SISU dokument

Data Warehousing

–

Ny teknik för bättre beslutstöd

Peter Rosengren

nr
24

Sammanfattning

Denna rapport ger en kortfattad introduktion till begreppet *data warehousing*. Syftet med data warehousing är att ge bättre beslutstöd till en bred skara användare i en organisation. Med ett datavaruhus menas en separat kopiedatabas, till vilken data från operativa produktionsdatabaser kopieras och transformeras för att bättre anpassas till frågeställande och analyser.

Ett datavaruhus brukar definieras som *ett ämnesorienterat informationslager utformat specifikt för beslutstödstillämpningar*. Ett datavaruhus är *skräddarsytt* för en viss kund.

Datavaruhuset används uteslutande för att återsöka och analysera data, inte för uppdatering. Datavaruhuset fylls typiskt på en gång om dygnet, oftast på natten när belastningen på produktionsdatabaserna är låg.

I begreppet data warehousing ingår tre komponenter:

- *Beslutstöd*, det vill säga de program som finns tillgängliga för slutanvändarna i olika beslutsituationer. Ett nyckelord är *flerdimensionell analys*.
- *Databasteknik*. Ny parallell databasteknik gör det möjligt att hantera mycket stora datavaruhus. 500 Gbyte är idag en vanlig storlek på ett datavaruhus. Datavaruhus byggs ofta upp kring en relationsdatabas, men vissa förespråkar en speciell typ av databasteknik, MDBMS, *flerdimensionella databaser*.
- *Konstruktion och underhåll av datavaruhus*. Speciella produkter finns för att hantera överflyttning och transformation från operativa data till beslutstödsinformation, men man kan också bygga detta själv.

Ett viktigt område inom data warehousing är hantering av metadata, det vill säga information om data, hur olika begrepp i varuhuset är definierade och var informationen har hämtats från.

När det gäller beslutstöd är *Cognos*, *Business Objects* och *Andyne* framstående leverantörer. På databassidan har *Oracle* och *Informix* framgångar på grund av sina parallella databasarkitekturer. De stora aktörerna idag inom konstruktion och underhåll av datavaruhus är *Prism Solutions* och *Carleton*.

Analytiker uppskattar att produkter för datavaruhus kommer att omsätta 1 miljard dollar redan i år. Undersökningar pekar på att 90 procent av företagen på USAs Fortune 2000-lista planerar att bygga ett datavaruhus inom ett – två år.

Innehållsförteckning

1 Introduktion

Data lagrade i stora databaser utgör en värdefull resurs i de flesta företag. I stora databassystem lagras enskilda transaktioner, t ex varje bankautomatuttag, varje telefonsamtal, varje inköpt produkt m m. Dessa databaser brukar kallas produtionsdatabaser, d v s det är där data produceras. Den stora mängden data i ett produktionssystem utgör en guldgruva för den som vill analysera företagets försäljning, kundernas köpmönster, nätbelastning etc.

Problemet har hittills varit att om dataavdelningarna tillåter användarna att ställa frågor, göra grupperingar och summeringar direkt i en produktionsdatabas så belastar det systemet alldeles för mycket och äventyrar driftsäkerheten. En komplicerad SQL-fråga som involverar koppling mellan tre – fyra tabeller och sedan grupperingar och delsummeringar kan ta upp till en timme att exekvera i en riktig stor databas. Dessutom är en produktionsdatabas ofta utformad från ett teknisk perspektiv för att ha bra prestanda vid uppdateringar och är inte optimerad för frågeställande.

Detta problem har lett fram till tanken på att skapa så kallade datavaruhus, *eller data warehousing*, vilket tycks kunna bli ett av årets stora modeord. Med ett datavaruhus menas en separat kopiedatabas, till vilken data från operativa produktionsdatabaser kopieras över. Det är inte enbart en fråga om att kopiera utan datatransformationer görs också för att åstadkomma bättre anpassning till frågeställande.

