Datavarehus

(Connolly & Begg, ch. 31 – 34)

Hva?

Et datavarehus er en samling av data lagret slik at de egner seg for analyse f.eks.

- trendanalyse, konkurranseanalyse, kundeanalyse og annen form for markedsanalyse (mest vanlig bruk)
- analyse f.eks. i forbindelse med forsikring og -premier
- data f.eks. i forbindelse med geologiske data etc.

kort sagt: beslutningsstøtte på en eller annen måte.

Karakteristika:

- ofte mye data
- samlet opp over et lenge tidsrom.
- fra ulike kilder
- summer er ofte viktigere enn enkeltdata
- separat fra den vanlige driftsdatabasen

NB! Warehouse = lagerhus, ikke "supermarked".

> Data Warehousing:
> ”A subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management’s decision-making process” (Inmon, 1993)

Bruk av data kan deles i data for transaksjonsbruk og data for analyse bruk:

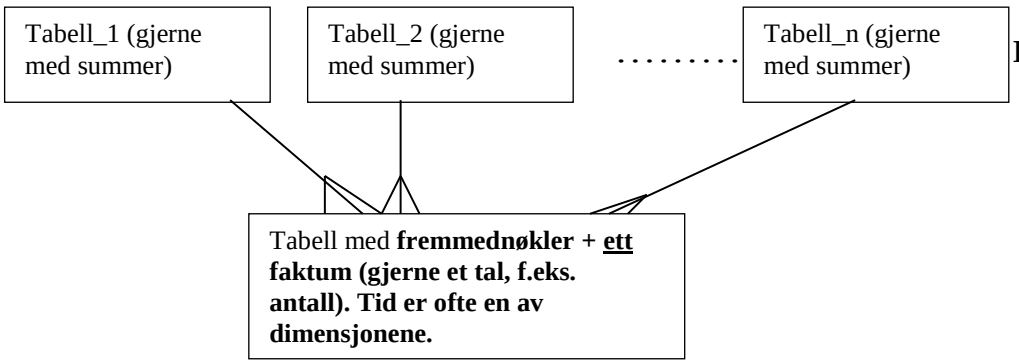
- OLTP - On Line Transaction Processing
- OLAP - On Line Analytical Processing. Det er her Datavarehus kommer inn.

Hvordan - planlegging

- svært forenklet:

Planlegging:

- Klargjøre hensikten med datavarehuset
- Klargjøre størrelsen og omfanget av datavarehuset
- Lage felles mønster fra heterogene systemer og kilder (felles metadata).
  Blir et supersett av alle kilder - ofte med i utgangspunktet inkompatible data.
  Ofte nødvendig med surrogatnøkler m.m.
- Laging av datavarehus-datamodell (ofte dimensjons-og faktatabeller, ofte denormalisert, summeringer m.m.)



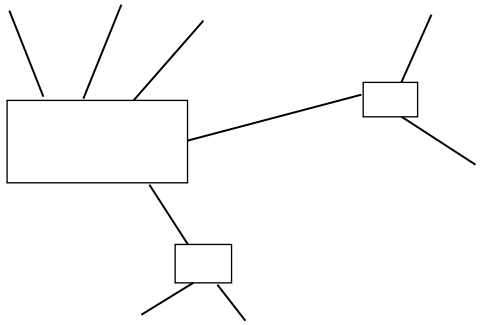
- Lage regler for vasking/rensking av data
- Lage regler for selve overføringen (når, hvorfra, samme/ulike data de ulike gangene, samtidighetsspørsmål)
- Lage regler for bruk, bl.a. adgangskontroll
- Skal data noen gang slettes fra datavarehuset, skal de i tilfelle overføres til summer?
- Hvilke spørringer vil være de typiske
- NB! Alt dette henger sammen og må betraktes samtidig

Dimensjonsmodellering.

Det å lage en slik struktur kalles ofte for dimensjonsmodellering.

