Tieto-osaava

TP2. Tiedon varastointi ja mallintaminen

Mitä tietovarastolla tarkoitetaan

- *Tietovarasto on keskitetty ja organisoitu kokoelma tietoa, joka on tallennettu ja saatavilla yhdessä paikassa. Se voi sisältää monenlaisia tärkeitä tietoja organisaation toiminnasta. Tietovarasto on suunniteltu helpottamaan tietojen hallintaa, analysointia ja käyttöä.*

- *Tietovarasto kerää tietoa eri lähteistä, kuten operatiivisista järjestelmistä, tietokannoista, ulkoisista tiedonlähteistä ja muista tietolähteistä. Tämä tieto integroidaan ja muunnetaan usein - muttei aina - yhtenäiseen muotoon, jotta sitä voidaan helpommin käsitellä ja analysoida.*

- *Tietovarastojen tarkoitus on parantaa päätöksentekoa ja liiketoiminnan ymmärtämistä. Tietovarastosta tuotetut jalostetut analyysit, listaukset ja visuaalisoinnit voivat tuoda tähän avun.*

Tietovarastoinnin tavoista ja termeistä

- ***Relaatiotietovarasto*** *on tietovarasto, joka käyttää relaatiotietokantatekniikkaa, ja sen avulla organisaatiot voivat tallentaa ja analysoida tietoa suhteellisen jäsennellyssä ja tehokkaassa muodossa. Tietovarastossa oleva data on tallennettu relaatiotietokantojen tauluihin ja sen käsittelyyn ja kyselyihin voidaan käyttää esim. SQL (Structured Query Language) -kieltä.*

- ***Tietoallas*** *(data pool tai data lake) on joustava tietovarasto, joka voi tallentaa erilaisia raakadataa ilman ennalta määriteltyä rakennetta. Se mahdollistaa suurien tietomäärien keräämisen ja myöhemmin päätöksen siitä, mitä dataa analysoidaan ja käytetään. Useimmiten käyttötarkoitukset liittyvät esim. sensoridataan, ostotapahtumadataan jne.*

-

Inmonin ja Kimballin tietovarastofilosofiat

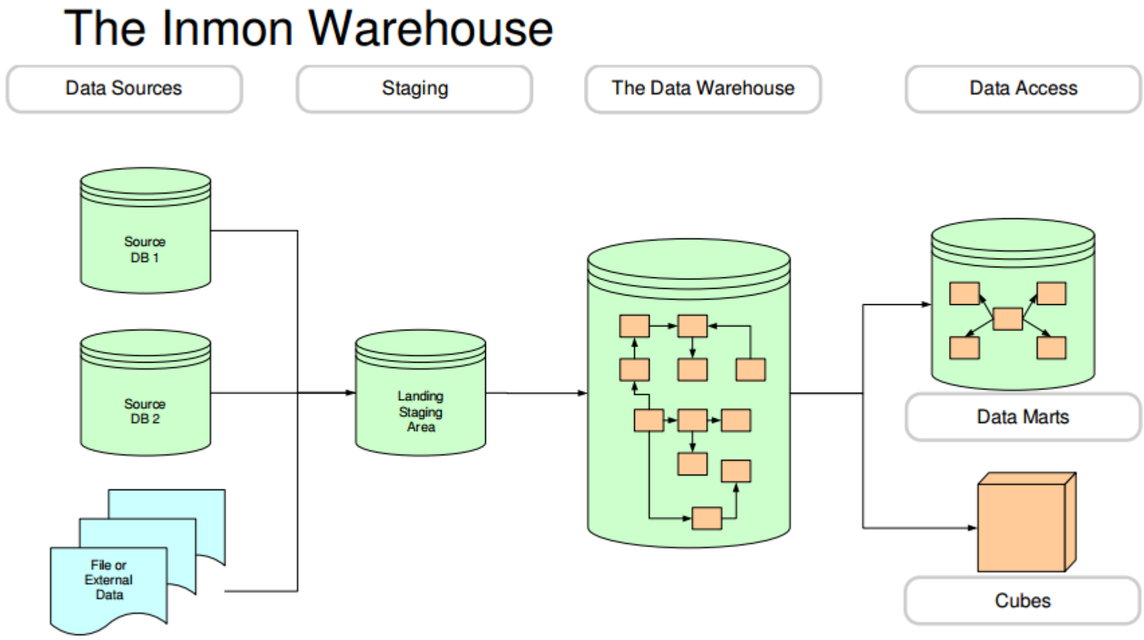
- *Bill Inmon alkoi kehittää 1980-luvulla tietovaraston keskitettyyn ja normaaliin muotoon perustuvaa arkkitehtuurimallia. Inmon korostaa tietojen normaalin muotoon muuttamista (tunnetaan myös nimellä normalisointi) ja hänen ajatus eriyttää operatiiviset tietokannat ja tietovarastot. Hänen mallinsa painottaa tietojen yhtenäisyyttä ja keskitetyn tietovaraston merkitystä.*

