Haaga-Helia
ammattikorkeakoulu Oy

Tiedon virtauksen hallinta metadatan avulla

Janne Laaksonen

Opinnäytetyö

Tietojenkäsittelyn koulutusohjelma

2019

Haaga-Helia ammattikorkeakoulu Oy

Tiivistelmä

26.09.2019

Tekijä(t)
Janne Laaksonen

Koulutusohjelma
Tietojenkäsittelyn koulutusohjelma

Opinnäytetyön otsikko
Tiedon virtauksen hallinta metadatan avulla

Sivu- ja liitesivumäärä
25 + 4

Opinnäytetyön otsikko englanniksi
Controlling Data Flow with Metadata Management

Oli kyseessä sitten henkilötiedot, pörssiyrityksen liiketoimintatiedot tai somessa liikkuvat tiedot – kaikki kaipaavat avoimuutta ja näkyvyyttä siihen, miten, missä, milloin ja kuka on tiedon luonut tai kuka tietoa käyttää.

Yksinkertaisesti sanoen metadata on tietoa tiedosta, mutta yhden tuottama metadata on toiselle liiketoiminnan tietoa. Metadata pitää sisällään esimerkiksi tietoa siitä, onko tieto numeraalista vai tekstiä, milloin tieto on luotu tai kuka sen on luonut. Metadataa on selkeimmin esillä tietokannoissa, mutta metadataa voidaan myös kerätä muualta, kuten esimerkiksi ohjelmointikoodista (muuttujat), kuvista tai sosiaalisesta mediasta.

Tiedon virtaus (engl. data flow) on tiedon siirtämistä tietojärjestelmästä toiseen. Tiedon siirrossa yleisimmin käytetty tapa on ETL (engl. Extract, Transform & Load). Tällä tarkoitetaan tiedon siirron jakamista kolmeen osaan: tiedon hakeminen lähdejärjestelmästä, tiedon muokkaaminen ja tiedon lataaminen kohdejärjestelmään.

Data Lineage on tietovirran hallintaa metadatan avulla siten, että lähdejärjestelmän metadata liitetään siirrossa tehtäviin tiedon muokkaussääntöihin ja kohdejärjestelmän tietokannan metadataan.

Verkostoituneissa järjestelmissä Data Lineage muodostaa ”putken” tai ”puun”, josta näkyy mistä järjestelmistä tieto on peräisin ja missä muodossa se on talletettu, mutta yksittäistä tietoelementtiä ei näe. Asiaa voisi verrata kodin vesiputkeen, josta putkea seuraamalla päätyy aina isompaan putkeen, ja lopulta vesilaitokseen, ja sieltä sinne mistä laitos ottaa vetensä, mutta silti ei pystyttäisi tunnistamaan yksittäistä vesipisaraa, mutta silti pystytään kuvaan koko putkisto, mistä se lähtee ja minne kaikkialle se päätyy.

Data Lineage voidaan katsoa kahdesta eri suunnasta. Lähdejärjestelmästä katsoen putki näyttää ”puulta”, ja siitä näkee mihin kaikkialle yksittäistä tietoa käytetään. Tätä voi käyttää hyväkseen esimerkiksi muutoksen hallinnassa tai data-arkkitehtuurisessa tietojen hallinnassa. Raportoinnista katsoen Data Lineage näyttää, mistä järjestelmistä kyseinen arvo on tullut, ja millä laskusäännöillä se on muodostettu. Tämä on tiedon avoimuuden kannalta tärkeää, jotta voidaan todeta, että tieto on laskettu oikein.
Yhdistämällä Data Lineage muuhun tietoon, voidaan myös saavuttaa tietoa siitä, että kuka omistaa, käyttää ja auditoi kyseisen tiedon tai minne tieto on tallennettu. Tämä auttaa organisaatioita hallitsemaan tietoa tietolähtöisesti (esim. GDPR, Brexit vaikutukset, maakohtaiset tietosuojasäännöt.)

Asiasanat
metadata, data lineage, tiedon hallinta, data governance

Sisällys

1 Johdanto

Pilvipalvelujen käyttöönotto on luonut tilanteen, jossa uusien resurssien ja teknologioiden käyttöönotto on nopeaa ja halpaa verrattuna siihen mitä olemassa olevien järjestelmien muuttaminen on hidasta ja kallista. Samalla kun tiedon määrä kasvaa on epäselvää, miten tieto on laskettu ja miksi kahdesta eri järjestelmästä saadut tunnusluvut eroavat toisistaan.

Verkostoituneissa tietojärjestelmissä tiedon virtauksen hallinta on alkutekijöissään. Yritykset ovat ottaneet käyttöön tietojen operatiivista hallintaa, ydintiedon hallintaa, mutta tiedon laadun hallinta ja tiedon virtauksen hallintaa ei ole tai on karkeasti konseptitasolla (liite 3).

Metadatan hallinta on tietojen hallinnan perusta. Jos sitä vertaisi talon rakentamiseen, metadatan hallinta olisi sokkeli, minkä päälle kaikki muu rakennetaan. Metadatan hallinnan avulla voidaan rakentaa perusteet tiedon hallintaan. Missä tieto luodaan, kenen toimesta, missä kaikkialla tietoa säilytetään, miten sitä muokataan, kuka käyttää tietoa – kaikkeen tähän metadatan hallinnalla voidaan luoda tietopohjainen näkymä, jossa oikeuksia ja valvontaa ei tehdä järjestelmäpohjaisesti, vaan tiedon pohjalta mahdollistaen kokonaisnäkymän välittämättä alla olevista teknologiaratkaisusta.

Metadatan hallinta on hidasta ja monesti vaikeaa, koska työkalut sen hallintaan eivät ole kovinkaan kehittyneet. Data Lineage (liite 1) on metadatasta visualisointi, jolla tietovirrat järjestelmistä toiseen saadaan kirjattua, hallittua, mahdollistaa suunnitelmallisen tietopohjaisen muutoksenhallinnan ja.

Data Lineage antaa näkyvyyden tieto attribuutin tasolle. Tällä on huomattava synergia etu ydintietojen hallinnan ja tiedon laadun hallinnan kanssa. Molemmissa hallinnoissa tarvitaan näkyvyyttä, miten ja missä tietoattribuuttia käytetään ja miten tietoelementtiä muokataan siirtojen yhteydessä. Tämä mahdollistaa myös tiedon avoimuuden, jolloin tiedon lähteet, muokkaukset ja rajaukset voidaan esittää.

Tiedon omistajuuden ja käyttäjäoikeuksien näkökulmasta miettien, tiedon omistajalla olisi hallinta omistamaansa tietoon järjestelmästä huolimatta ja näkyvyys missä kaikkialla hänen omistamaansa tietoa säilytetään ja käytetään.

Data Lineage luo organisaatioille sekä kustannussäästöjä että ketteryyttä. Kun tiedon attribuuttien metadata ja lähteet ovat visualisointuja, uusien yhteyksien luominen on nopeampaa ja vanhojen yhteyksien muuttaminen on kokonaisvaltaisempaa, koska visualisointi

luo paremman kuvan muutoksen vaikutuksista. Tämä vähentää muutoksenhallinnassa ennalta ennakoimattomien virheiden (engl. unknown unknown problems) määrää.

1.1 Opinnäytetyön tuotos

Opinnäytetyöni tuotoksessa haluan selkeyttää, mitä on metadata management ja sen työkalu Data Lineage ja sen eri käyttötarkoitukset. Lisäksi luon toimintamallin, mistä rajapinnasta ja mitä tietoa pitää kerätä luodakseen toimivan tietomallin Data Lineageen, sekä erilaisia raporttimalleja Data Lineagesta.

Ratkaisu on teknologiasta riippumaton ja toimintamalli mahdollisesti julkaistaan creative commons lisenssimallin alaisesti esimerkiksi Github:ssa. Opinnäytetyön toimintamalli ei ole täydellinen malli Data Lineagen käyttötavoista, vaan pitää sisällään vain rajallisen määrän attribuutteja, joilla voidaan todeta, että malli ja raportit toimivat.

Opinnäytetyöni rajaukset:

- Opinnäytetyö ei ota kantaa Data Governanceen, paitsi kun se risteää metadata hallinnan aluetta.
- Opinnäytetyössä ei kerrota tietovarastoinnin perusteista, mallintamisesta, suunnittelusta, käyttöönotosta tai ylläpidosta.
- Opinnäytetyössä ei oteta kantaa Data Lineage -malleihin ja näkymiin, jossa metadataa yhdistetään tietoelementtiin.

Opinnäytetyön seminaarin jälkeen kävi selvitetyksi, että tämä opinnäytetyö saattaa johtaa patenttihakemukseen. Tämän takia alkuperäisestä opinnäytetyötä on yksinkertaistettu huomattavasti, ja pohdintaosuudesta on jätetty pois johtopäätöksiä ja jatkokehitysmahdollisuuksia.

