Gagnasafnsfræði

Vöruhús gagna

Hallgrímur H. Gunnarsson

Kynning

Vöruhús gagna: (e. data warehouse)

Vöruhús gagna er gagnasafn sem er ætlað fyrir skýrslugerð og greiningu á gögnum.

Gögn sameinuð úr mörgum áttum á einn stað, svokallað vöruhús, til að vinna með þau. Gögn oftast tímatengd.

Áður en gögn eru sett í vöruhús þá eru þau oft hreinsuð, samhæfð og forunnin að e-u leyti.

Önnur tæknileg sjónarmið. Ekki lagðar sömu áherslur þegar gagnasafn er hannað fyrir vöruhús og fyrir venjulegt gagnasafn.

Hugmyndafræðin byggir meira á töflureikningslíkani frekar en stöðluðu (normalized) venslalíkani

Notkun á vöruhúsi

Dæmi um notkun:

Upplýst ákvarðanataka (e. decision support)

Viðskiptagreind (e. business intelligence)

Greina stefnu (e. trend analysis)

Spá fyrir um stefnu (e. forecasting)

Finna svikara (e. fraud analysis), t.d. tryggingagögn, símtalagögn, skattagögn

Greina símtalafærslur (call detail record analysis), úrvinnsla fyrir reikningagerð

Greina netmælingargögn, úrvinnsla fyrir reikningagerð

Tveir frasar: OLTP vs. OLAP

OLTP: (on-line transaction processing)

Einkennist af mörgum stuttum hreyfingum (INSERT/UPDATE/DELETE/SELECT) sem vinna með lítið af gögnum í mjög stöðluðu (normalized) gagnasafni.

OLAP: (on-line analytical processing)

Einkennist af fyrirspurnum sem vinna með stórt safn af static gögnum og reikna ýmiskonar samantektir og fleira.

Ný gögn koma í hollum úr mörgum áttum (batch import og consolidation), t.d. einu sinni á dag eða einu sinni á klukkutíma. Gömlum gögnum hent út reglulega, t.d. gæti verið ákveðið að geyma 2 ár aftur í tímann af sögulegum gögnum.

Gagnasafnið hannað með það í huga að gera fyrirspurnir einfaldar og hraðar. Meiri áhersla á svokallað víddarskipulag, minni áhersla á staðalskipulag, allt í lagi að endurtaka gögn þegar það hentar.

Kostir við vöruhús

Söguleg gögn:

Vöruhúsið viðheldur sögulegum gögnum þó upprunaleg kerfi geri það ekki.

Einn staður og eitt líkan:

Öll gögn á einum stað. Hægt að búa til skýrslur þvert yfir gögn úr mörgum kerfum.

Hægt að fella mörg skyld gögn í sama mótið og vinna með þau á einn hátt. Erfitt að vinna þvert með gögn sem eru með mjög mismunandi snið/uppbyggingu/töflustrúktúr jafnvel þó þau séu mjög skyld.

Léttir undir rekstrarkerfum:

OLAP kerfi fær oft gögn frá mörgum OLTP kerfum. Léttir á þeim. Viljum ekki keyra út þungar skýrslur og trufla rekstur.

Hönnun á vöruhúsi

Víddarskipulag:

Staðalskipulag er lausn við ákveðnum vandamálum (e. anomalies). Höfum vanalega engar áhyggjur af endurtekningu gagna í vöruhúsi. Notum víddarskipulag frekar en staðalskipulag.

Gagnatafla: (e. fact table)

Hver lína inniheldur e-r gildi (e. measures) og svo dálka sem venslast við víddartöflur.

Víddartafla: (e. dimension table)

Fyrir hverja vídd er ein tafla sem gagnataflan venslast við. Sú tafla inniheldur nánari upplýsingar um víddina.

Skipulag

Stærð á töflum:

Gagnatöflur eru vanalega grannar með margar mjög línur en víddartöflur eru vanalega víðar með færri línur

Stjörnuskipulag: (e. star schema)

Ein tafla per vídd.