- *Ralph Kimball tuli tietoisuuteen 1990-luvulla omalla dimensiohierarkiapohjaisella mallillaan. Hän korosti tietovaraston rakentamista liiketoiminnan tarpeiden pohjalta, ja hänen mallinsa keskittyy enemmän tietovaraston nopeaan toteuttamiseen ja joustavuuteen. Kimballin ajattelu on ollut erityisen suosittu käytännönläheisenä lähestymistapana.*

- *Nämä kaksi mallia ovat kehittyneet vuosien varrella, ja on olemassa myös monia yhdistettyjä ja sovellettuja lähestymistapoja. Nämä kaksi tapaa ajatella tietovarastointia on syytä tuntea, kun tietovarastoa alkaa suunnittemaan ja toteuttamaan.*

Inmonin malli

- *Inmonin tietovarastomalli korostaa keskitettyä ja normalisoitua tietorakennetta. Mallissa data tallennetaan yhtenäiseen muotoon ja jaetaan tieto martteihin. Se korostaa tietojen yhtenäisyyttä, mutta voi olla monimutkaisempi ja vaatii enemmän aikaa suunnitteluun. Inmonin malli on perinteisempi ja keskittyy tietovaraston rakentamiseen yhtenä kokonaisuutena.*



Kuvan lähde: https://arjunjune.wordpress.com/2017/03/20/bill-inmon-vs-ralph-kimball/

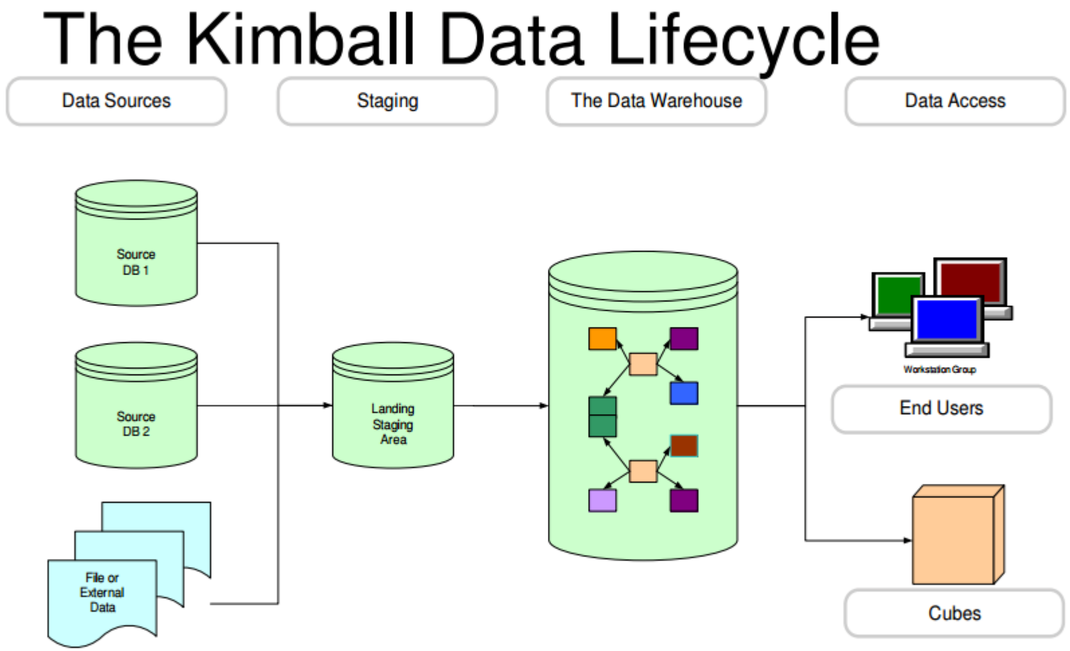
Inmonin mallin ominaisuuksista

- *Tietovarasto on erittäin joustava muutoksiin.*
- *Liiketoimintaprosessit voidaan ymmärtää hyvin helposti.*
- *Eri liiketoimintaprosesseille on helppo rakentaa omia data martteja keskitetyn tiedon yhteyteen, joita on sitten helpohko hallita omina kokonaisuuksinaan.*
- *ETL-prosessi (Extract, Transform, Load) on huomattavasti vähemmän altis virheille.*