1.2 Käsitteet suomenkielellä

Tässä opinnäytetyössä pyrin käyttämään suomennoksia, jos sellaista on. Koska englanninkieliset termit ovat laajasti käytössä, käytän niitä sen sijaan että keksisin termeille suomenkieliset vastineet. Suomenkielisissä termeissä on ongelmia, koska viralliset suomenkieliset termeillä on jo muu merkitys (esim. suomenkielinen termi tietovirta viittaa suoratoistoon (YSA - Yleinen suomalainen asiasanasto 2019)).

Samoin Julkisen hallinnon tietohallinnon neuvottelukunta käyttää termiä tietovirta kuvaamaan informaation virtausta ja tietovuokaavio kuvaa ohjelmien rakentamisessa tarvittavaa

kaaviota. Kumpikaan ei ole hyvä termi kuvaamaan tietovirtaa kahden eri tietojärjestelmän välillä. (Julkisen hallinnon tietohallinnon neuvottelukunta 2019).

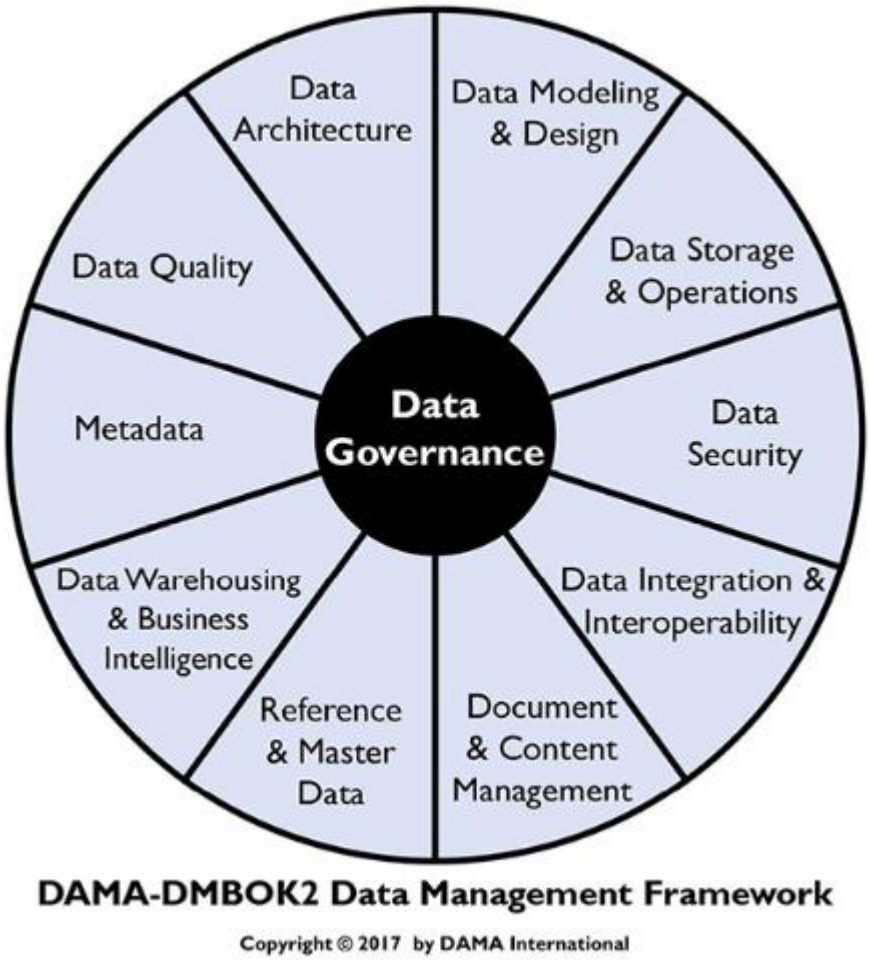
2 Data Governance

2.1 Kehys

Data Governancen määritelmän mukaan se on auktoriteetti ja valvoja kaikessa tietoon liittyvässä asioissa. Kaikki organisaatiot tekevät päätöksiä tietoon perustuen, riippumatta siitä onko heillä tunnistettua Data Governancea vai ei (DAMA International 2017, luku 3).

Data Governancea pidetään aikaa vievänä ja monimutkaisena tehtävänä, josta on organisaatiolle äärimmäisen vähän hyötyä pidemmällä aikavälillä. Kuitenkin CDO:n (engl. Chief Data Officer) rooli on yrityksissä yleistymässä, joka myös vaikuttaa siihen, että integroitu ja yhtenäinen tiedonhallinta on nousemassa datalähtöisissä organisaatioissa strategian keskeiseksi osatekijäksi (Ahlroth R. 2017).

Data Governance jakautuu 10 alueeseen (DAMA International 2017, luku 3):



Kuva 1. Data Management Framework (DAMA International 2017)

- Data Architecture on suunnitella, sijoittaa ja valvoa tietokantamallien, tietomääritelmien ja tietovirtojen eri tasoja.
- Tiedon mallintaminen ja suunnittelu (engl. Data Modeling & Design) on prosessi löytää, analysoida ja rajata tiedon vaatimuksia, ja myöhemmin esittää ja välittää nämä tiedon vaatimukset tarkassa muodossaan tiedon mallille. Tämä prosessi on iteratiivinen ja voi pitää sisällään konseptiivisen, loogisen ja fyysisen mallin.
- Tietokantojen tilanhallinta ja ylläpito (engl. Database Storage & Operations) on suunnitella, toteuttaa ja tukea, maksimoimalla tallennetun tiedon arvo.
- Data Security on määritellä, suunnitella, kehittää ja suorittaa tietoturvapolitiikkaa ja -menettelyjä tuottaakseen asianmukaiset luvat, todennukset, pääsyn ja tiedon valvonnan ja tiedot itsessään.
- Tiedon integrointi ja yhteen toimivuus (engl. Data Integration & Interoperability) on tiedon siirtoa ja keskittämistä ohjelmistojen ja organisaatioiden välillä.

- Dokumenttien ja sisällön hallinta (engl. Document and Content Management) on suunnitella, toteuttaa ja ohjata tiedon elinkaaren hallinnan toimintaa ja tietoa missä tahansa muodossa tai muotojen välissä.
- Viite- ja ydintieto (engl. Reference & Master Data) on yhteisen tiedon hallintaa saavuttaakseen organisatoriset tavoitteet, tietojen irrallisuuteen liittyvien riskien vähentäminen, saavuttaakseen paremman tiedon laadun ja vähentääkseen tiedon integroinnin kustannuksia.
- Data Warehousing & Business Intelligence Management on suunnitella, toteuttaa ja valvoa prosesseja tuottaakseen päätöksiä tukevaa tietoa ja tukea tietotyöläisiä raportoinnissa, kyselyissä ja analyyseissä.
- Metadatan hallinta (engl. Metadata Management on suunnitella, toteuttaa ja valvoa toimintaa mahdollistaakseen korkea laatuisen integroidun metadatan.
- Tiedon laadun hallinta (engl. Data Quality) on suunnitella, toteuttaa ja valvoa toimintoja, jotka lisäävät laadunhallintatekniikoita tietoon varmistaakseen, että se sopii käytettäväksi ja saavuttavat tiedon käyttäjien tarpeet. Tiedon laadulla on tunnistettu 118 erilaista attribuuttia, jotka on jaettu 20 eri ulottuvuuteen, jotka puolestaan on jaettu 4 pääulottuvuuteen.

  Tiedon laadun pääulottuvuudet ovat (Wang R., Strong D. & Sharpe M.E. 26.10.2013):
  - Luontainen tiedon laatu (engl. Intrinsic Data Quality)
  - Kontekstuaalinen tiedon laatu (engl. Contextual DQ)
  - Kuvaava tiedon laatu (engl. Representational DQ)
  - Saavutettava tiedon laatu (engl. Accessibility DQ)

3 Metadata Management

Yksinkertaisesti selitettynä metadata on tietoa tiedosta. Tämä voidaan esittää taulukon (kuva 2) avulla seuraavasti:

Etunimi	Sukunimi	Osoite	Kaupunki	Syntymävuosi
Sauli	Niinistö	Mariankatu 2	HELSINKI	1948
Janne	Laaksonen		VANTAA	1973
John	Doe	NULL	NULL	NULL

Metadata

Data

Kuva 2. Metadata esimerkki

Metadatalle on kuitenkin kilpailevia tai laajempia selityksiä:
Metadata on tietoa, joka kuvaa minkä tahansa yrityksen tietovarojen osan ja mahdollistaa organisaatiota käyttämään ja hallitsemaan näitä tietovaroja (Malcolm C. 07/2008).