Datavaruhuset används sedan uteslutande för att återsöka och analysera data, inte för uppdatering. Datavaruhuset fylls typiskt på en gång om dygnet, oftast på natten när belastningen på produktionsdatabaserna är låg. Datavaruhuset fylls alltså på med nya poster hela tiden, medan existerande poster i databasen aldrig ändras.

Ett datavaruhus brukar definieras som *ett ämnesorienterat informationslager utformat specifikt för beslutstödstillämpningar*. Det är viktigt att notera att ett datavaruhus inte är något man köper utan något som byggs, det vill säga ett datavaruhus är *skräddarsytt* för en viss kund.

Datavaruhus är egentligen ingen ny företeelse. Dataavdelningar, speciellt i stora företag, har sedan länge gjort kopior av produktionsdatabaser för att tillåta användare att ställa frågor utan att störa produktionssystemens drift.

Redan för några år sedan började begreppet data warehousing dyka upp. Förespråkarna för datavaruhus hävdade då att det inte räcker med att göra direkta kopior av produktionssystemen. De senare innehåller data med en för hög detaljeringsgrad. I ett datavaruhus måste data förberedas för sökning och analyser genom att summeras och aggregeras till större enheter. Annars blir svarstiderna alldeles för långa.

En del hävdar till och med att det krävs en speciell typ av databas, så kallade *flerdimensionella databaser*. Ett fåtal sådana finns, t ex *Red Brick VPT* och *IRI Express*. En mindre drastisk ansats är att använda en traditionell relationsdatabas som kärna i datavaruhuset, men specialprogram som hanterar kommunikation, kopiering och överflyttning av data från produktionssystemen till datavaruhuset.

Det finns idag några så kallade *Data Warehouse Management Systems*, DWMS, som gör precis detta. Exempel på sådana är *Carleton Passport*, *Extract Tool Suite* och *Prism Warehouse Manager*. Dessa verktyg kostar 75 000 - 250 000 dollar.

Bakom Prism Warehouse Manager står Prism Solutions, ett företag som grundats av W.H. Inmon, mannen som anses ha uppfunnit begreppet data warehouse. Han har också skrivit två böcker i ämnet "Building the Data Warehouse" och "Using the Data Warehouse".

Nu upplever datavaruhusen återigen en ny vår. Det nya nu är att allt fler börjar säga att ett datavaruhus kan byggas helt med standardteknik utan att man behöver köpa dyr specialprogramvara. Därmed blir också tekniken tillgänglig för många fler och användarna kan fortsätta arbeta med verktyg de är vana vid från sitt dagliga arbete.

Några vanliga argument för att bygga ett datavaruhus brukar vara:

- Att mer kostnadseffektivt ge beslutstöd.
- Bättre omvärldsbevakning.
- Förbättrad kundservice.
- Att identifiera nya marknadsmöjligheter.
- Öka konkurrenskraften.

Vilka problem kan man stöta på, när man börjar bygga ett datavaruhus? Det största problemet är i allmänhet att bestämma varuhusets utbud, det vill säga vilka data som ska kopieras över till datavaruhuset från produktionsystemet.

Det kan bara avgöras genom en noggrann informationsanalys tillsammans med användare. Detta kan vara en tidsödande process, speciellt om man ger sig i kast med att försöka kartlägga hela organisationens informationsbehov och försöker ta fram en företagsövergripande informationsmodell. Därför rekommenderar flera experter att man börjar i mindre skala och inför begränsade datavaruhus som sedan tillåts växa allt eftersom användarna upptäcker vilken information de kan få ut och därmed ställer nya krav.

Om datavaruhuset ska innehålla data från olika produktionssystem så är det ofta så att samma data är definierat på olika sätt i olika system. Begreppet "Kund" kan vara tolkat på ett sätt i ett kundregister men på ett annat sätt i ett försäljningssystem. Kunder kanske också identifieras på olika sätt i två system, t ex om man har två olika nummerserier som identifikationsnycklar. Detta trots att det i verkligheten rör sig om samma kunder. Sådana här semantiska olikheter måste klaras ut av ett datavaruhus.