Kimballin malli

- *Dimensiohierarkiapohjainen malli. Mallissa korostetaan tietovaraston rakentamista liiketoiminnan tarpeiden pohjalta. Malli keskittyy enemmän tietovaraston nopeaan toteuttamiseen ja joustavuuteen. Kimballin ajattelu on ollut erityisen suosittu käytännönläheisenä lähestymistapana.*

- *Jossain esimerkeissä Data Martit eli tietokokonaisuudet piirretään tietovaraston etupuolelle. Oleellisinta on kuitenkin se, että ne ovat erillisiä.*



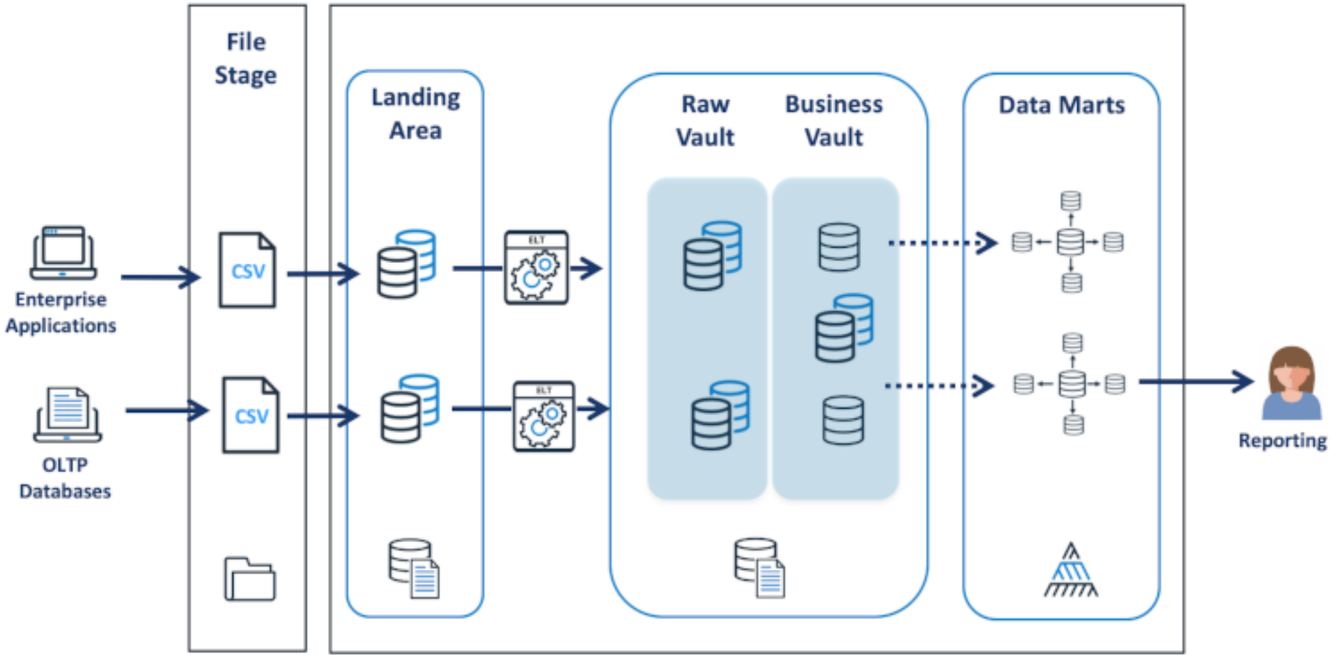
Kuvan lähde: https://arjunjune.wordpress.com/2017/03/20/bill-inmon-vs-ralph-kimball/

Kimballin mallin ominaisuuksista

- *Kimballin tietovarastomalli korostaa dimensioiden, kuten aika, tuote ja asiakas, tärkeyttä. Se käyttää tähti- ja lumipallokuviota, yksinkertaisia taulukoita ja keskittyy käytännöllisyyteen. Mallissa tiedot jaetaan ymmärrettäviin osiin, mikä tekee kyselystä tehokasta. Kimballin malli keskittyy liiketoiminnan tarpeisiin ja helppoon ymmärrykseen.*
- *Asentaminen ja aloittaminen on nopeaa*
- *Raporttien luominen tähtiskeemojen avulla on helppoa.*
- *Tietokantatoiminnot ovat erittäin tehokkaita.*
- *Vie vähemmän tilaa tietokannassa ja hallinta on helppoa.*

Data Vault 2.0

- *Dan Linstedtin kehittämä 2000-luvun alkupuolella.*
- *Data Vault -malli sisältää yksilöllisesti linkitetyn joukon normalisoituja taulukoita, jotka tukevat yhtä tai useampaa liiketoiminta-aluetta. Mallia voi pitää jossain määrin Kimbalin ja Inmonin mallien hybridinä.*
- *Lisätietoja: https://en.wikipedia.org/wiki/Data_vault_modeling*