Laajin selitys on, että metadata pitää sisällään teknisen ja liiketoiminnan prosessit, tieto säännöt ja rajoitukset ja loogiset ja fyysiset tietorakenteet. Se määrittää tiedon itsessään (esim. tietokannat, tiedon elementit ja tietomallit), käsitteet, joita tieto edustaa (liiketoiminnan prosessit, ohjelmistojärjestelmät, ohjelmointikoodin, teknisen infrastruktuurin) ja yhteydet tiedon ja struktuurin välillä (DAMA International 2017, luku 12).

Se mitä yhdelle on metadataa, on toiselle liiketoiminnallista tietoa (esim. puhelimen soitto-aika ja kesto ovat metadataa käyttäjälle, mutta liiketoiminnallista tietoa puhelinyhtiölle).

Metadataa voi kerätä monesta (Burbank, D. 2017):

- tietokantamalleista
- tietokannoista
- dokumenteista
- tiedon siirtämisestä (esim. json, xml)
- taulukoista
- media (kuvat, videot)
- sosiaalisesta mediasta (esim. Twitter, Facebook, Reddit)
- IoT
- ohjelmointi koodista (muuttujat)
- pilvestä tai vastaavista palveluista (esim. Azure, Sharepoint)
- open data

Metadata vastaa seuraaviin kysymyksiin (Burbank, D. 2017):

Kuka	Mikä	Minne	Miksi	Milloin	Miten
Kuka loi tämän tiedon?	Mikä on business metadata määritelmä tälle tiedolle?	Minne tieto on tallennettu?	Miksi tallennamme tätä tietoa?	Milloin tämä tieto luotiin?	Kuinka tämä tietoa on muodostettu?
Kuka hallinnoi tätä tietoa?	Mitkä on business säännöt tällä tiedolle?	Mistä tieto tulee?	Mikä on käyttötarkoitus?	Milloin tämä tieto viimeksi päivitettiin?	Kuinka monessa paikassa tämä yksittäinen tieto on tallennettu?
Kuka käyttää tätä tietoa?	Mikä on turvallisuusluokitus tai yksityisyyssäännöt tälle tiedolle?	Miten tietoa käytetään ja/tai viedään?	Mikä on business syyt tämän tiedon käytölle?	Kuinka pitkään tietoa tarvitaan?	
Kuka omistaa tämän tiedon?	Mikä on lyhenne tai akronyymi tälle tiedolle?	Missä on tiedon varmuuskopiot?		Milloin tieto pitää poistaa?	
Kuka säännöstelee tai auditoi tätä tietoa?	Mitkä ovat tekniset nimeämisstandardit?	Mitkä ovat lainsäädännöstä tulevat säännöt tälle tiedolle?			

Kuva 3. Metadatan kysymykset

Metadata jakautuu kolmeen pääluokkaan: Business metadataan, tekniseen metadataan ja operatiiviseen metadataan. Liite 6 kuvaa tietokannan mallinnuksen, arkkitehtuurin, teknisen metadatan ja business metadatan väliset riippuvaisuudet.

3.1 Business metadata

Business metadata kuvaa Business Glossary:n, säännöt ja kontekstin tiedolle (Inmon, W. H., Fryman, L., Inmon, W. & O'Neil, B. K. 2008)
Esimerkiksi ”liikevaihto” on yksiulotteinen metadata tieto, jolla on sama käyttötarkoitus ja sen laskentatapa on sama eri yrityksen osastoissa, ja se kuvaa yrityksen myyntimäärää.

Business Metadataa ovat (DAMA International 2017, luku 12):

- ei-tekniset nimet ja kuvaukset konsepteista, aihepiireistä, kokonaisuuksista ja attribuuteista

- ei-tekniset nimet ja kuvaukset attribuuttien tietotyypeistä ja muista tiedon attribuuttien ominaisuuksista, kuten tietojen välin kuvaukset, laskusäännöt ja algoritmit
- liiketoiminnan säännöt
- sallitut voimassa olevat arvot ja niiden kuvaukset.
- määritelmät ja kuvaukset tiedosta, tauluista ja kolumneista
- liiketoiminnan säännöt, muutossäännöt, laskusäännöt ja johtamissäännöt
- tietomallit
- tiedon laadun säännöt ja mittausten tulokset
- aikataulut, joilla tietoa päivitetään
- tietojen alkuperä (engl. Data Provenance) ja Data Lineage
- tiedon standardit
- tietoelementtien kirjaussäännöt

3.1.1 Business Glossary

Business Glossary on yksi Business Metadatan merkittävimmistä ulottuvuuksista. Tämä on kerätty sanasto, jota yrityksessä käytetään. Esimerkiksi sana ”asiakas” tarkoittaa eri asioita riippuen organisaatiosta: myyntiosastolla se tarkoittaa tuotteita ostavaa asiakasta, IT-osastolla asiakas on pääasiassa yrityksen oma henkilöstö ja tuotesuunnittelun asiakas voi olla puolestaan toinen organisaatio.

Business Glossary:n kerätään liiketoiminnan termien attribuutteja kuten:

- termien nimet, määritykset, akronyymit, lyhenteet ja synonyymit
- mikä organisaatio tai henkilö on vastuussa tiedon terminologiaan liittyvistä tiedoista
- termin kuvanneen henkilön nimi ja aika
- termin luokittelu
- ristiriitaiset määritykset termille, jotka vaativat ratkaisun tai kuvaavat ongelman
- yleiset väärinkäsitykset termille
- muutoshistoria
- viralliset tietolähteet termille

3.2 Tekninen metadata

Tekninen metadata (engl. Technical Metadata) kuvaa tiedon attribuutteja: tiedolle varatun tilan tyyppiä ja pituutta, null-sääntöjä, yksilöllisyyttä, hyväksyntä sääntöjä, luontietoja ja tiedon hallinta tietoa.
Tekninen metadata pitää sisällään (DAMA International 2017, luku 12):

- fyysisen tietokantamallin, mukaan lukien taulujen nimet, avaimet ja indeksit
- fyysisen taulun sarakkeiden nimet
- sarakkeiden ominaisuudet
- tietokannan objektien ominaisuudet
- tiedon CRUD (create, replace, update, delete) säännöt
- relaatiot tietokantamallin ja tietokantojen tiedostojen välillä
- lähdejärjestelmistä kohdejärjestelmään kohdistus dokumentaatio
- Data Lineagen dokumentaatio
- ohjelmien ja sovellusten nimet ja kuvaukset
- ETL töiden tiedot
- sisällön päivitys töiden aikataulut ja riippuvaisuudet
- palautusten ja varmistusten säännöt ja aikataulut
- käyttöoikeudet
- tiedon pääsyn oikeudet, ryhmät ja roolit

3.2.1 Data Lineage

Data Lineage on tekniikka tiedon virran kuvaamiseen verkostoituneissa tietokannoissa. Data Lineageta voidaan katsoa sekä loogisella tasolla, jolloin kuvataan ”As Designed Lineage”, että fyysisen tason kuvauksella, jolloin kuvataan ”As Implemented Lineage” (DAMA International 2017, luku 12).

Yksinkertaisesti Data Lineage on visualisointi teknisestä metadatasta. Tietokannoista kerätään tauluista metadataa, tiedon siirrossa käytetyistä koodista ja ETL:n aikana syntyneestä metadatasta. Tästä kaikesta muodostuu tiedon attribuuttitasolta katsoen ”putki” (liite 1), josta voidaan seurata yksittäisen attribuutin siirtymisestä järjestelmästä toiseen, sekä mahdolliset tiedon muokkaussäännöt matkan varrella.

Laajemmin selitettynä Data Lineage on dokumentaatio (liite 1), miten organisaation käyttämää tietoa muutetaan siirrettäessä (DAMA International 2017, luku 12). Tämä on tärkeää tiedon esimerkiksi tiedon avoimuuden kannalta, jolloin pystytään yksittäisellä visualisoinnilla esittämään mistä lähteistä kyseinen tieto on lähtöisin ja minkälaisia rajauksia tai muokkauksia tietoon on tehty sitä siirrettäessä.

Data Lineage -puuta voidaan tarkastella molemmista päistä. Raportointijärjestelmästä katsoen se näyttää, mistä järjestelmistä tietoattribuutti on lähtöisin, ja minkälaista muokkausta siihen on tehty matkan varrella. Lähdejärjestelmistä katsoen se näyttää, missä kaikissa järjestelmissä kyseistä tietoattribuuttia käytetään.

Data Lineage -tekniikalla voidaan vastata moneen sivulla 9 oleviin kysymyksiin. Lisäämällä operatiivista metadata tietoa (esim. käyttöoikeudet, kyselyiden käyttäjätiedot) Data Lineageen, voidaan vastata vielä useampaan kysymykseen.

