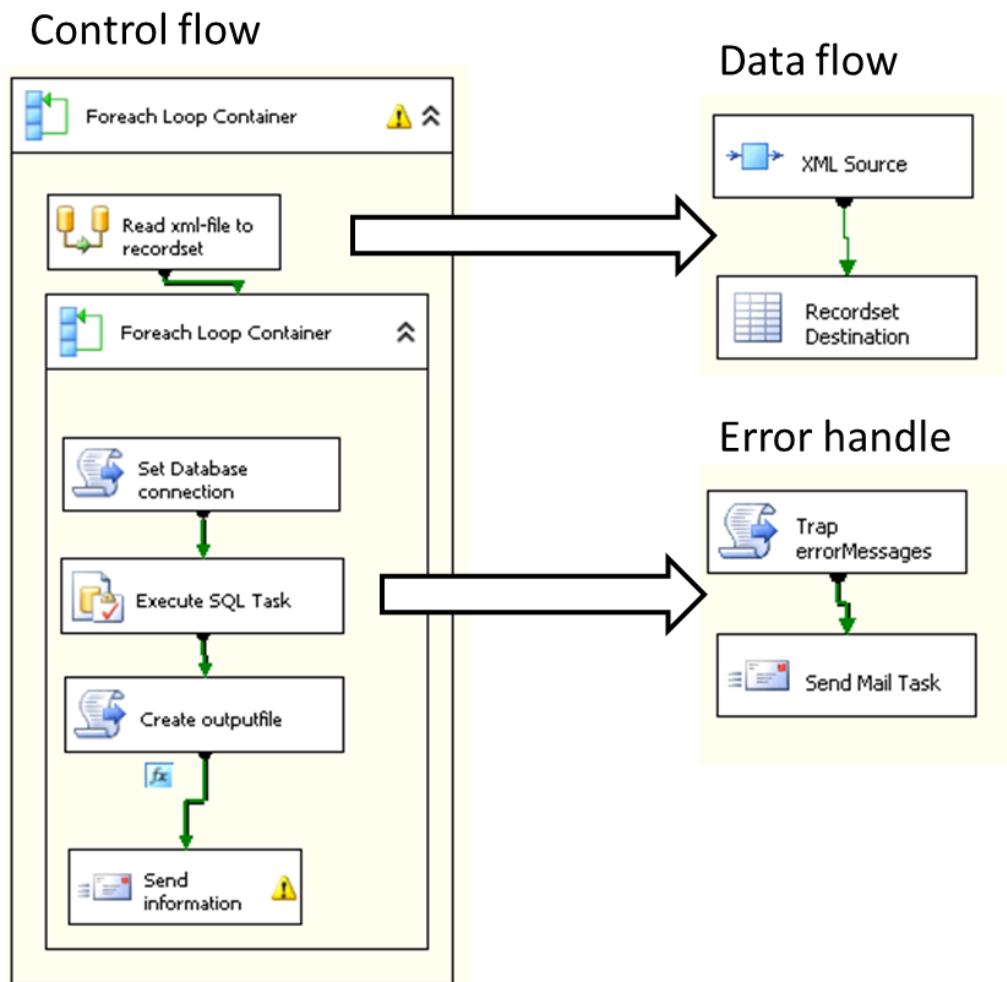
```
<DocumentElement>
  <Result>
    <PERIOD>20131024</PERIOD>
    <ASUB>044313xxx</ASUB>
    <TYPE>POSTPAID</TYPE>
    <PCS>13</PCS>
    <MINUTE>2</MINUTE>
    <PULSE>3395</PULSE>
    <CHARGE>228.48</CHARGE>
  </Result>
  <Result>
    <PERIOD>20131024</PERIOD>
    <ASUB>044275xxxx</ASUB>
    <TYPE>POSTPAID</TYPE>
    <PCS>9</PCS>
    <MINUTE>1</MINUTE>
    <PULSE>3200</PULSE>
    <CHARGE>215.36</CHARGE>
  </Result>
  <Result>
  </Result>
  <Result>
  </Result>
</DocumentElement>
```

Kuva 29. Valvontatyökalun kirjoittama XML-tiedosto

Itse valvontakäsky tietokantaan tehdään Execute SQL Task -tehtäväkomponenttia hyväksikäyttäen. Valvontapaketissa määritellyt SQL-muotoiset valvontasäännöt välitetään tehtäväkomponentilla, joka huolehtii kyselyn suorittamisesta annettua tietokantayhteyttä vasten ja tietokannasta saadun vasteen käsittelystä.