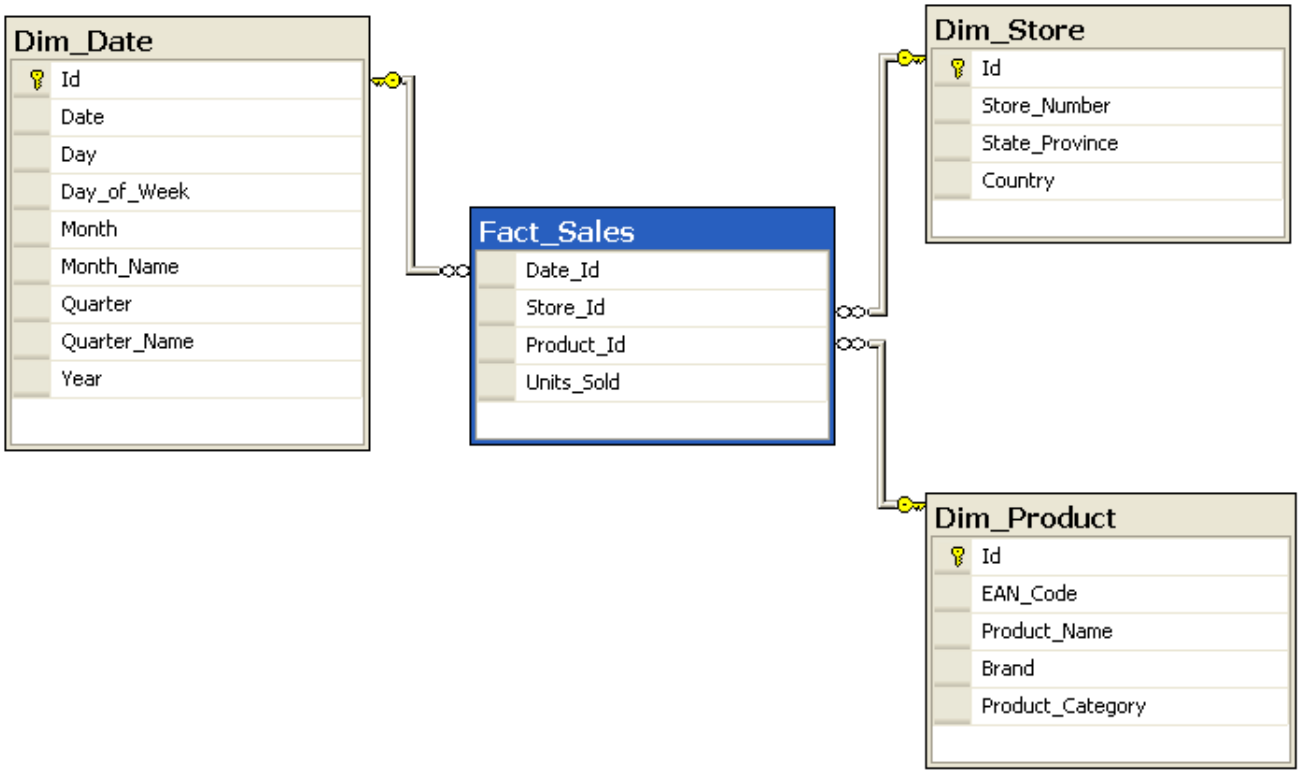
Snjókornaskipulag: (e. snowflake schema)

Víddartaflan stöðluð (normalized) yfir í margar töflur.

Flatt skipulag: (e. flat schema)

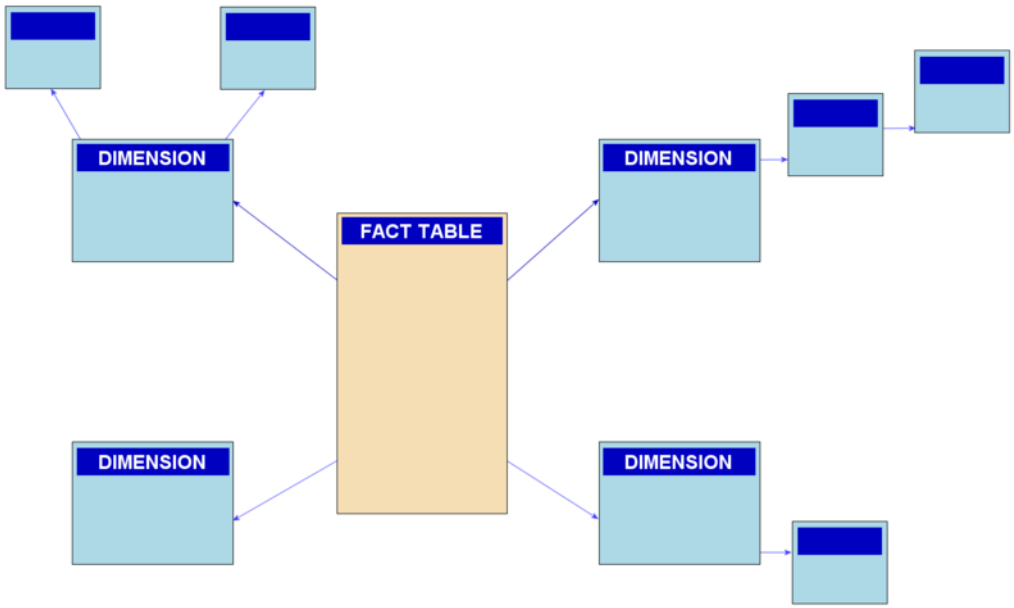
Víddartöflur felldar inn í gagnatöflu (denormalized). Flatur strúktúr. Tekur meira pláss.