Esimerkki: ”Mistä tieto tulee?”
Liitteen 1 esimerkki läpikäytynä: Eri OLTP-järjestelmien tietokantamallit voivat erota toisistaan, mutta jokaisessa käsitellään asiakastietoja järjestelmien omien vaatimusten pohjalta. ETL-prosessilla lähdejärjestelmistä haetaan customer-entiteetti, muokataan se Staging-tietokantamalliin ja viimeiseksi ladataan se Staging-tietokantaan.

Staging-tietokannasta tehdään ETL-prosessin kautta vertailu, mitkä rivit ovat uusia ja mitkä rivit on päivittyneitä verrattuna Data Market -tietokantaan (ellei tämä käy selville jo lähdedata metadatasta esim. viimeksi muokattu -attribuutista). Loogiset tietokantamallit Staging and Data Market -tietokannoissa ovat identtiset. ETL-prosessi lataa uudet rivit Data Market alueelle, ja päivittää päivittyneet rivit. Viimeisessä vaiheessa raportointityökalu lataa tiedon Data Market tietokannasta, ja luo siitä visualisoinnin.

Data Lineage on attribuutti-tason “putki”, josta esimerkissä näkyy mistä lähdejärjestelmistä ja millä laskusäännöillä raportoinnin etunimi ja sukunimi ovat tulleet.
Putkea voidaan myös lukea alhaalta-ylös. Tällöin näkyy, että missä järjestelmissä (ja raporteissa) Sales Database:n etunimi ja sukunimi -attribuutteja käytetään.

Erilaisia Data Lineage visualisointeja (Dennis A. 2017):

- Useampitasoisen Data Lineagen ymmärrys (engl. Understand multi-layer data lineage):
    - liiketoiminnan seurattavuus
    - kriittisen tietoelementin sukupuu (engl. Critical Data Element Lineage)
    - useampitasoinen tekninen sukupuu (engl. Multi-Layer Technical Lineage)
    - ongelmien hallinta ja löytäminen Data Lineagesta
    - pura ETL – ymmärrä muutoksen yksityiskohdat
    - analysoi ohjelmistojen muutokset
- Esitä visualisointi käyttötapausmuodoissa (engl. Present in Use-Case Related Formats)
    - liiketoiminnan seurattavuus
    - tietoelementin lineage (engl. Data Element Lineage Reports)
    - vaatimusten vienti (esim. projekteihin)

- snapshot Data Lineagesta – ymmärrä muutokset & seuraukset data lineagessa (engl. Lineage Snapshots)

3.2.2 Data Catalog

Data Catalog on osa teknistä metadata hallintoa, käytännössä luettelo tiedoista, jossa on tieto missä ko. tietoa säilytetään, sen käyttötarkoitus, mahdolliset akronyymit (Bond S. 2018).

Data Catalogilla on tarkoitus vastata seuraaviin kysymyksiin:

- Missä tieto on, ja mistä se tulee?
- Mitä se merkitsee?
- Kenellä on pääsy siihen?
- Koska tieto on viimeksi validoitu?
- Miksi kyseinen tieto on olemassa?
- Mikä on tiedon käyttötarkoitus?
- Mitkä ovat tiedon suhteet muuhun tietoon?

3.3 Operatiivinen metadata

Operatiivinen metadata kuvaa tiedon prosessointia ja tiedon käyttöä (DAMA International 2017, luku 12). Operatiivinen metadata on tiedon käytön yhteydessä muodostuvaa metadataa, eli käytännössä pääasiassa käyttölokeja ja aikataulutietoja. Näistä merkittävimpiä kerättävät ovat alkamisajat, päättymisajat, suorittavan käyttäjän tiedot, missä operaatio ajettiin, mitä operaatiossa ajettiin, virheilmoitukset.

- eräohjelmien työ lokit
- tiedon kyselyiden ja tulosten historia
- aikataulujen poikkeamat
- tarkastusten ja valvontamittausten tulokset
- virhelokit
- kyselyiden käyttäjätiedot (engl. query access patterns), säännöllisyys ja suoritusaika
- korjausten ja versioiden ylläpitosuunnitelmat ja toteutukset
- varmuuskopioit ja niiden luontipäivä ja säilyttämisaika sekä palauksia koskevat säännökset
- SLA vaatimukset ja määräykset
- tilavuus ja käyttökaavat

- tiedon arkistointi ja säilyttämisen säännöt ja niihin liittyvät arkistot
- tiedon puhdistamiskriteerit
- tiedon jakamisen säännöt ja sopimukset
- tekniset roolit ja vastuut, yhteydet

3.4 Metadatan Hallinta ydintietojen hallinnassa

Ydintieto (engl. Master Data) on organisaatiossa sovittua yhteistä tietoa, jotta saavutetaan organisaation kriittisen tiedon yhdenmukaisuus, oikeellisuus, hallinta, semanttisen oikeellisuus ja vastuut (DAMA International 2017, luku 10).

Ydintieto jakautuu kahteen alueeseen:

- Ydintietojen hallinta (engl. Master Data Management) on syntynyt tarpeesta saada eri järjestelmiin sirpaloituneet, moneen kertaan talletetut ja usein eri tasoilla olevat perustiedot parempaan hallintaan. MDM on pitkäikäistä, hitaasti muuttuvaa, monia yrityksen tai organisaation yksikköjä kiinnostavaa tietoa; ikään kuin ”perusrekistereitä”, kuten tuote- ja asiakastiedot (Hovi A. 2018).
- Viitetietojen hallinta (engl. Reference Data Management) on määrättyjen arvojen ja niiden määritysten valvontaa. RDM:n tarkoitus on varmistaa, että organisaatiolla on pääsy tarkkaan ja ajantasaiseen arvoihin jokaisen edustetun käsitteen osalta. Kuten muukin tietojenhallinta, myös RDM vaatii metadatan hallintaa. Yksi tärkeimmistä metadatan attribuutteista RDM:lle pitää sisällään tiedon lähteen. (DAMA International 2017, luku 10).

3.5 Metadatan hallinta tiedon integroinnissa ja yhteen toimivuudessa

ETL (liite 4) tarkoittaa datan siirtämistä ja muokkaamista ja lataamista: tiedot haetaan (engl. Extract) lähdejärjestelmästä, niitä muokataan (engl. Transform) ja ladataan (engl. Load) lopulta tietovarastoon (Hovi A. 2018).

ETL luomisprosessi vaatii ja luo metadataa. ETL:ssä käytettävä koodi on käytännössä SQL:llä (tai vastaavalla strukturoidun tiedon käsittelykielellä), joka on teknistä metadataa itsessään. Kun koodi ajetaan, muodostuu operatiivista metadataa lokeihin.
Nämä metadatat on syytä hallita läpi tiedon elinkaaren: tiedon löytämisestä käytön hallintaan saakka. Mitä täydellisempää ja tarkempaa on organisaation metadatan hallinta, sen parempi on organisaation kyky hallita riskejä ja kustannuksia, jotka liittyvät tiedon integraatioon ja yhteen toimivuuteen (DAMA International 2017, luku 10).

Data Lineagea voidaan myös käyttää ETL pakettien ajastukseen kokonaisvaltaisesti. Metadatasta näkyy pakettien aloitusajat ja lopetusajat, kesto ja odotusajat voidaan laskea. Tällä voidaan saavuttaa nopeampi tiedon virtaus lähdejärjestelmistä raportointijärjestelmiin. Ja yhdistämällä operatiivista metadataa voidaan myös seurata ELT prosessien toimivuutta tiedon attribuutin näkökulmasta katsoen.

3.6 Metadatan hallinta tiedon laadun hallinnassa

Metadata on tärkeää tietoelementtien laadun hallinnassa. Tiedon laatu perustuu kuinka hyvin se täyttää tiedon käyttäjän vaatimukset. Metadata määrittää miten tieto on rakennettu.

Yrityksellä tulee olla prosessi, jolla tiedot määritellään. Tämä tukee organisaation kykyä muodostaa ja dokumentoida tiedon standardit ja vaatimukset, jolla tiedon laatua voidaan mitata. (DAMA International 2017, luku 13).

Tiedon laadun profilointi antaa tilastollisen näkyvyyden tiedon rakenteelle, sisällölle ja tiedon laadulle. Näistä yhteneväisiä attribuutteja teknisen metadatan kanssa ovat muun muassa:

- null-arvojen määrä
- tiedon minimi ja maksimi pituudet
- tiedon tyyppi ja muoto (engl. pattern)
- uniikkien arvojen prosenttimäärä

Vielä suuremman yhteyden luo Data Lineage tiedon laadun korjaamisessa. Data Lineagen avulla saa selville missä kaikkialla kyseistä attribuuttia tai entiteettiä käytetään verkostoituneissa järjestelmissä, selvittäen attribuutin tai entiteetin lähteet. Ja jos tiedon laadun ongelma korjataan lähteessä, niin samasta Data Lineagesta saadaan selville, missä kaikkialla kyseistä attribuuttia tai entiteettiä käytetään. Tämä on selkeä etu muutoksenhallinnassa, jolloin muutoksenhallinta voidaan suoraan kohdistaa jokaiseen kohtaan suunnitellusti, ja saavutetaan laadullisesti korkeampitasoista työtä.