Hälytysten generointiin käytetään Send Mail Task -tehtäväkomponenttia, jonka avulla voidaan lähettää hälytykset ja mahdolliset liitetiedostot tarvittaville tahoille.

Kuvassa 30 on esitetty toteutetun valvontatyökalun SSIS-paketin komponentit ja niiden suoritusjärjestys.



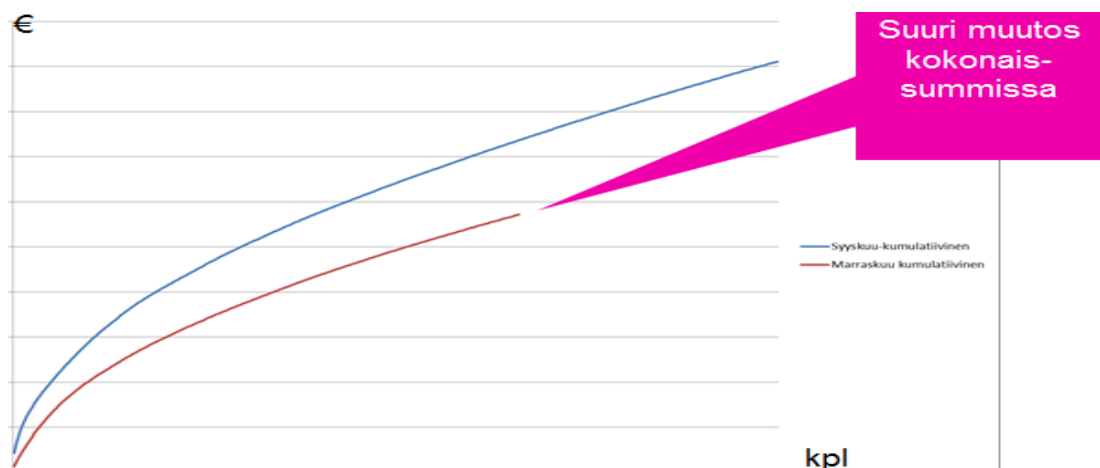
Kuva 30. Valvontatyökalun SSIS-paketti komponentteineen

7 POHDINTA

Alan kirjallisuuden mukaan SSIS on erittäin monipuolinen ja tehokas, mutta kompleksinen ja vaikeasti opittava tuote. Saatujen kokemusten perusteella edellä mainittuihin näkemyksiin on helppo yhtyä. "Drag-and-Drop" -tyylinen SSIS-pakettien kehitystyö vaikuttaa näennäisesti helpolta lähestymistavalta, mutta vaatii todellisuudessa pakettien kehittäjiltä monipuolista ja laajaa tietoteknistä osaamista niin sovelluskehityksestä, tietokannoista kuin itse SSIS-palvelinohjelmistostakin. Ilman riittävää osaamis- ja koulutuspohjaa pakettien kehittäjä saattaa nopeasti törmätä aikaa vieviin ja turhauttaviinkin ongelmiin.

SSIS soveltuu kuitenkin erinomaisesti televerkkojen väärinkäytösten havainnointiin. SSIS:n ja SQL Serverin arkkitehtuuri mahdollistaa monipuolisten työnkulkujen rakentamisen, helpon pääsyn eri tietolähteisiin, tehokkaan tietojen prosessoinnin sekä valmiit algoritmit niin ohjattuihin kuin ohjaamattomiin tiedonlouhintatekniikoihin. SSIS:n käyttö ei myöskään pääsääntöisesti vaadi erillisiä investointeja lukuun ottamatta mahdollisia palvelimia, sillä SSIS tulee osana SQL Server palvelintuotetta, joka teleoperaattoreilla tyypillisesti on jo valmiiksi osana yrityksen järjestelmäarkkitehtuuria.