En del faktorer innen dimensjonsmodellering:

- må ofte ta hensyn til tidligere strukturer, f.eks. varegrupperinger som ikke finnes lenger - hvorledes transformere disse
- må ta hensyn til endringer i framtiden
- må ofte ta hensyn til ulike kilder, f.eks. fra sammenslåtte firmaer
- inkommensurable (usammenlignbare) størrelser, evt. konvertering av størrelser
- granulariteten av data
- det skal stadig fylles på med data, ingen data skal bort, men må muligens aggregeres etter hvert. Er nyere data viktigere enn gamle?
- bruken av data blir viktig, f.eks.: skal man ha med salgssted i et med et datavarehus for et grossistfirma eller ikke.
- må også ta hensyn til presentasjon av dataene, f.eks. grafisk etc.
- ”stjerne”/star - de-normalisering (mest vanlig: nedflating av et hierarki) er nødvendig for å få rask analyse, og er ufarlig, fordi data kun er generert fra (forhåpentligvis) konsistente strukturer.
- "snøflak"/snowflake - når man ikke kan/bør denormalisere mye

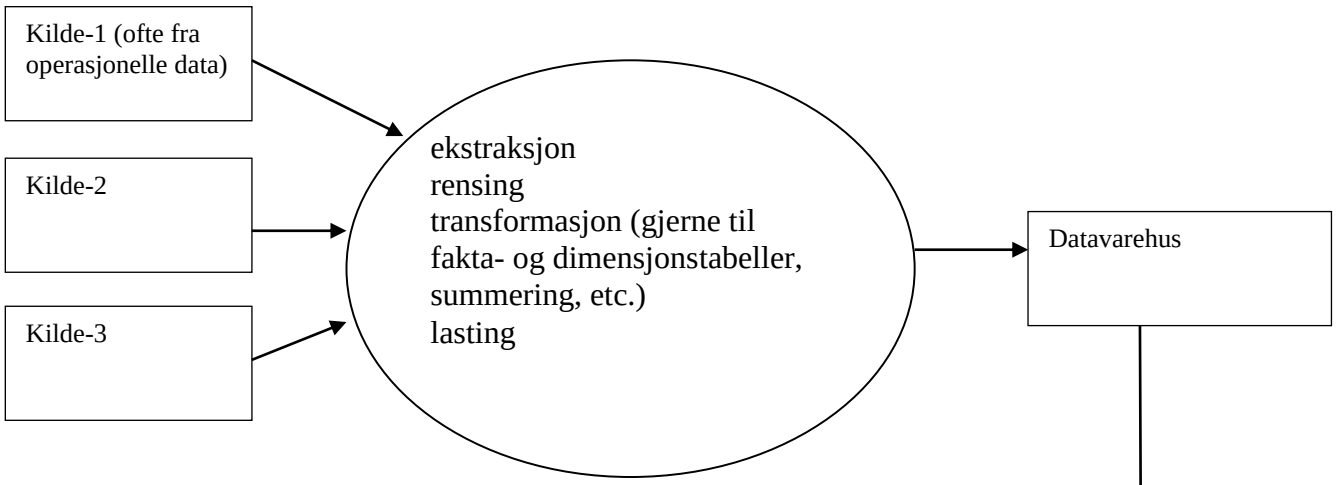


- ”stjerneflak”/starflake - blanding
- størrelse, omfang og bruk: "data warehouse" vs. "datamart"

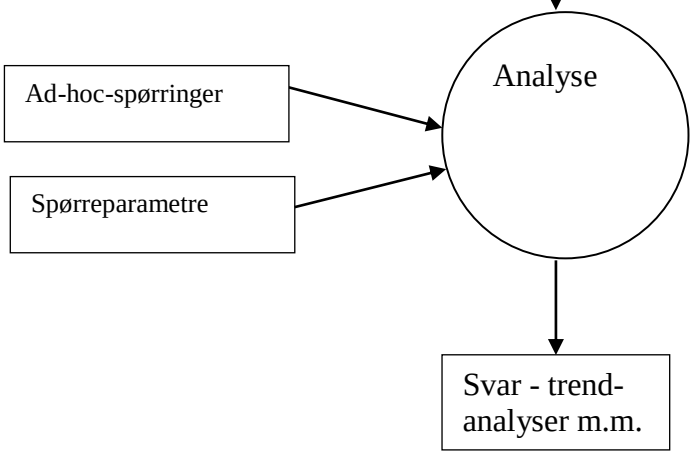
Altså:

å bygge et datavarehus er temmelig forskjellig fra å bygge en normalisert, bruks/applikasjons-uavhengig database for mindre datamengder.