4 Toimintamalli

4.1 Opinnäytetyön tavoite

Toimintamallin tavoite on toteuttaa malli, jossa näkyy mitä metadata attribuutteja pitäisi kerätä mistäkin kerroksesta, luoda looginen tietokantamalli, johon tämä tieto voidaan tallentaa.
Lisäksi tehdä muutama visualisointi ja testata, että malli toimii ja mitä lähteä kehittämään.

4.2 Opinnäytetyön toteutus

Opinnäytetyö toimintamalli pitäisi vastata kysymyksiin:

- mistä tietoa kerätään?
- mitä tietoa kerätään?
- minne tieto tallennetaan?
- miten kerättyä tietoa käytetään?

4.3 Opinnäytetyön tuotos

Liite 2 eri tasoista on tarkoitus kerätä erilaisia asioita tavoitteen saavutettavaksi.

Data Table kerroksesta voidaan kerätä seuraavia asioita:

- Taulujen metadatatiedot
    - nimi
    - schema
    - taulun ominaisuudet
    - storage
    - sarakkaiden nimet
    - sarakkaiden tietotyyppi ja tietotyypin pituus
    - sarakkaiden null-arvot sallittuja -sääntö
    - sarakkaiden tietotyypin ominaisuudet
    - Primary Key
    - Constraint
    - Määritelmä
    - kommentit, muistiinpanot
- Indexien metadatatiedot
    - nimi

    - indexin ominaisuudet (unique, clusterointi)
    - storage
    - sarakkeet nimet
    - sarakkaiden tietotyypit ja tietotyyppien pituus
    - sarakkaiden null-arvot sallittuja -sääntö
    - sarakkaiden tietotyypin ominaisuudet
    - Määritelmä
    - kommentit / muistiinpanot
- Foreign Keys
    - Nimi
    - Parent Key
    - Parent Entity
    - Parent Columns
    - Child Entity
    - Child Columns
    - Relationship type
    - Existance
    - Cardinality
    - FK muut ominaisuudet (nocheck, not for replication)
    - Triggers
    - Määritelmä
    - kommentit / muistiinpanot

Data Flow kerroksesta voidaan kerätä seuraavat tiedot:

- ETL paketin nimi
- käyttäjätunnus, jolla ajetaan
- Ajastusaika
- Logi (kesto, loppustatus)

Metadata Field kerroksesta voidaan kerätä seuraavat tiedot:

- Extract rules
    - select conditions per attribute
    - from per attribute
- Transmutation rules
    - format changes
    - structure changes
    - semantic conversions

    - de-dupping
    - re-ordering
- Load rules
    - insert/update conditions per attribute
    - from per attribute

Avoimia asioita mistä haetaan:

- Tiedon luokitus
- Tallennuspaikkojen luokitus
- Tiedon omistaja (säännöt? lähde? attribuutti?)

Data Lineage kerros on missä edellisistä kerroksista luodut tiedot yhdistetään samaan tietokantamalliin (liite 5), josta voidaan rakentaa raportointijärjestämällä visualisointi.

4.4 Tietokantamalli

Katso liite 5.

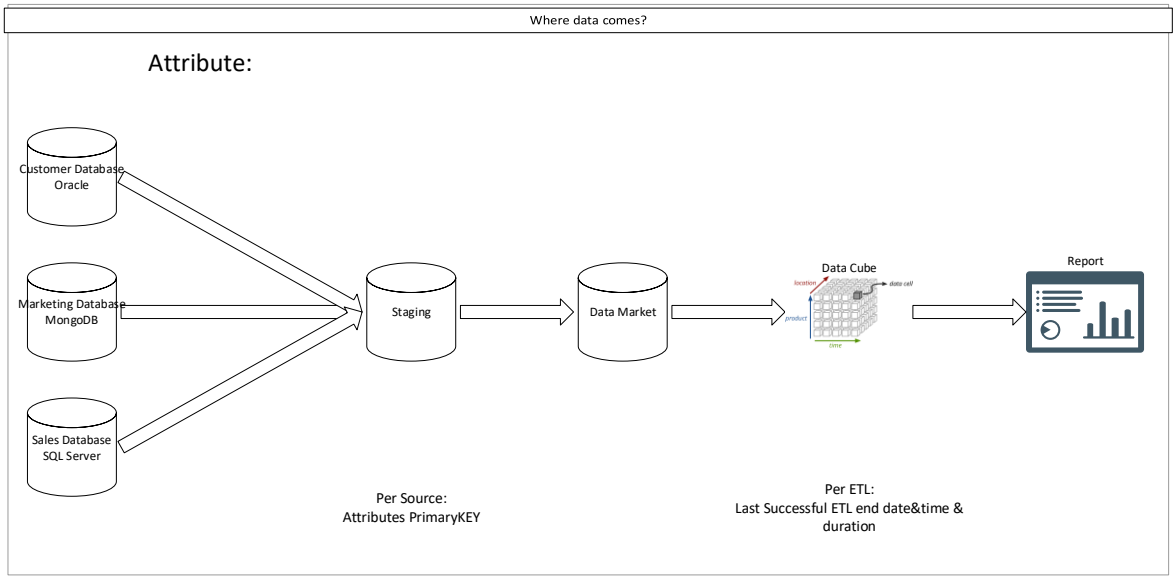
4.5 Data Lineage Visualisointeja

Ideoida metadata kysymysten vastauksiksi erilaisia raporttipohjia, jotka pystyisivät vastaamaan yhteen tai useampaan kysymykseen. Jaoin kysymykset viiteen luokkaan.

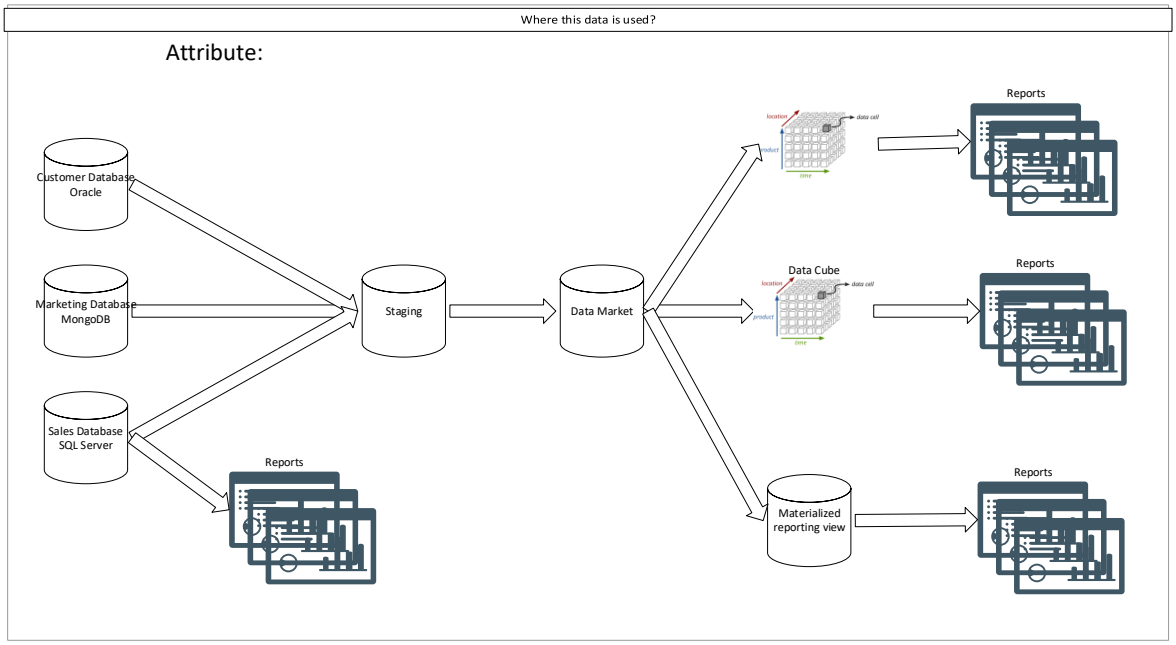
4.5.1 Tieto attribuutin seurattavuus

Tieto attribuutin seurattavuudella seurataan yksittäistä attribuuttia, ja sen muutoksia eri järjestelmissä. Avoimuuden nimissä, kun esitellään uutta raporttia, pitäisi samalla esitellä raporttia mistä järjestelmistä kyseinen tieto on kerätty (ja millä säännöillä).