SSIS:lle toteutetun työkalun käyttö palvelunumeroiden käytön seurannassa antoi lupaavia tuloksia. Työkalun generoimien valvontaraporttien perusteella tehtiin toimeksiantajan televerkoissa estotoimenpiteitä siten, että kahden kuukauden vertailujaksolla väärinkäytösepäilyjen kappalemäärä laski 34 prosenttia ja kumulatiivinen euromäärä 37 prosenttia suhteessa vertailujakson alkuun (kuva 31).



Kuva 31 Väärinkäytösepäilyjen kappale ja euromäärien kehitys vertailujaksolla

Työkalun avulla löydettiin myös mobiilikeskuksista konfigurointivirhe, joka mahdollisti tietyissä tapauksissa palveluiden ilmaisen käytön. Sääntöpohjaisen työkalun käyttöä kannattaakin jatkossa laajentaa myös muihin erikseen määriteltäviin palveluihin vastaavanlaisten virheiden havaitsemiseksi.

Työkalu generoi testijaksolla kohtuullisen paljon myös aiheettomia väärinkäytösepäilyjä erityisesti prepaid-liittymistä palvelunumeroihin soitettujen puheluiden osalta. Mainittujen väärinkäytösepäilyjen määrää pystytään joiltakin osin vähentämään sääntöjä tarkentamalla. Mukana on kuitenkin puhelutapauksia, joihin ei sääntöpohjaisilla työkaluilla voida vaikuttaa. Näiden puhelutapausten suhteen harkitsemisen arvoinen lähestymistapa voisi olla ohjattaviin tiedonlouhintamenetelmiin perustuvien luokittelumallien määrittäminen ja koulutus, ja pyrkiä tätä kautta löytämään muutoksia asiakkaiden palvelunumeropalveluiden oletuskäyttäytymisessä. SSIS ja SSAS tarjoavat tähän hyvät työkalut.

LÄHTEET

Augustin, S., Gaißer, C., Knauer, J., Massoth, M., Piejko, K., Rihm, D., & Wiens, T. (2012, May). Telephony Fraud Detection in Next Generation Networks. In AICT 2012, The Eighth Advanced International Conference on Telecommunications (pp. 203–207)

Beck Computer Systems. 2003. Breaking the back of telephone fraud.

Bihina Bella, M.A., Olivier, M.S., & Eloff, J.H.P. 2005a. Using the Internet Protocol Detail Record standard for NGN billing and fraud detection, in: Proceedings of the 5th Information Security South Africa (ISSA) Conference 2005, Sandton, South Africa, 29.6. - 1.7.2005.

Bihina Bella, M. A., Olivier, M. S., & Eloff, J. H. P. 2005b. A fraud detection model for Next-Generation Networks. In Proceedings of the 8th Southern African Telecommunications Networks and Applications Conference (SATNAC 2005), Central Drakensberg, KwaZulu-Natal, South Africa.

Bihina Bella, M.A., Eloff, J. H., & Olivier, M. S. 2009. A fraud management system architecture for next-generation networks. Forensic science international, 185 (1), 51–58.

Communications Fraud Control Association (CFCA) Announces Results of Worldwide Telecom Fraud Survey. 2011. Luettu 25.11.2013, http://www.cfca.org/pdf/survey/Global%20Fraud_Loss_Survey2011.pdf

Communications Fraud Control Association (CFCA) Announces Results of Worldwide Telecom Fraud Survey. 2013. Luettu 25.11.2013, http://www.cfca.org/pdf/survey/Global%20Fraud_Loss_Survey2013.pdf

Gong, L. 2011. The application of Naive Bayesian Classification in anti-fraud system of telecommunications, Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on , vol.2, no., pp.1061, 1064, 26.7. - 28.7.2011