Hvordan arbeide med et datavarehus - laging og bruk – svært forenklet:



Bruk av et datavarehus



Kortform: ETL = Extract – Transform - Load

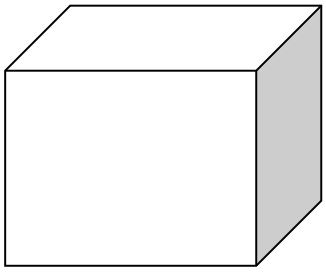
Organiseringsformer for data

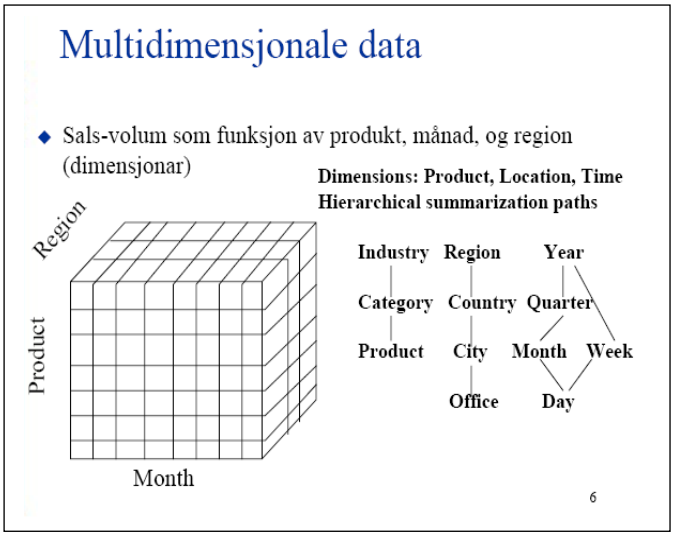
ROLAP – Relational OLAP.

Dataene lagres relasjonelt, dvs. som tabeller, med f.eks. et stjerneskjema som logisk modell. Bruker SQL (gjerne med utvidelser) som databasespråk.

MOLAP – Multidimensjonal OLAP (datakuber)

Dataene lagres som kuber (dvs. på samme måte som 2-, 3- eller flerdimensjonale arrays i programmeringsspråk). Bruker spesialspråk i kombinasjon med SQL.





Eksempel på dimensjoner[1]

- Kan godt være 4 eller flere dimensjoner, selv om det kan ikke tegnes.
- Tilsvarer multivariat analyse i statistikk
- Lite effektivt ved glisne data.

HOLAP – Hybrid OLAP.

Kombinasjoner.

NB! Systemene kan være laget slik at de fra brukernes side kan ses på som et MOLAP, men være ROLAP og motsatt.

[1] http://www.idi.ntnu.no/emner/tdt4150/foilar/OLAP.pdf

Noen variasjoner og begrep

Datavarehus og datamarked

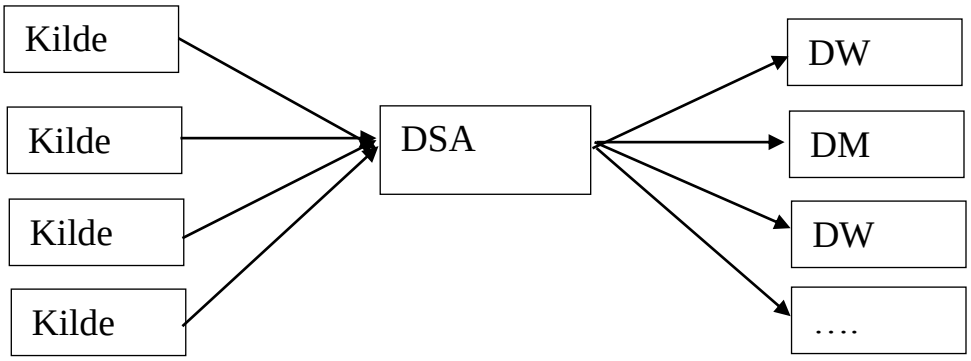
Et **datamarked** (data mart) er en minivariant at et datavarehus.

”Data Staging Area”

(stage = plattform, stillas)

- i stedet for å kjøre data rett over i et datavarehus, lager man en tradisjonell, normalisert relasjonsdatabase som inneholder alle data som skal inn i datavarehuset. Denne brukes så til å lage datavarehus (DW) eller data marts (DM).

- Dermed lettere å lage DW og DM mer på ad-hoc-basis.