Kuvan lähde: https://www.analytics.today/blog/when-should-i-use-data-vault

Data Vault 2.0 -mallin ominaisuuksista

- *Mahdollistaa ketterän kehittämisen. Voi ensin tuoda osan datalähteistä ja myöhemmin helposti lisätä uusia.*

- *Koska kaiken tiedon ei tarvitse olla 3. normaalimuodossa, tuo se tähän malliin joustavuutta. Koska tietovarastoon tallennetaan tieto sekä alkuperäisessä että muokatussa liiketoiminnan mukaisessa muodossa, on toteutusta helppo muuttaa jos liiketoimintasäännöt muuttuvat.*

- *Suhteellisen korkea oppimiskäyrä varsinkin alkuvaiheessa. Projektissa ja ylläpidossa tulee huolehtia, että kehittäjät ja ylläpitäjät ovat riittävästi koulutettuja.*

- *Tietoja käsitellessä joudutaan paljon tekemään liitoksia (join). Vaatii paljon onnistunutta suunnittelua, jottei toteutuksesta tule liian kompleksinen ja vaikeasti ylläpidettävä.*

Arkkitehtuurivalinta ei aivan yksinkertainen

- *Tietovarastoprojekti, jonka tavoitteet ovat epäselvät, on suuressa vaarassa epäonnistua*
- *Jos tietovarastotyyppi, -tuote ja sen arkkitehtuuri päätetään ensin, voi käydä niin ettei se tue tarpeita*
- *Johtamisen käyttötarpeet (strateginen, operatiivinen, pääkäyttäjätarpeet) on syytä suunnitella ja dokumentoida ennen teknologia valintoja.*

Kyselyn (n=33) vastausten yhteenveto

- Tietovarasto on käytössä noin puolella oppilaitoksista.
- Vastausten perusteella ymmärrys tietovarastoista, niiden ominaisuuksista ja mahdollisuuksista näyttäisi vaihtelevan.
- Jokin raportointiratkaisu on käytössä suurimmalla osalla oppilaitoksia. Yleisen käytössä oleva raportointiratkaisu on PowerBI.

- Kysely kohdennettiin hankkeessa mukana oleville koulutuksen järjestäjille.

1. kysymys

”Millaisella projektiryhmällä kannattaa tietovarastototeutusta lähteä pystyttämään? Mitä osaamista omassa organisaatiossa olisi syytä olla, mitä taas kannattaa yleensä hankkia ulkopuoliselta asiantuntijalta?”

Vastaus: *tarvitsette erinomaisen tietovarastointia tuntevan projektipäällikön, hankkeen omistajan joka sitoutuu aidosti (kiinnostus, resursointi), prosessiomistajien omistautumisen ja työaikaa sekä hyvän teknisen tiimin.*

2. kysymys

”Organisaatioissa on erilaisia tapoja toteuttaa tietovarastoja, ja niiden käyttöönotto vaihtelee. Minkälaisia ratkaisuja eri organisaatiot ovat tehneet tietovarastojen käyttöönotossa?”

Vastaus: *Osallistujat voivat kertoa tähän kysymykseen liittyen chatissä lisätietoja.*

”Miten organisaatiot ovat perustelleet henkilötietojen käsittelyä tietovarastoissa? Millaisiin haasteisiin tähän liittyen on törmätty ja miten ne on ratkaistu. Joillakin organisaatioilla on järjestelemien käyttöönoton yhteydessä tehtyä muutosvaikutusten arviointi, joilla avataan tietovaraston ja siihen liitettävien lisäosien vaikutuksia mm. henkilötietoturvaan, organisaation muihin prosesseihin ja riskeihin - miten muissa organisaatioissa?”

Vastaus: *Kaikki tietämäni tahot perustelevat tietovaraston käyttämisen lainsäädännöstä tulevien perusteiden ja vaatimusten kautta. Esimerkiksi laki ammatillisesta koulutuksesta ja oppilas- ja opiskelijahuoltolaki määrittelevät oppilaitosten oikeuksia ja velvollisuuksia, myös rekisterinpidon näkökulmasta. Tietovarastolla toteutetaan oppilaitoksen tukipalveluita lakisääteisten velvoitteiden suorittamiseen. Luonnollisesti tietoturvaan ja -suojaan liittyvät asiat ja näkökulmat sekä riskienhallinta on syytä olla osana tietovaraston kehitys- ja käyttöönottoprojektia.*

Tieto-osaava

Tiedon mallintaminen

Käsitteellinen tietojen mallinnus

- Käsitteelliseen tietojen mallinnuksessa määritellään organisaation ja tietojen yleinen rakenne. Näitä ovat esimerkiksi opiskelijoihin, opettajiin, muihin työsuhteisiin sekä opiskeluun liittyvät tiedot. Nämä tiedot muodostavat entiteeteiksikin mainittuja tietoluokkia ja asiakokonaisuuksia.