- Cortes, C., & Pregibon, C. 2001. Signature-based methods for data streams. *Data Mining and Knowledge Discovery* 5.3 (2001): 167–182.
- Gosset, P., Hyland, M. 1999. Classification, Detection and Prosecution of Fraud on Mobile Networks. *Proc. of ACTS Mobile Summit, Sorrento, Italy 1999*. Tulostettu 12.11.2013, [ftp://ftp.cordis.europa.eu/pub/ist/docs/ka4/10187.pdf](http://ftp.cordis.europa.eu/pub/ist/docs/ka4/10187.pdf)
- Halkidi, M., Batistakis, Y. & Vazirgiannis, M. 2001. On clustering validation techniques. *J. Intell. Inf. Syst.*, 17 (2-3): 107–145
- Hartgrove, K., Intrevado, P. & Abel, S. R. 2008. Validation Study: Clarity Multistrip Urocheck. *Journal of the American Society for Clinical Laboratory Science*. *Clinical Laboratory Science* 21 (3), 158–161.
- Haselden, K. 2009. *Microsoft SQL Server 2008 Integration Services Unleashed*. 1. painos. USA: Pearson Education, Inc., Indiana.
- Hearne, S. 2004. A Fraud Detection Framework for Next-Generation Telecommunications Networks, M S. thesis. Waterford Institute of Technology.
- Hilas, C.S. 2012. Data mining approaches to fraud detection in telecommunications. A short description of ongoing research. 2nd Pan-Hellenic Conference on Electronics and Telecommunications - PACET'12, 16. - 18.3.2012, Thessaloniki, Kreikka. Luettu: 10.11.2013, http://www.pacet.gr/index_htm_files/S25_.pdf
- Hinde, S. F. 1996. Call record analysis. *IEE Seminar Digests*. Vol. 8. No. 1996. 1996.
- Hovi, A., Hervonen, H., & Koistinen, H. 2009. *Tietovarastot & business intelligence*.
- Jacobs, R. 2002. Telecommunications Fraud: the single biggest cause of revenue loss for telecommunications providers, White Paper, Dimension data.
- Kamtsan, C., Tirkkonen, T., & Sihvola, S.K. 2010. Asiakkaan tunnistaminen, tunteminen ja väärinkäytösseuranta. Rajoitettu saatavuus.

Kantardzic, M. 2011. Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.

Knight B., Veerman E., Moss J., Davis M., & Rock C. 2012. Professional Microsoft SQL 2012 Integration Services. USA: Wiley & Sons, Inc. Indianapolis.

Kvarnstrom H., Lundin, E., and Erland, J., Combining fraud and intrusion detection – meeting new requirements, in proc. Fifth Nordic Workshop on Secure IT Systems (NordSec), Reykjavik, Iceland, October 12-13, 2000.

Nanda, A. 2008. Hands-On Microsoft SQL Server 2008 Integration Services. 2. painos. USA: McCraw-Hill Companies.

Rosset, S., Murad, U., Neuman, E., Idan, Y., & Pinkas, G. 1999. Discovery of Fraud Rules for Telecommunications: Challenges and Solutions KDD-99: 409–413.

Sähköisen viestinnän tietosuojalaki 16.6.2004/516.

SQL Server 2008 R2 Books Online. Luettu 24.11.2013, [http://msdn.microsoft.com/en-us/library/ms141026\(v=sql.105\).aspx](http://msdn.microsoft.com/en-us/library/ms141026(v=sql.105).aspx)

Tan, P., Steinbach, M., & Kumar, V. 2005. Introduction to data mining. 1. painos. USA: Addison-Wesley.

Taniguchi, M., Haft, M., Hollmén, J., & Tresp, V. 1998. Fraud detection in communication networks using neural and probabilistic methods. In Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on (Vol. 2, pp. 1241–1244). IEEE.

Tindal, S. 2009. Voip hackers strike Perth business. ZDNet. Luettu 12.11.2013, <http://www.zdnet.com/voip-hackers-strike-perth-business-1339294515/>

Theodoridis, S., & Koutroubas, K. 1999. Pattern Recognition. Academic Press.