Data Mining

Bruk av datavarehus til å oppdage skjulte sammenhenger, ofte ved å bruke statistiske og/eller heuristiske metoder

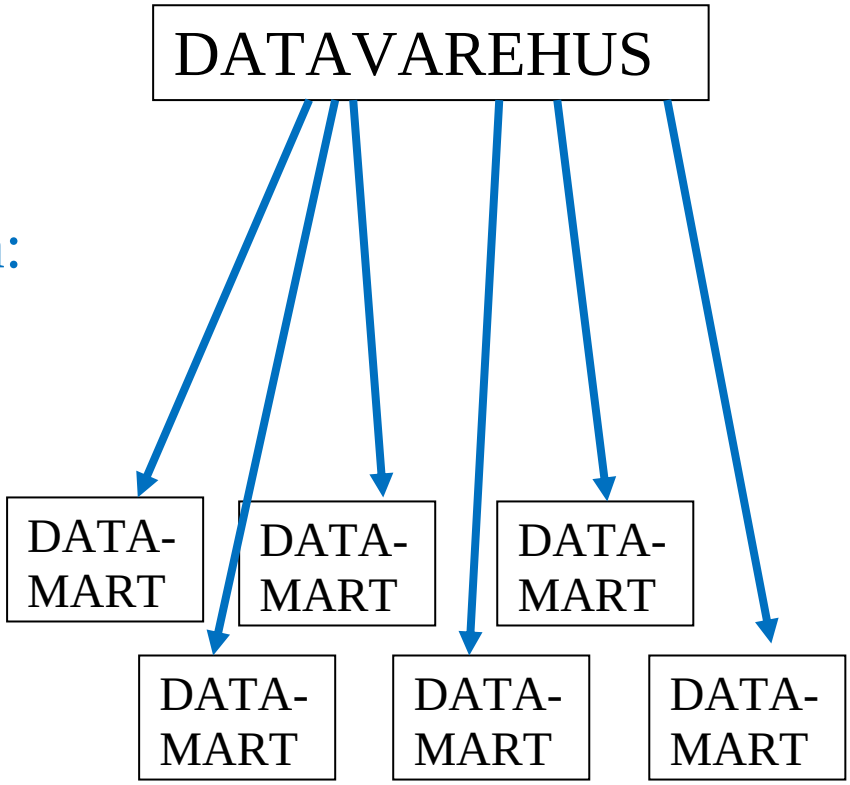
Business Intelligence

Blir gjerne brukt som strategisk bruken av en datavarehus.Ofte:

- datafolk snakker om Datavarehus,
- økonomer snakker om Business Intelligence.

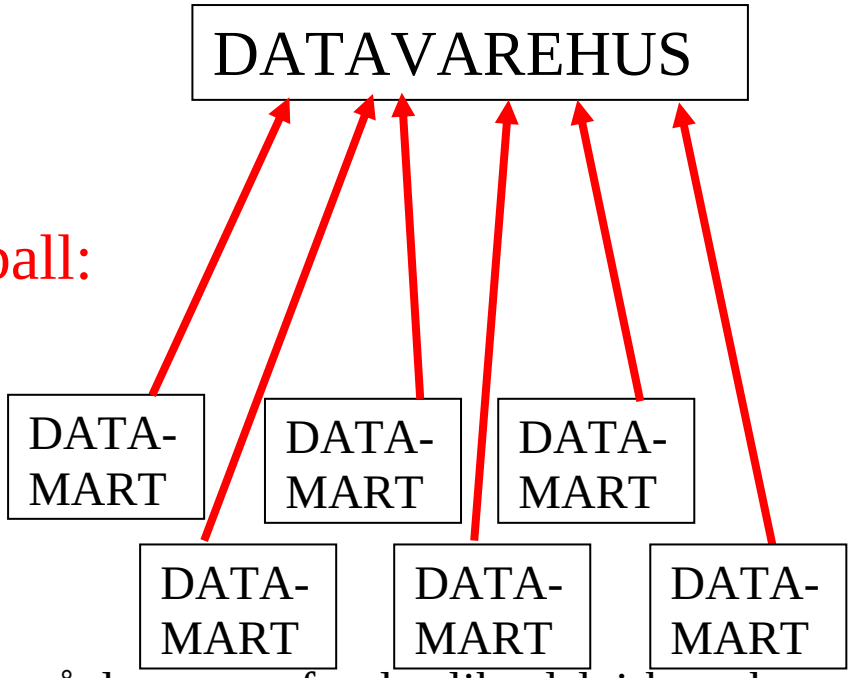
Top-down eller bottom up?