Vad kostar det att bygga ett datavaruhus? Eftersom datavaruhus skräddarsys för varje organisation varierar detta beroende på behov och ambition, men det finns uppgifter som pekar på att ett datavaruhusprojekt i snitt kostar 3 miljoner dollar att genomföra, drygt 20 miljoner svenska kronor. Det finns dock också uppgifter om datavaruhus som kostat det tiodubbla att bygga. Kostnaderna ökar om datavaruhuset är stort och kräver speciella parallella datorer etc.

Avslutningsvis kan sägas att datavaruhus är ett område som kommer att få mycket uppmärksamhet framöver. Enligt en undersökning som Meta Group gjort planerar 90 procent av USAs största företag (Fortune 2000-lista) att bygga ett datavaruhus under de närmsta två åren. Analysföretaget Aberdeen Group i Boston uppskattar att produkter för datavaruhus kommer att omsätta 1 miljard dollar under 1995.

Denna rapport ger en introduktion till data warehousing och de tekniker som ligger bakom begreppet. Rapporten ger också information om vilka produkter som idag finns tillgängliga för att bygga datavaruhus.

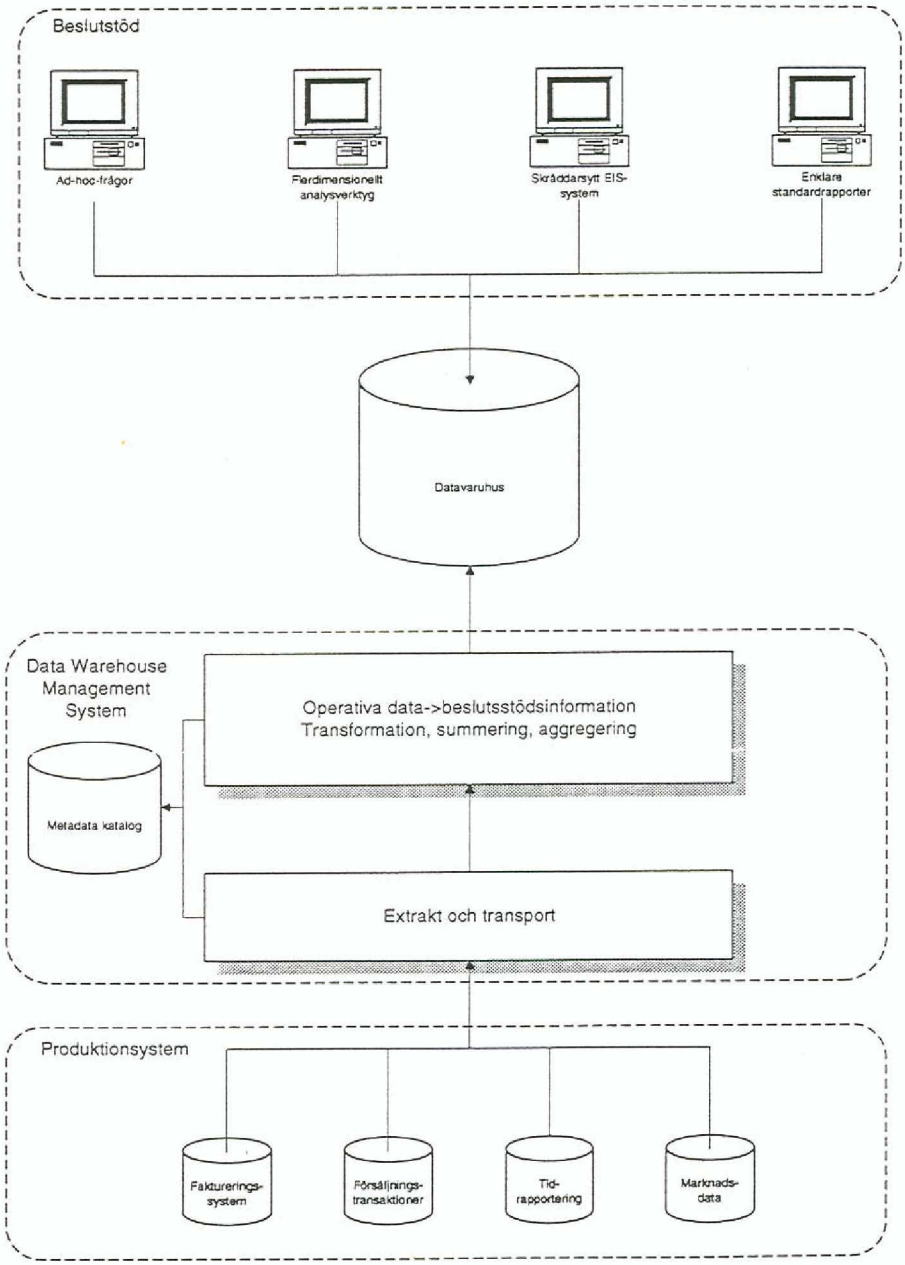
I kapitel 2 behandlar vi principer för hur ett datavaruhus är uppbyggt. Kapitel 3 ger en översikt av vilka olika typer av program för beslutstöd som finns tillgängliga. I kapitel 4 förklarar vi vilka olika typer av databasteknik som kan användas i ett datavaruhus. Kapitel 5 beskriver hur konstruktion och underhåll av ett datavaruhus görs och vilka olika kategorier av produkter som finns på marknaden för detta. I kapitel 6 sammanfattar vi våra slutsatser och i kapitel 7 ges en produktsammanfattning. Appendixet innehåller en kort ordlista.