- Lisäksi useimmiten entiteetillä on suhteita muihin entiteetteihin. Esimerkiksi opiskelija-tyyppinen entiteetti linkittyy tutkinto-entiteettiin, tutkinto-entiteetti linkittyy organisaatio-entiteettiin.

- Käsitteellisessä mallinnuksessä määritetään nämä eetiteetit että edellä mainitut entiteettisuhteet.

Looginen tietojen mallinnus

- Looginen tietomalli laajentaa käsitteellistä mallia lisäämällä erityisiä tietomääritteitä jokaiseen entiteettiin sekä suhteita jokaisen määritteen välille.

- Loogisessa tietomallissa määritetään esimerkiksi opiskelija-entiteettiin liittyviä erityispiirteitä. Esimerkki tästä on vaikka opiskelijasekvenssien (oppivelvolliset, jatkuva oppiminen) erot. Loogiset säännöt periytyvät lainsäädännöstä ja asetuksista määrittäen, että oppivelvollinen on alle 18-vuotias, jolla ei ole vielä toisen asteen tutkintoa.

- Raportoinnin suunnittelu ilman loogisen tason tietojen mallinnusta on vaikeaa ellei mahdotonta. Loogisen tietomallin lisääminen jälkikäteen aiheuttaa yleensä enemmän työtä kuin sen suunnittelu ja huomioiminen ennen ensimmäisen toteutusversion tekemistä.

Fyysinen tietojen mallinnus

- Fyysinen tietomalli on loogisen tietomallin tekninen toteutus esimerkiksi tietovarastoon.

- Fyysisen tietokanta voidaan luoda joko manuaalisesti eli käsin tietovarastoon, käyttäen jonkinlaista välimallia tai puolittaista automatisointia vaikkapa exceliin tehdystä loogisesta tietomallista tietokantascripteiksi tai erillisin työkaluin, jotka muodostavat automaattisesti loogisesta tietomallista tietokannan.

- Tässä asiassa vaaranpaikat liittyvät yleensä siihen, että teknisillä ei-kontekstia tuntevia henkilöillä voi olla vaikeuksia tunnistaa onko looginen tietomalli täysin valmis, pahasti kesken vai jotain siltä väliltä. Valitaanko käyttöön työkalu, jolla voidaan useita kertoja iteroida loogisen tietomallin siirtämistä tietovarastoon vai riittääkö yksi manuaalinen kerta. Tärkeää arvioida työkalujen arvoa suhteessa käytetyn työajan arvoon.

Mitä on syytä huomioida, jos tiedot mallinnoidaan mallinnustyökalulla?

- Onko mallinnustyökalu helppo käyttää? Prosessien omistajat eli tiedon tuottajien tulisi mallinnusprosessissa vastata siitä, että tieto on mallinnettu oikein ja laadukkaasti. Kuten edellä todettiin, että tietohallinnolla ei edes aina ole osaamista arvioida tätä asiaa.

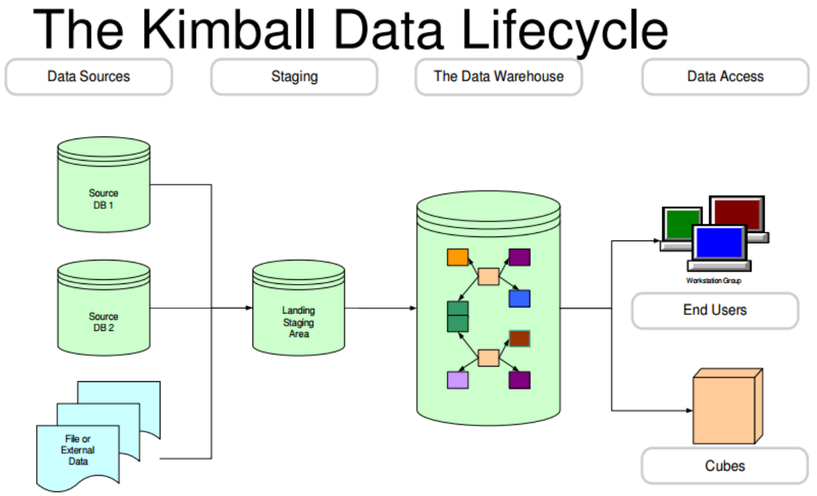
- Jos mallinnustyökalu valitaan vain teknisten ominaisuuksien perusteella esim. kuinka helppoa tiedot on viedä tietovarastoon, voi tulla haasteita. Asioiden teknisen tekemisen helppoudesta on vain vähän hyötyä, jos käsitemalli on laadultaan huono eikä vaikeakäyttöisyyden vuoksi laatua saada helposti parannettua.