Bill Inmon:



Fokuser på felles, «enterprise» datavarehus, lag data marts hvis det trengs.

Ralf Kimball:



Fokuser på data marts for de ulike delvirksomhetene, lag felles enterprise DW hvis nødvendig.

Verktøy for DW, OLAP, BI ....

Det finnes en rekke verktøy på området:

- **Integrert eller som tilleggsverktøy til databasesystemer**, f.eks. Oracle Oracle Data Warehouse, OLAP, Data Mining.
  Features (C & B):
  - summary management
  - analytical functions
  - bitmapped indexes
  - advanced join metods
  - sophisticated SQL optimizer
  - resource management

- Egne **kommersielle verktøy**, som ofte vanligvis lagrer data i andre kilder, f.x. et vanlig databasesystem (et eksempel: SAS Institute, se www.sas.com)

- **Gratisverktøy**, se f.eks. http://butleranalytics.com/free-olap-tools/
  Men OBS: mange «gratisvertøy» er «free download», ikke «freeware»

  Pentaho er blant de mest kjente freeware, og brukes også en del i undervisning, bl.a. fordi det er lett å lære.

- **Regnearksystemer** har også en god del analysefunksjoner for OLAP, datamining m.m.

Typisk bruk av datavarehus:

- multivariat analyse
- tidsserieanalyse
- "datadrilling" (drill down)
- "oppsummering" (drill up)

God struktur på datavarehuset kan være avgjørende for et firmas suksess eller ikke.

Sammenligning mellom OLTP og Datavarehus (Connolly & Begg):

OLTP	**Data warehouse systems**
Holds current data	Holds historical data
Stores detailed data	Stores detailed, lightly and highly summarized data
Data is dynamic	Data is largely static
Repetitive processing	Ad-hoc, unstructured, and heuristic processing
High level of transaction throughput	Medium to low level of transaction throughput
Predictible patterns of usage	Unpredictible pattern of usage
Transaction-driven	Analysis driven
Application oriented	Subject-oriented
Supports day-to-day decisions	Supports strategic decisions
Serves large number clerical/operational users	Serves relatively low number of managerial users

«Big data» (noen få ord bare ……..)

Begrepet «Big data» er et nytt begrep som har noen likheter med datavarehus og data mining, men også andre aspekter.

- Ønske om å lagre og se mønstre i enormt store mengder data
  - ikke bare GB, 10^9
  - men også TB, 10^{12}
  - petabyte, PB, 10^{15}
  - exabyte, EB 10^{17}
- Ofte komplekse data
- Vanlige relasjonsdatabaser holder bare delvis, eller ikke i det hele tatt
- Ofte «skreddersydde» databasesystemer for enkeltanvendelser
- Ofte avansert matematikk og statistikk for å finne mønster i dataene ➔ finne skjulte sammenhenger, se trender ……..
- Ofte data fra **mange steder** over **lengre tidsrom,** ofte **spatiale data**, ofte **realtimedata** (f.eks. fra sensorer).
- Fra natur, teknikk, økonomi, samfunnsforhold, informatikk (bl.a. Internett generelt og sosiale media spesielt)
- Strekker alle grenser til det ytterste, både lagringsmessig, prosesseringsmessig (bl.a. hardware, algoritmer, parallellisering), nettverksytelse, algoritmer og presentasjonsmessig.

Sjekk

http://www.oracle.com/pls/ebn/swf_viewer.load?p_shows_id=5671642&p_referred=undefined&p_width=800&p_height=600

http://www.oracle.com/technology/products/warehouse/SHORT_intro_to_owb10gR2/SHORT_intro_to_owb10gR2_viewlet_swf.html

http://www.oracle.com/technology/products/warehouse/11gr1/presentations/owb11gr1-overview.ppt

Søk etter begrepene:

- Business Intelligence
- ETL
- Data Mart
- OLAP
- data cubes
- Big data

Søk hos

- SAS Institute
- Oracle & datavarehus
- IBM & datavarehus
- Microsoft & datavarehus
- MySQL & datavarehus
- Pentaho. http://www.pentaho.com/products/demos/showNtell.php?tab=demos.
  Pentaho brukerveildening