Rapporten är framtagen inom ESPRIT-projektet Intuitive där SISU deltar som en partner. I projektet utvecklas bland annat nya söksystem för datavaruhustillämpningar.

2 Arkitektur för datavaruhus

I detta avsnitt går vi igenom principerna för hur ett datavaruhus är uppbyggt. I princip kan man säga att ett datavaruhussystem består av tre olika delar:

- Beslutstödsprogram.
- Datavaruhus.
- Program för drift och underhåll av datavaruhus, Data Warehouse Management System.



Figur 1. Principskiss av hur ett datavaruhus är uppbyggt.

Beslutstödsprogrammen är det som användarna arbetar med. Det finns olika typer av beslutstödsprogram som kan användas var för sig eller tillsammans. Typiska funktioner som ett beslutstödsprogram erbjuder är möjligheter till spontana frågor,

möjligheter att analysera data i flera dimensioner, nedbrytning av data till olika detaljeringsgrader.

Datavaruhuset är en databas där beslutstödsinformation finns lagrad. Datavaruhuset kan vara baserad på en relationsdatabas men det finns också en speciell typ av databas som ibland används och som kallas för MDBMS, *Multidimensional Database Management System*.

DWMS, *Data Warehouse Management System*, är den del av systemet som ansvarar för att ur produktionsdatabaserna automatiskt extrahera data och omforma det till beslutsinformation. Det finns idag speciella DWMS-produkter som tar hand om alla steg i denna process, men en del företag väljer att utveckla motsvarande funktioner själva för att skräddarsy lösningen för den egna organisationen. DWMS-produkter innehåller funktioner för att definiera hur data ska översättas och transformeras till beslutstödsinformation. Utifrån det skapas automatiskt ett program, ofta ett Cobol-program, som sedan sköter transformationerna.

En viktig komponent i ett DWMS är *metadata*, det vill säga information om data i produktionssystemen. Metadata behövs för att översättningen från produktionsdata till beslutstödsinformation ska kunna göras.

Ett problem för närvarande är att metadatahantering inte är standardiserad. Produktionssystemen innehåller ofta någon form av datakatalog med eget format. DWMS-produkter har sina egna metadatakataloger. Dessutom använder beslut-stödsprogram ofta någon form av intern metadatahantering. Det innebär att inkonsistenser kan inträffa om metadata ändras.