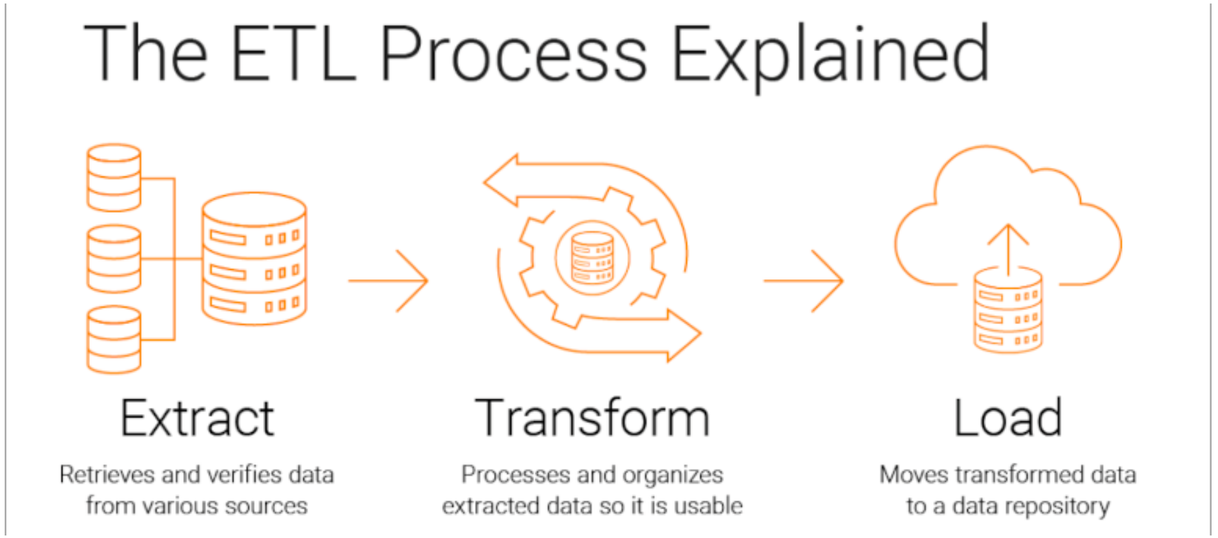
- Aiemmin käytiin läpi tietovarastotyyppejä (Inmon, Kimball, Data Vault 2.0). Mahdollisesti löytyy työkaluja, jotka toimivat parhaiten jonkun tai joidenkin tietovarastotyyppien kanssa.

ETL (Extract, Transform, Load)

- *ETL tarkoittaa datan siirtämistä ja muokkaamista ja lataamista: tiedot haetaan (Extract) lähdejärjestelmästä, niitä muokataan (Transform) ja ladataan (Load) lopulta tietovarastoon.*

- *Latausprosessissa tiedot muunnetaan tietovaraston rakenteen muotoon, integroiden samalla eri lähtöjärjestelmien tietoja.*