- mistä tieto tulee (kuva 4)?
  Tiedon lähteiden selvittämisen lisäksi on oleellista nähdä aikajana, milloin tieto on haettu ja viimeksi päivitetty.
- kuinka monessa paikassa tätä tietoa käytetään (kuva 5)?
- kuka omistaa tämän tiedon?
  Edellistä raporttia voidaan käyttää myös tiedon omistajuuden määrittelyyn. Periaatteellinen kysymys onkin, että onko tiedolla omistaja per lähdejärjestelmä, vai omistaako joku tietyn tyyppisen tiedon (tuotetiedot, asiakastiedot?), vai jonkinlainen yhdistelmä (omistaja määräytyy sen perusteella, missä tieto on luotu).
- missä ovat tiedon varmuuskopioit?
  Edelleen edellistä raporttia voidaan käyttää myös tiedon varmuuskopioiden tutkimiseen, varmuuskopioiden kattavuuteen ja varmuuskopioiden vanhentumistietojen tarkistamiseen. Tämän tyyppinen visualisointi antaa



Kuva 4. Esimerkki Data Lineage visualisoinnista



Kuva 5. Esimerkki Data Lineage visualisoinnista

4.5.2 Tiedon virtauksen seuranta

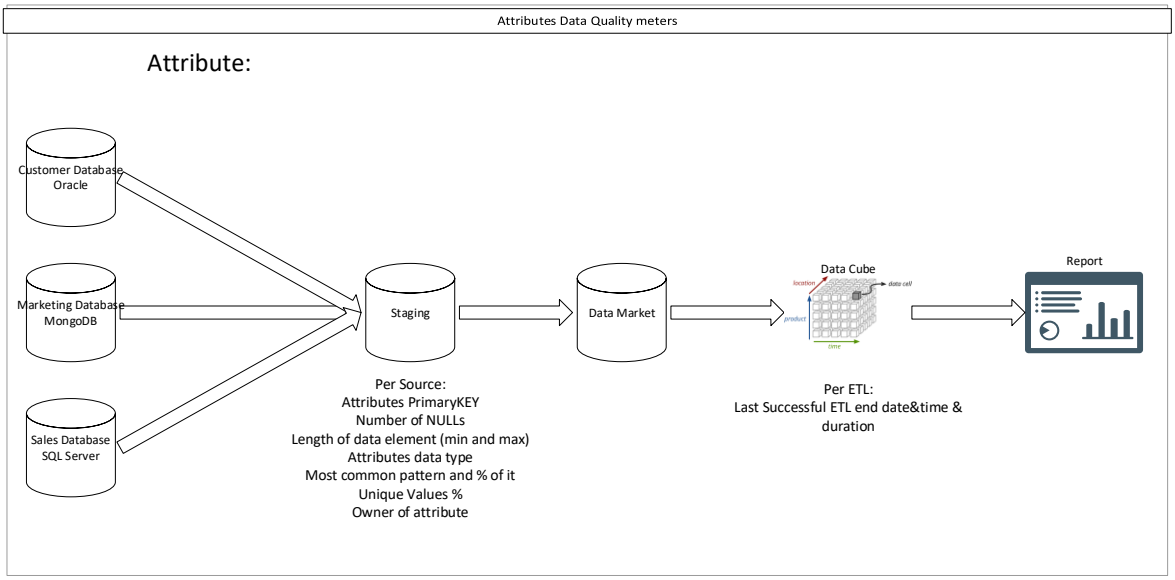
- Mitä kaikkea tietoa tämä serveri välittää?
- Miten tiedon virtauksen aikataulutus on toteutunut?

4.5.3 Yhdistelmiä

- Toteutuuko tiedon luokitus (minne tallennettu/luokitus / tiedon luokitus)

4.5.4 Tiedon laadun seuranta

Tiedon laadun selvittämisessä pystytään keräämään kappaleessa 3.6 määriteltyjä teknisiä tiedon laadun mittareita (kuva 6).



Kuva 6. Esimerkki Data Lineage visualisointi yhdistetty tiedon laadun seurantaan

5 Pohdinta

Data Lineage tekniikan käyttöönoton kannalta on muutama merkittävä tekijä, mitä pitää ottaa huomioon tehdessä data arkkitehtuuria. Jokaiselle attribuutille pitäisi löytää vähintään:

- kuka on luonut tietoelementin?
- koska tietoelementti on luotu?
- koska tietoelementti on viimeksi päivitetty?

Data Lineage tekniikkana luo selkeän visualisoinnin, miten tietoattribuuttia käytetään monimutkaisissa, verkostoituneissa tietojärjestelmissä. Käyttämäni visualisointi esimerkit ovat yksinkertaistettuja ja oikeassa käyttötapauksissa yksittäinen puu voi olla hyvinkin pitkä ja monihaarainen. Varsinkin organisaatoissa, jossa käytetään hallitusti Master Dataa, syntyy visualisoituja silmukoita. Haasteellisuutta muodostaa myös neuroverkkojen avulla rakennetut raportit ja analyysit, ja niissä tietoattribuuttien seuranta. Neuroverkoista voidaan kuitenkin nähdä mitä attribuutteja ja millä ehdoilla niitä käytetään, mutta ei miten visualisointiin päästään.

5.1 Avoimia kysymyksiä/jatkokehitys mahdollisuuksia.

Tuottaa MVP tuote opinnäytetyöstä. Tässä pitää lähteä selkeästi liikkeelle ehkä vain parilla tietokantayhteensopivuudella, ja parilla ETL-tuote yhteensopivuudella. Opinnäytetyö määrittää mitä pitää kerätä, joten vaatimukset ovat kunnossa siltä osin.
Samoin kerätyn tiedon käyttö vaatii fyysisen tietokantamallin johtamisen (sekä mahdollisen raportointikuution luomisen) ja visualisoinnin rakentamisen raportointityökaluun (en usko, että löytyy valmiita raporttipohjia, vaan nekin pitää tuottaa).

Teknisen metadatan kerääminen ja tutkiminen, varsinkin mahdollisuus rakentaa data lineage ohjelmointikoodista.

Some-alustojen kehittäminen, miten Data Lineage -tekniikalla voisi nähdä mihin kaikkeen omaa tietoasi käytetään.

Tietoelementin seurattavuuden rakentaminen Data Lineageen. Tämä suoraa jatkokehitystä tietoattribuutin Data Lineagelle, koska tietoelementin Data Lineage käyttää hyväkseen suoraan attribuuttitason Data Lineagea. Tietoelementin Data Lineagessa on käytännön ongelmia, koska yksittäisen tietoelementin yhdistäminen järjestelmästä toiseen on varsin

haasteellista, jos ETL:ssä tapahtuu muutoksia kyseisessä tietoelementissä. Samoin tiedon laadun näkökulmasta, ei-uniikkien tietoelementtien seuraaminen luo haastetta lisää.

6 Sanasto

attribuutti	loogisen mallinnuksen rivi, vastaa fyysisen mallinnuksen sarake (ks. liite 6)
CRUD	tulee sanoista create/read/update/delete
entiteetti	looginen mallinnuksen kokonaisuus, vastaa fyysisen mallinnuksen taulua (ks. liite 6)
ETL	tulee sanoista extract/transform/load. Katso kappale 3.5.
Kontekstuaalinen tiedon laatu (engl. Contextual DQ)	kuvaa tiedon laadun merkityksellisyyttä, täydellisyyttä ja ajantasaisuutta
Kuvaava tiedon laatu (engl. Representational DQ)	kuvaa tiedon laadun tulkittavuutta, helppoa ymmärtävyyttä ja johdonmukaisuutta
Luontainen tiedon laatu (engl. Intrinsic Data Quality)	kuvaa tiedon laadun tarkkuutta, objektiivisuutta ja uskottavuutta
OLAP	tulee sanoista online analysis processing. Tietokanta, joka on suunniteltu tiedon analysointiin ja raportointiin.
OLTP	tulee sanoista online transaction processing. Tietokanta, joka on suunniteltu tapahtumien käsittelyyn
Saavutettava tiedon laatu (engl. Accessibility DQ)	kuvaa tiedon laadun saatavuutta ja käyttöturvallisuutta
sarake	fyysisen tietokannan taulun rivin otsikko. Toimii tarvittaessa viiteavaimena taulujen välillä.
staging	kohdejärjestelmän tietokantaan kopioidaan taulut samassa fyysisessä mallissa, jolla saavutetaan se etu, että tiedon lataus (extract) ja tiedon muutos (transform) pystytään eriyttämään ja ongelmatilanteet on helpompi ratkaista
taulu	fyysinen tietokantataulu, fyysisen mallinnuksen taulu. Sisältää yhden tai useamman sarakkeen.
tietoelementti	tiedon atomitaso, yksittäinen arvo.
tietovirta	tieto virtaa järjestelmästä toiseen, siirtyen ETL:n avulla
tietue	yksittäinen rivi, jokainen sarake pitää sisällään tietoelementin.
visualisointi	jonkin asian tekemistä havainnolliseksi näköaistille.