3 Beslutstöd

I detta avsnitt ger vi en översikt av de alternativ som finns för användarna av ett datavaruhus. I princip kan beslutstödsprogrammen delas upp i tre kategorier:

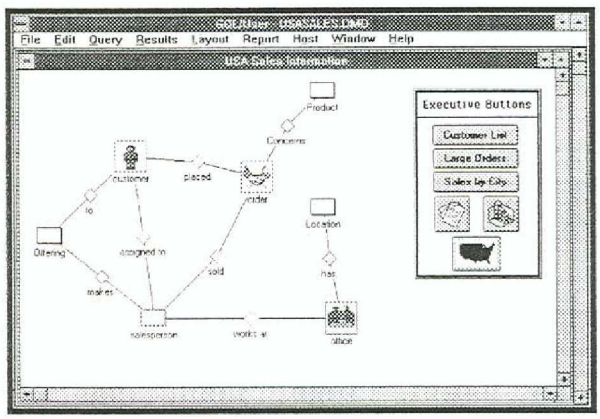
- Verktyg för ad-hoc-frågor.
- Flerdimensionella analysverktyg.
- Skräddarsydda Executive Information System.

I följande sektioner går vi igenom vad som karaktäriserar dessa olika produktkategorier.

3.1 Verktyg för ad-hoc frågor

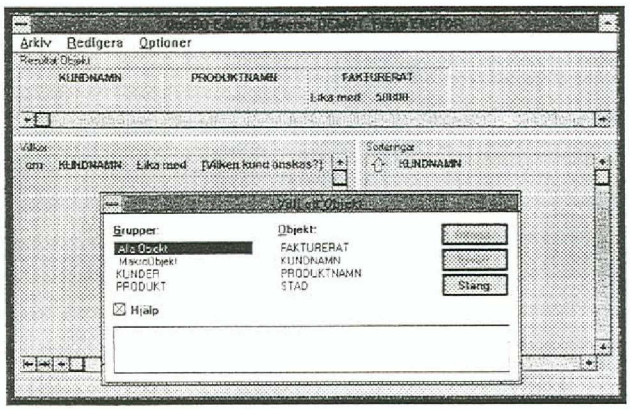
Med en ad-hoc fråga, eller spontan fråga, avses en fråga mot en databas som användaren själv formulerar utan att utnyttja fördefinierade frågor eller rapporter. Tidigare var användarna tvungna att skriva frågor direkt i SQL, vilket i allmänhet är för svårt och komplicerat för normala användare.

Nu finns andra, enklare alternativ. Det vanligaste idag är att erbjuda någon form av grafiskt användarsnitt baserat på *visuell frågeteknik*. Exempel på sådana produkter är *GQL* och *Dataprism* och SISUs egna projekt *Hybris* och *Intuitive*.



Figur 2. Ett exempel på ett program för spontana frågor som baseras på visuell frågeteknik.

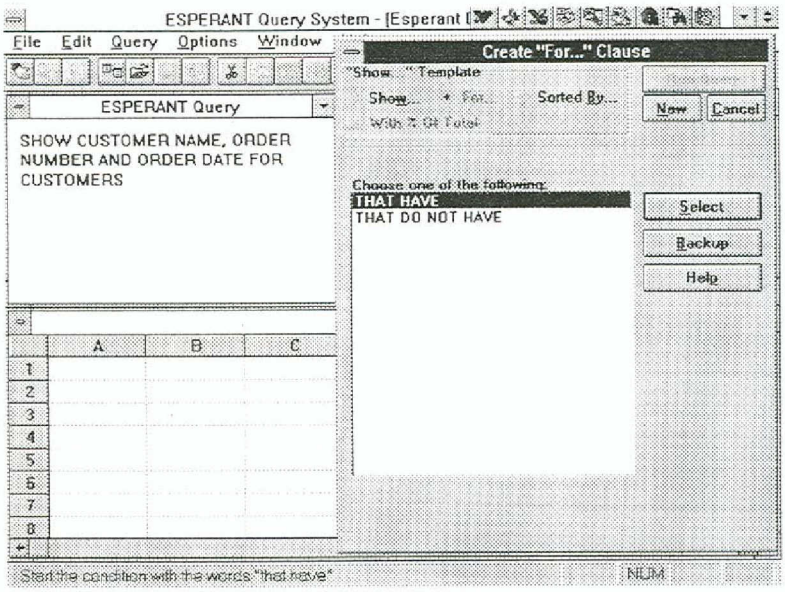
Figur 2 visar ett exempel på ett gränssnitt baserat på visuell frågeteknik. Användaren formulerar frågor genom att peka och klicka i en datamodell.