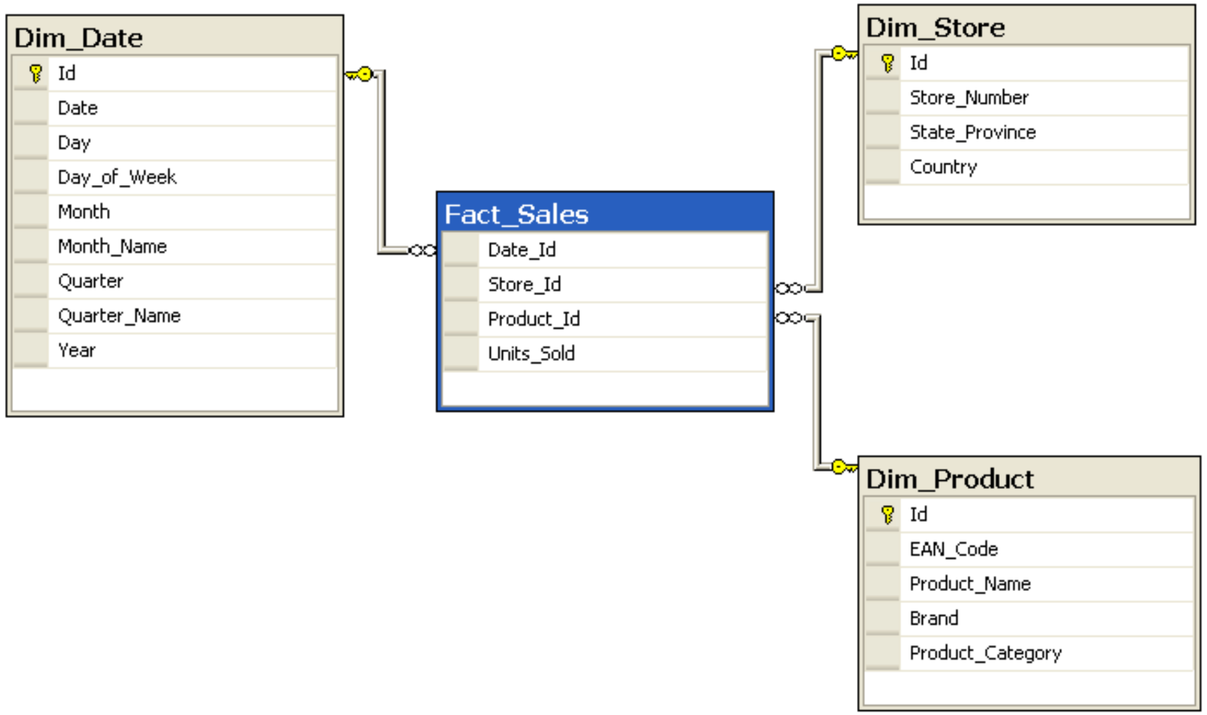


Tähtimalli (Star schema)

- *Tähtimalli on yksinkertaisin ja eniten käytetty tapa kehittää tietovarastoja ja dimensionaalisia data martteja (esim. Kimbalin malli).*

- *Tähtimalli sisältää yhden tai useita faktatauluja ja useita siihen tai niihin liittyviä dimensiotauluja.*

- *Tietovarastoissa ja raportointiratkaisuissa ei voida käyttää lähdejärjestelmistä tuttuja oliopohjaisia tietorakenteita niiden hitauden vuoksi.*



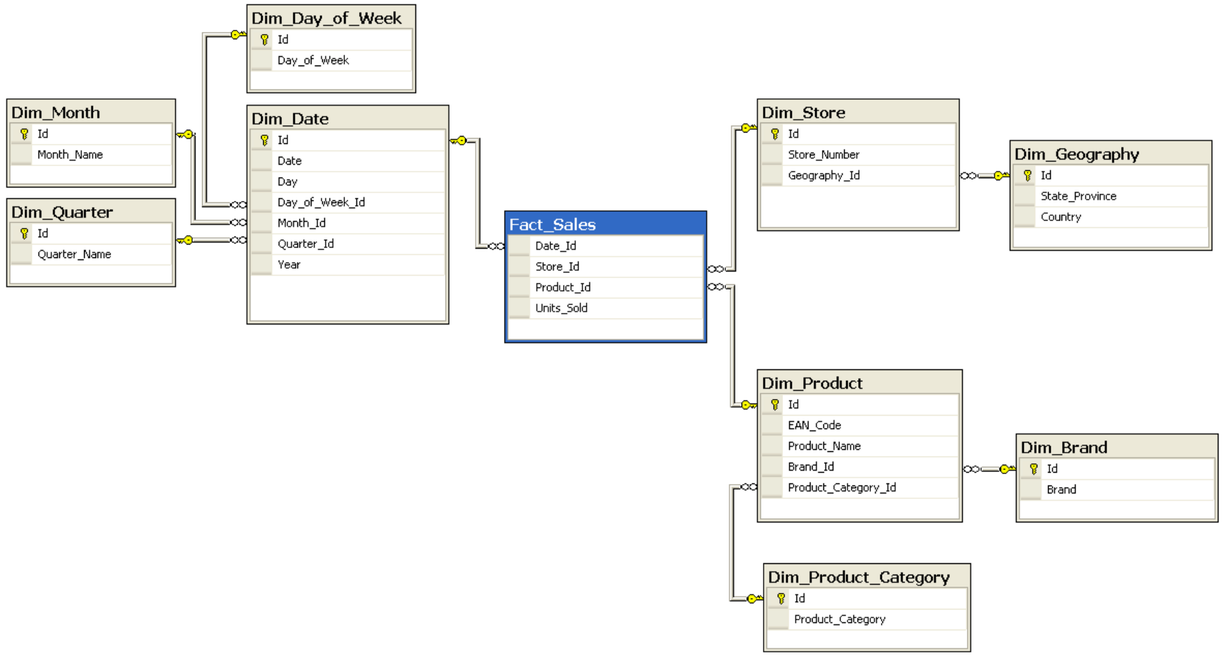
Kuvan lähde: https://en.wikipedia.org/wiki/Star_schema