Stjörnuskipulag (e. star schema)



(Mynd af Wikipedia)

Snjókornaskipulag (e. snowflake schema)



(Mynd af Wikipedia)

Þróun vídda með tímanum

"Hæg breyting vídda" vandamálið: (e. slowly changing dimensions problem)

Gögn í víddum geta tekið breytingum með tímanum. Dæmi: gagnatafla yfir sölu, víddartafla fyrir sölumenn sem inniheldur staðsetningu sölumanns og fleira. Hvað gerist þegar sölumaður er fluttur milli staða?

Dæmi um mögulegar lausnir:

1. Yfirskrifa gömlu gildin (töpum gamla), getur leitt til vandamála í skýrslum
2. Skrá nýju gildin sem nýja útgáfu og halda utan um margar útgáfur (gildistími fyrir hverja útgáfu)
3. Nota flatan strúktúr á gagnatöfluna og geyma víddirnar þar líka

OLAP verkfæri

Hópun: (e. aggregation)

Nánast allar fyrirspurnir í OLAP vinnslu fela í sér hópun. Þá er til dæmis verið að reikna samantekt gilda samkvæmt tilteknum víddum. Dæmi: heildarsölu, heildarsölu fyrir hverja borg eða hvert land, o.s.frv.

Rúlla upp: (e. roll-up)

Að *rúlla upp* er að framkvæma hópun upp á hærra stigveldi víddar. Til dæmis farið úr því að skoða sölur per mánuð yfir sölu per ár (tímavíddin).

Bora niður: (e. drill-down)

Að *bora niður* er andstæðan við að rúlla upp. Til dæmis farið úr því að skoða sölur per land niður í að skoða sölur per borg í landi.

OLAP verkfæri framhald

Vending: (e. pivoting)

Vending byltir niðurstöðunum, dálkar verða línur og öfugt. Til dæmis, förum úr því að skoða summur gilda fyrir ár (línur) á móti borgum (dálkar) í summur gilda fyrir borgir (línur) á móti árum (dálkar).

Skera og hluta: (e. slice and dice)

Skorður settar á gildi í víddum, til dæmis ár=2011 eða skoða síðustu 10 ár, o.s.frv.

Innsetning gagna í vöruhús

ETL: (Extract, Transform, Load)

Ferlinu er skipt í þrjú skref: Extract, Transform, Load.

Sækja gögn úr ytri kerfum: (e. extract)

Sækja gögn úr ytri kerfum og lesa þau. Ytri kerfi: gagnagrunnur, FTP svæði, vefþjónusta, o.s.frv. Mismunandi gagnasnið, t.d. CSV, binary, custom snið, o.s.frv.

Umbreyta gögnum: (e. transform)

Hreinsa gögn. Fletja út (denormalize). Henda út því sem er ekki notað. Forvinna, t.d. summera eða bæta inn static upplýsingum.

Static upplýsingar: gildi sem eru háð gögnunum en er erfitt eða dýrt að reikna með fyrirspurn seinna

Innsetning gagna í vöruhús frh.

Innsetning: (e. load)

Innsetning þarf að halda bókhald til að tryggja skráningu (ekkert má tapast/gleymast) og koma í veg fyrir tvískráningu.

Dæmi um bókhald fyrir innsetningu:

Skrá kemur úr ytra kerfi með raðnúmer: 1.dat, 2.dat, 3.dat, o.s.frv.

Höldum bókhald utan um raðnúmer síðustu skrá sem var sett inn.

Ef næsta skrá er ekki raðnúmer+1 þá vantar skrá.

Ef við fáum skrá með raðnúmer $<$ síðasta raðnúmer, sleppum henni.

Innsetning með hreyfingu (í lægsta einangrunarstigi) til að tryggja all-or-nothing skráningu.

Töfluskipting (e. partitioning)

Vandamál:

Gagnatafla getur orðið mjög stór. Þurfum að geta sótt gögn hratt og líka hent gömlum gögnum án þess að það sé mjög dýr aðgerð.

Töfluskipting:

Vanalega eru gögn tímatengd, t.d. sölur eða pantanir.

Sniðugt að skipta töflum upp í tímatöflur, t.d. ein tafla per mánuð. Fyrirspurnir sækja bara úr þeim töflum sem þær þurfa. Auðvelt að droppa gömlum töflum.

Stundum er stuðningur fyrir töfluskiptingu innbyggður í gagnagrunnum, stundum þarf að nota heimasmíðaða lausn og handvirka skiptingu.

Views og materialized views

Sýndartafla: (e. view)

Búum til sýndartöflu (e. view) fyrir algengar fyrirspurnir. Slík sýndartafla er í raun bara *geymd fyrirspurn* sem lítur út eins og tafla. Reiknuð úr fyrirspurninni þegar sótt er úr sýndartöflunni.

Fyrirframreiknuð sýndartafla: (e. materialized view)

Ef það er oft sótt úr sýndartöflu þá getur borgað sig að reikna upp úr henni og geyma niðurstöðuna sem alvöru töflu. Slík sýndartafla er kölluð *materialized view*

Koma upp atriði eins og hversu oft á að endurnýja töfluna, hvaða flýtivísar eiga að vera á henni, o.s.frv.