Figur 3. Att ställa frågor med Business Objects.

Business Objects är ett franskt program för ad-hoc-sökning i databaser. Det är inte i någon större utsträckning baserat på visuell frågeteknik utan arbetar istället med fördefinierade verksamhetsobjekt som kan kombineras på olika sätt, se Figur 3.

Ett annat alternativ är att erbjuda ett naturligt språkgränssnitt. Exempel på sådana produkter är *Esperant* och *Quercus*. Figur 4 visar ett exempel på hur ett sådant användargränssnitt kan se ut.



Figur 4. Esperant erbjuder användaren att formulera ad-hoc-frågor direkt i naturligt språk.

3.2 Flerdimensionell analys

Det område inom datavaruhusmarknaden som utvecklas snabbast just nu är slutanvändarverktyg för *flerdimensionell analys*. Till viss del kan man säga att det är verktyg för flerdimensionell analys som gjort att datavaruhusmarknaden tagit fart.

En undersökning som industrianalytikerna Meta Group nyligen gjorde i USA visade att 90 procent av tillfrågade datachefer övervägde eller redan höll på att utveckla ett datavaruhus och att 65 procent av dessa ansåg att flerdimensionell analys hade högsta prioritet.

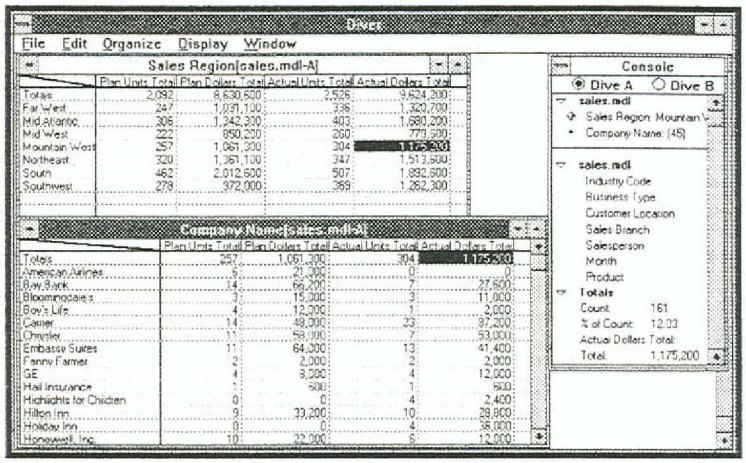
Några av de mest populära verktygen är *DSS Agent, IA Analysis, PowerPlay, Pablo* och *Crosstarget*. Ett verktyg värt att notera är *HAT* (Highspeed Analysis Tool) som är ett svenskutvecklat program. För att ge inblick i hur ett flerdimensionellt analysprogram fungerar ger vi här ett översiktligt exempel från Crosstarget. Programmen skiljer sig åt i detaljerna men erbjuder likartad funktionalitet.

Nyckelordet för den här typen av verktyg är *drill-down browsing*, eller *dive-down browsing*. Med det avses att man succesivt kan bryta ned data från helhet till detaljer. En annan viktig uppgift för ett analysprogram är att kunna göra prognoser.

Anta att vi vill arbeta med försäljningsinformation. På lägsta detaljnivå har vi då följande information – kund, region, produktnummer, säljare, belopp, datum etc. Crosstarget tar in sådan information som en flat fil och bygger sedan upp ett eget sökindex.