Lumihiutale (Snowflake schema)

- *Hiukan tähtimallia monimutkaisempi tapa mallintaa tietoa tietovarastoissa. Kun tähtimallissa dimensiot olivat yksihierarkisia, lumihiutale skeemassa ne voivat muodostaa monihierarkisen tason.*

- *Lumihiutale-mallissa datan hakeminen tietovarastosta on nopeampaa kuin tähtimallissa.*

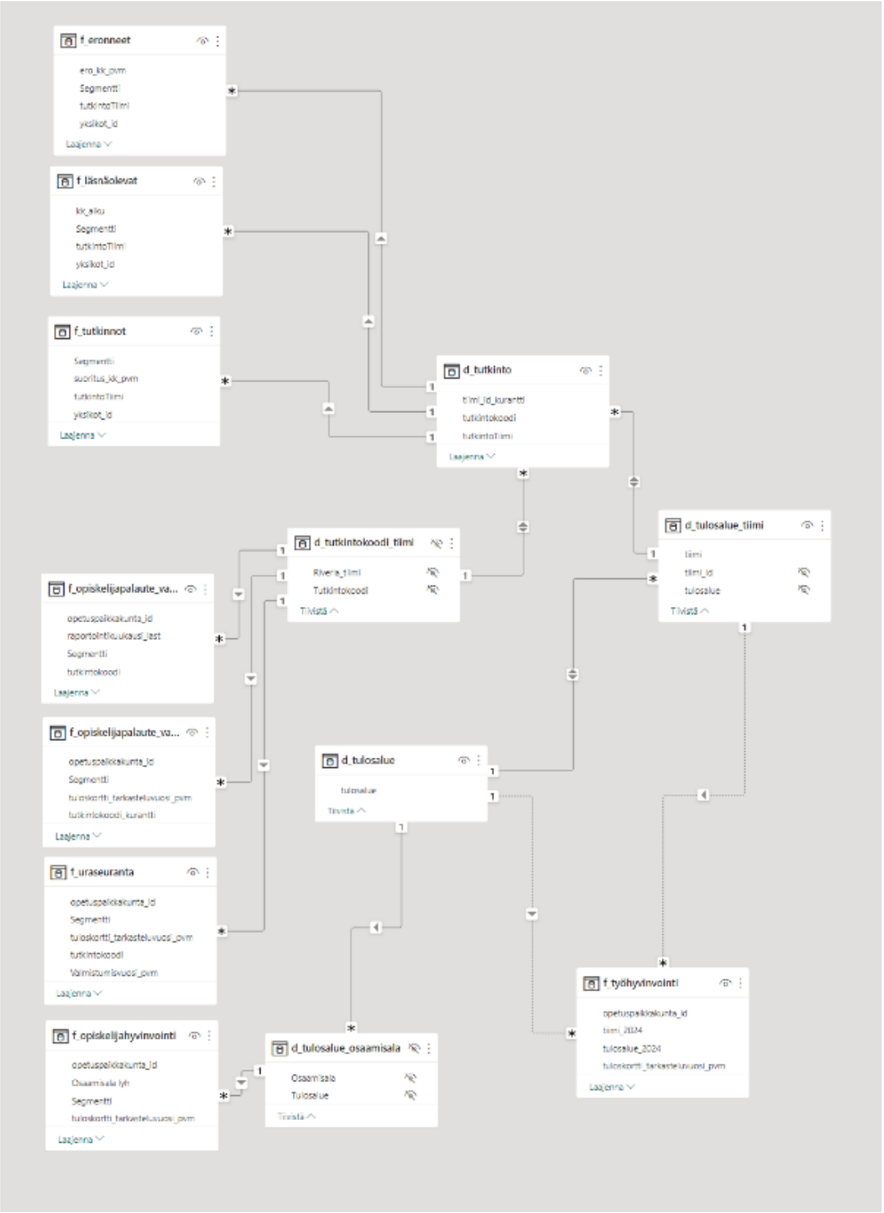
- *Lumihiutale-mallin rakentaminen on kuitenkin kompleksisempaa kuin tähtimallin rakentaminen.*



Kuvan lähde: https://en.wikipedia.org/wiki/Snowflake_schema

Esimerkki tietomallista PowerBI:ssä (case Riveria)

- Johtamisessa käytettävien mittareiden määrittely
- Mittareiden tarkennukset
- Tietomalli PowerBI:ssä



Tieto-osaava

Tieto-osaava

Kiitos!