6.1 Lähteet

Ahlroth, R. 2017. Tiedonhallinta voi olla tehostaja – eikä hidaste. Blogi. Riku Ahlroth. Luettavissa: http://www.alykassuomi.fi/2017/09/data-governance-voi-olla-tehostaja-eika-hidaste/ Luettu: 14.2.2019

Bond, S. International Data Corporation (IDC). IDC PlanScape: Data Intelligence Software for Data Governance. Luettavissa: Data%20Intelligence%20software%20for%20data%20governance.pdf

Burbank, D. 2017. Data Modeling & Metadata Management. Global Data Strategy Ltd. Luettavissa: https://www.youtube.com/watch?v=jyva44uHoR4
Luettu: 18.1.2019.

DAMA International. 2017. DAMA-DMBOK: Data Management Body of Knowledge (2nd edition). Technics Publications, Basking Ridge, New Jersey.

Dennis A. 2017. Data Lineage and Data Quality: Two Vital Elements for Enterprice Success. Luettavissa: https://www.dataversity.net/data-lineage-data-quality-two-vital-elements-enterprise-success/# Luettu: 28.2.2019.

Finto Suomalainen asiasanasto- ja ontologiapalvelu. YSA – Yleinen suomalainen asiasanasto. Luettavissa: https://finto.fi/ysa/fi/ Luettu 17.1.2019.

Hovi, A. 2018. Data-alan termien selitykset ja kuvaukset. Blogi. Ari Hovi. Luettavissa: https://www.arihovi.com/3274-2/. Luettu: 17.1.2019.

Inmon, W. H., Fryman, L., Inmon, W. & O'Neil, B. K. 2008. Business Metadata: Capturing Enterprise Knowledge. Burlington: Morgan Kaufmann.

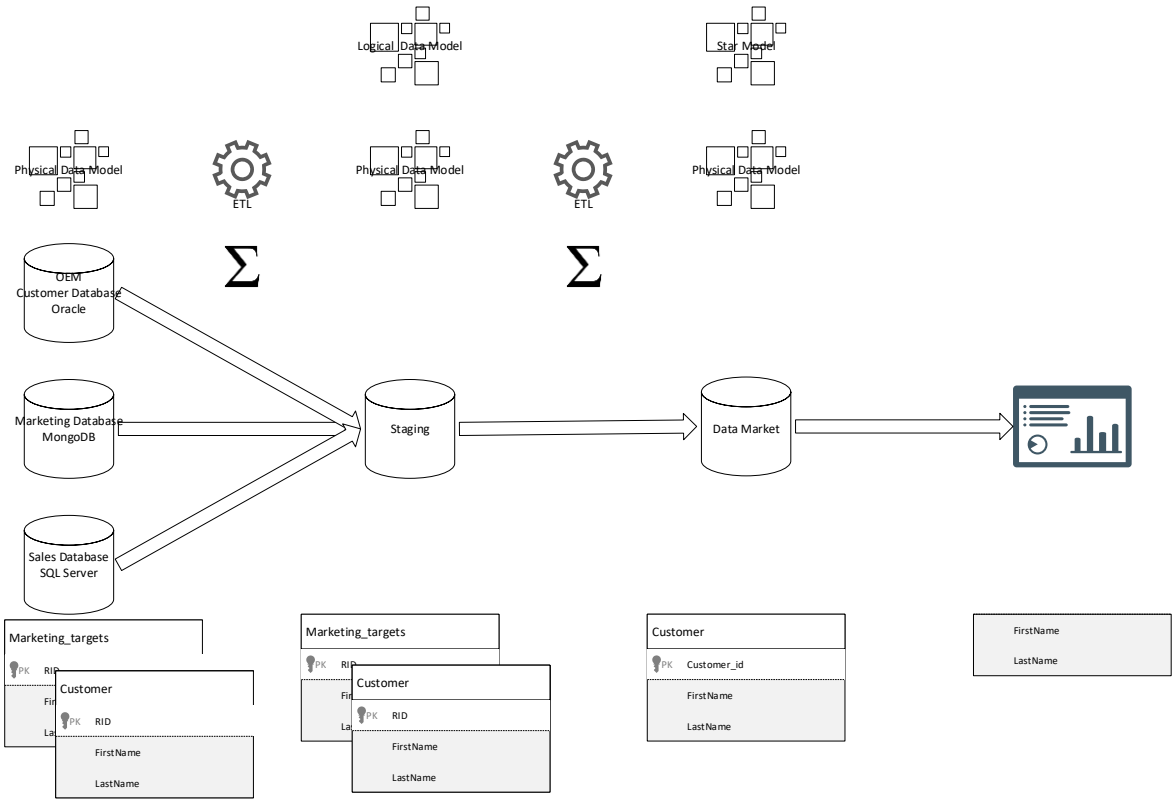
Julkisen hallinnon tietohallinnon neuvottelukunta 2019. JHS-sanasto. Luettavissa: http://jhs-sanasto.jhs-suositukset.fi/JHS/fi/page/c_0be84035 Luettu: 17.1.2019.

Malcolm C. 2008. Metadata is Master Data. Luettavissa: https://search.proquest.com/openview/c684b07cd243356a28503659ecadc2ee/1?pq-origsite=gscholar&cbl=51938. Luettu. 28.2.2019.

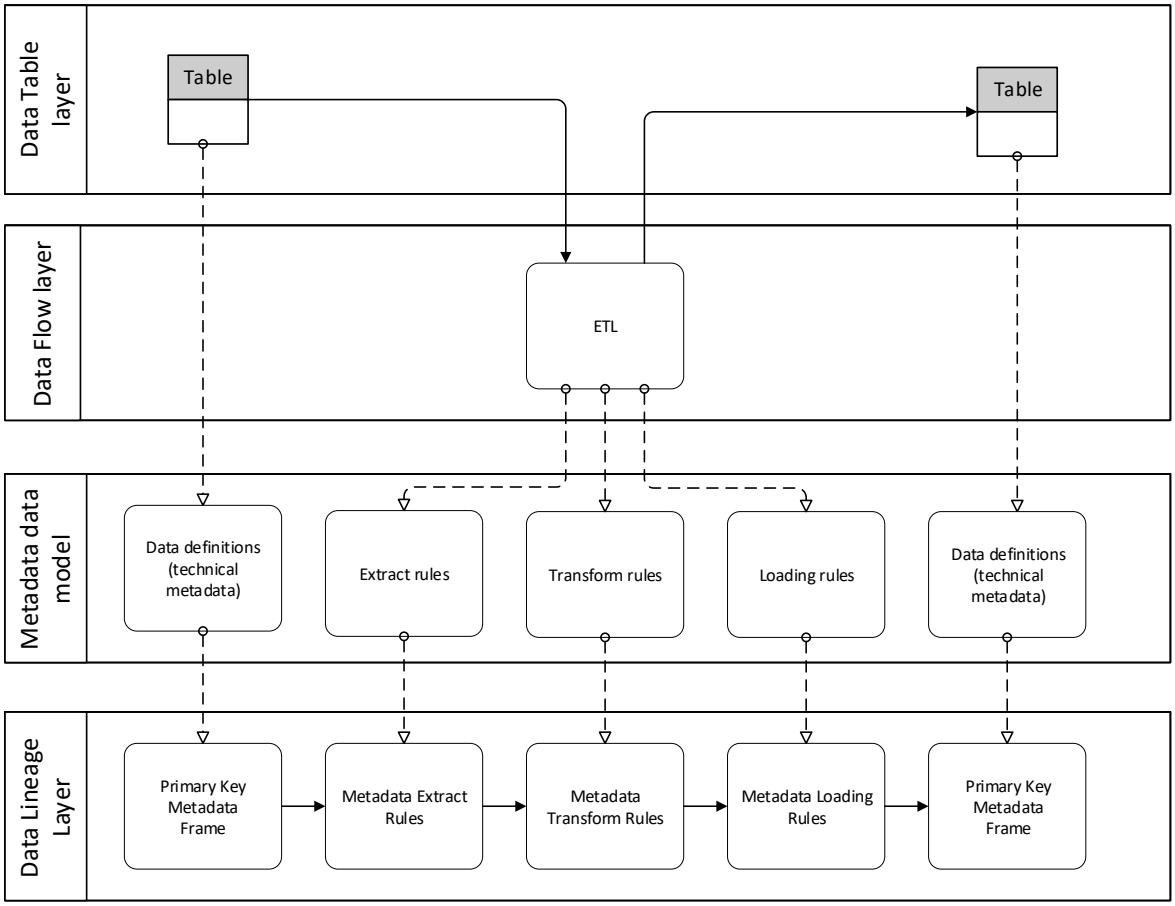
Wang R., Strong D. & Sharpe M.E. 2013. Beyond Accuracy: What Data Quality Means to Data Consumers. Luettavissa: http://courses.washington.edu/geog482/resource/14_Beyond_Accuracy.pdf Luettu: 20.5.2019.