En användare kan nu vända och vrida på denna information, t ex titta på försäljningsbelopp per region, per säljare, per månad etc. För en viss region kan man sedan gå ned ett steg och titta på hur försäljningsbeloppet fördelar sig per produkt, per säljare, per kund etc. För en viss kund i en viss region kan man sedan gå ned

ytterligare ett steg och titta på hur den kundförsäljningen fördelar sig per produkt, säljare, månad et c. Figur 5 visar hur försäljningen i region "Mountain West" fördelar sig per kund i området.

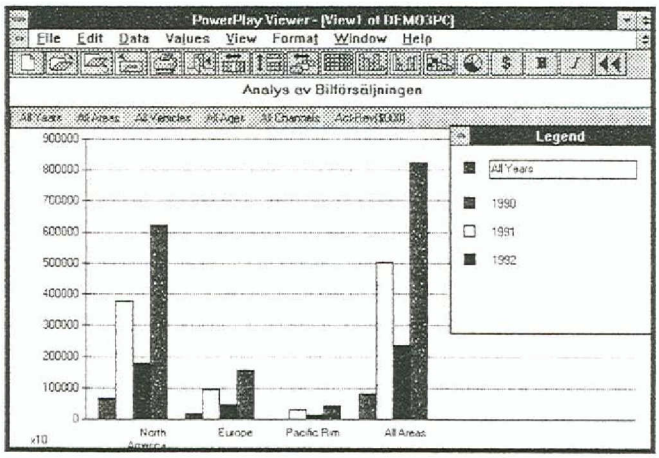


Figur 5. Ett exempel på verktyg för analys av data.

Användaren kan sedan gå vidare och bryta ned försäljningen per produkt för en viss kund i region "Mountain West". Ett analysprogram kan också presentera enklare typer av diagram som staplar, cirklar etc. Se Figur 6 som visar ett exempel från PowerPlay.

Enklare analysverktyg som PowerPlay, Crosstarget och HAT arbetar med extract av data som lagras på ett för verktyget eget format. Mer avancerade verktyg som DSS Agent och IA Analysis arbetar istället direkt mot en SQL-databas.

Typiska uppgifter som kan utföras med ett program för flerdimensionell analys är trendanalyser, avvikelseanalys, försäljningsuppföljning med mera.



Figur 6. Ett program för dataanalys har ofta möjlighet att visa enklare typ av grafik. Här ett exempel från PowerPlay.

3.3 Executive Information Systems

Ett begrepp som var en föregångare till datavaruhus är EIS, *Executive Information System*. Med ett EIS-system menas ett skräddarsytt beslutstödssystem. EIS-system kan implementeras på olika sätt. Verktyg som *Gentium*, *Lightship*, *Forest & Trees*, *Holos* och *Commander* är specialiserade för att utveckla EIS-system. Av de nämnda är Gentium, Holos och Commander stora allomfattande verktyg.

Det nyaste av dessa är Gentium som helt bygger på objekt-orienterad teknik. Gentium utvecklades ursprungligen för Next-datorer men har senare flyttats till Windows. Ett program utvecklat med Gentium kan enkelt flyttas från Windows till flera andra miljöer, t ex Macintosh. Gentium innehåller en egen flerdimensionell databas, se avsnitt 4.2.

Lightship och Forest & Trees kan ses som objektorienterade bygglådor med mindre byggstenar. I Lightship kan utvecklaren t ex ur en meny välja "tabellobjekt" och placera ut det på skärmen. Sedan kopplas tabellobjektet till en databas och en databasfråga specificeras. Till ett fält i tabellobjektet kan sedan ett diagramobjekt kopplas så att ett stapeldiagram ritas upp när en tabell fylls på med data.

Detta kan givetvis också åstadkommas med hjälp av verktyg som Visual Basic utvidgat med databaskopplingar. Skillnaden är att Lightship är mer specialiserat och erbjuder högre utvecklingseffektivitet för den här typen av system.

Statistikprogrammet *SAS* har under åren sakta men säkert byggts ut och SAS används ofta för att bygga EIS-system, speciellt i stora företag.