7 Liitteet

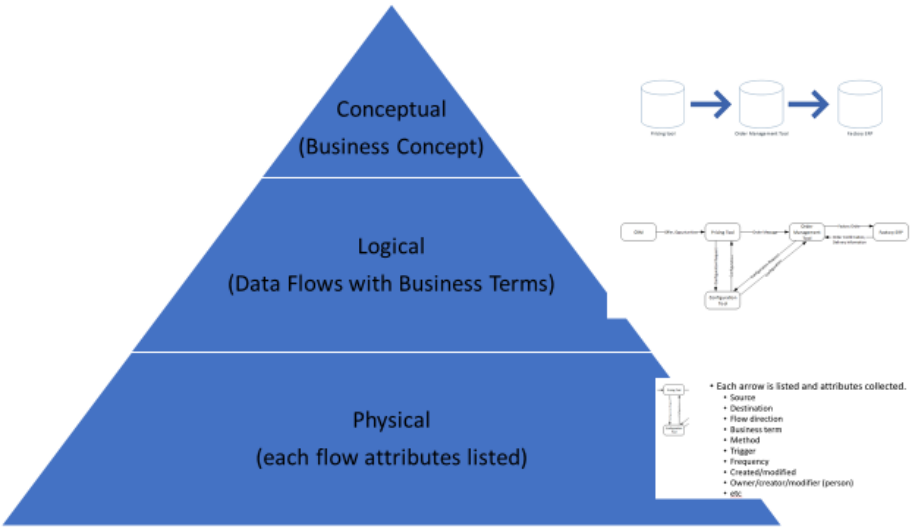
7.1 Liite 1. Data Lineage esimerkki 1.



7.2 Liite 2. Data Lineage Kerrokset



7.3 Liite 3. Tietovirtauksen tasot.



7.4 Liite 4. ETL Process Flow.

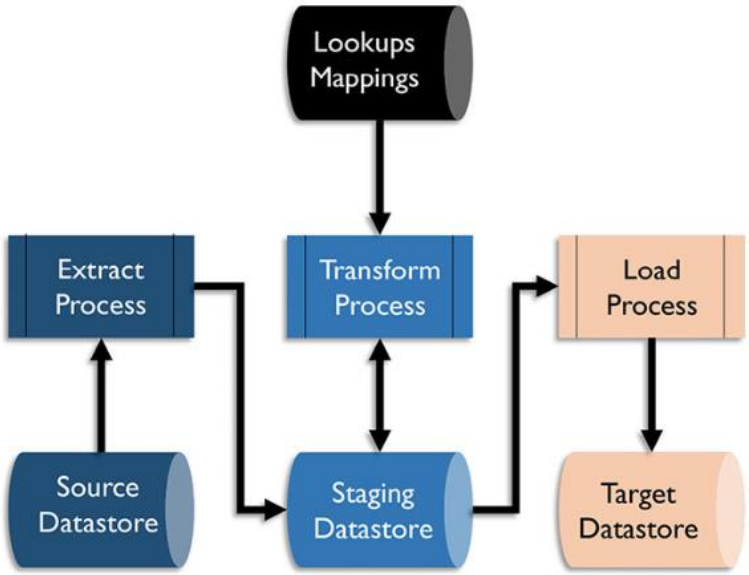
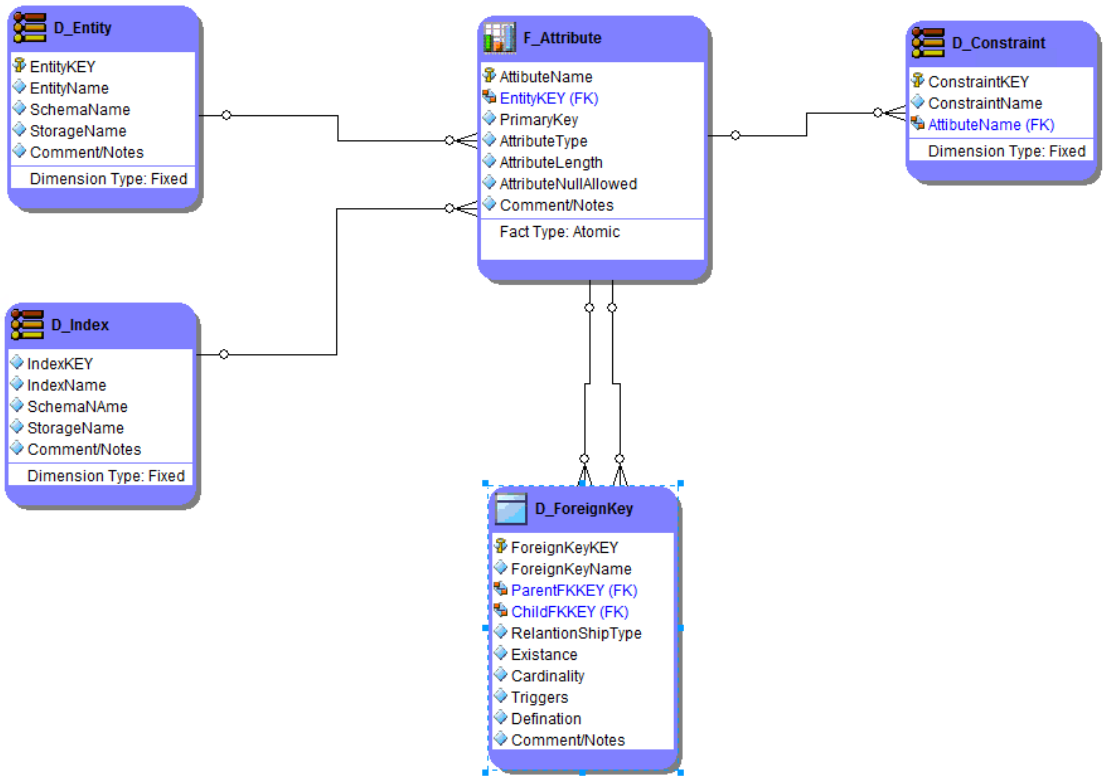


Figure 67 ETL Process Flow

lähde: (DAMA International 2017, luku 8)

7.5 Liite 5. Tietokantamalli.



7.6 Liite 6. Metadata Repository Metamodel

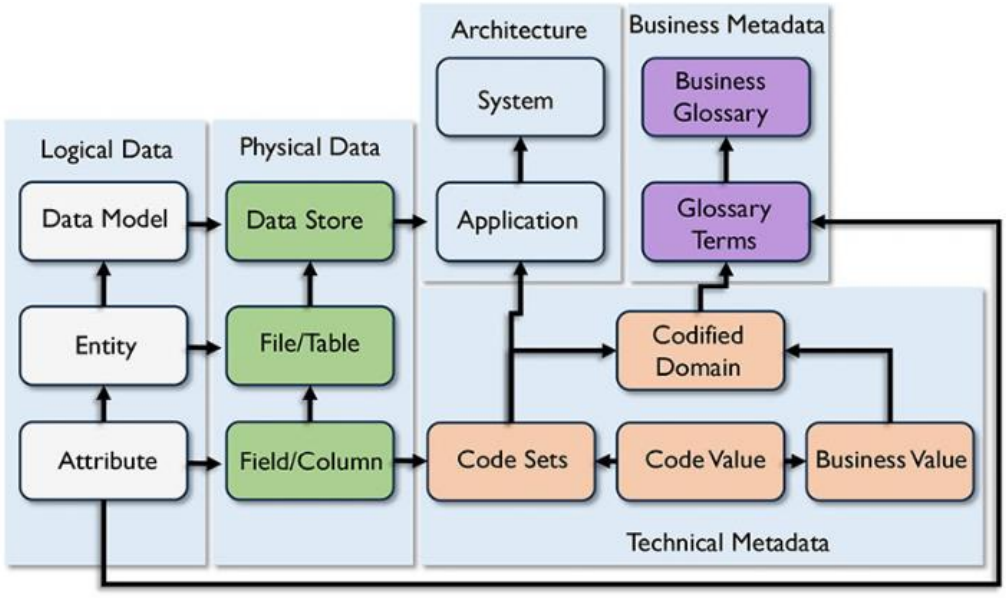


Figure 88 Example Metadata Repository Metamodel

lähde: (DAMA International 2017, luku 12)