4 Databasteknik

En viktig komponent i ett datavaruhus är själva databasen. Datavaruhus ställer nya krav på databashanterare. Dagens traditionella databashanterare har utvecklats och optimerats för *transaktionshantering*. Det innebär till exempel att stor möda lagts ned på att utveckla algoritmer för hantering av kolliderande uppdatering, det vill säga när två tillämpningsprogram samtidigt försöker skriva i en datapost.

I ett datavaruhus finns inte sådana problem eftersom användarna och tillämpningsprogrammen enbart läser data ur databasen. Detta innebär att en stor del av koden i en databashanterare inte behöver exekveras och att prestanda därmed kan förbättras.

Datavaruhus är dock ofta mycket omfattande. De flesta databashanterare klarar upp till 20-30 GByte stora databaser utan problem. Men många företag som implementerar datavaruhus har upptäckt att man snart hamnar i storleken 500 GByte. Databashanteraren måste också snabbt kunna ge svar på komplexa frågor över dessa datamängder.

Ett begrepp som har dykt upp på senare tid är OLAP, *On-Line Analytic Processing*. Det har myntats av Ted Codd, mannen bakom relationsteorin. Med OLAP menas databasteknik som är avsett för analyser och beslutstöd. Detta är för att markera skillnaden gentemot transaktionshanteringssystem där begreppet OLTP, *On-Line Transaction Processing*, ofta används. Ted Codd har definierat 12 regler för vad som krävs av ett OLAP-system:

1. *Flerdimensionella konceptuella vyer*. En analytikers bild av en verksamhet är flerdimensionell, därför måste ett OLAP-system kunna hantera skärningar i olika dimensioner, pivoteringar och roteringar av data.
2. *Transparens*. De underliggande datastrukturerna, nätverk och arkitekturer ska vara osynligt för användaren.
3. *Åtkomlighet*. Det måste vara möjligt för ett OLAP-system att kunna använda data från såväl relationsdatabaser som från äldre system. Användaren ska inte behöva bekymra sig om varifrån data kommer.
4. *Konsistenta svarstider*. Ett OLAP-systems prestanda får inte bero på underliggande databas storlek eller det antal dimensioner användaren vill analysera.
5. *Klient-server arkitektur*. Ett OLAP-system måste kunna stödja en klient-server arkitektur.
6. *Generisk dimensionalitet*. Alla datadimensioner måste behandlas likvärdigt av systemet.
7. *Dynamisk och gles matrishantering*. Ett OLAP-system måste på ett optimalt sätt kunna representera samband mellan olika dimensioner.
8. *Stöd för flera samtidiga användare*. Ett OLAP-system måste kunna hantera samtidig åtkomst, integritet och säkerhet.
9. *Obegränsade kors-dimensionella operationer*. Det innebär att data ska kunna brytas ned eller aggregeras obegränsat antal gånger och samtidigt kunna korsrelateras till varandra. Anta vi har en dimension som är organisation (land, region, sektor, kontor) och en annan som är försäljning (år, kvartal, månad, vecka, dag). Då ska det vara möjligt att få fram alla tänkbara kombinationer, t ex månadsförsäljning per region eller dagförsäljning per kontor.

10. *Intuitiva möjligheter att manipulera data.* Zoomningar, drill-down browsing och andra typer av slutanvändarfunktioner måste understödjas.
11. *Flexibla rapporteringsfunktioner.* Det måste vara möjligt att få data presenterat med logiska grupperingar som är naturliga för olika användare.
12. *Obegränsat antal dimensioner och aggregeringsnivåer.*

Det finns idag tre olika sätt att realisera OLAP-funktionalitet i ett datavaruhus:

- Relationsdatabas, RDBMS och sedan överlåts OLAP-funktioner på tillämpningsprogrammen.
- Flerdimensionella databaser, MDBMS.
- Relationsdatabaser med OLAP-motor.

I de kommande avsnitten behandlar vi för och nackdelar med de olika teknikerna.

4.1 Relationsdatabaser, RDBMS

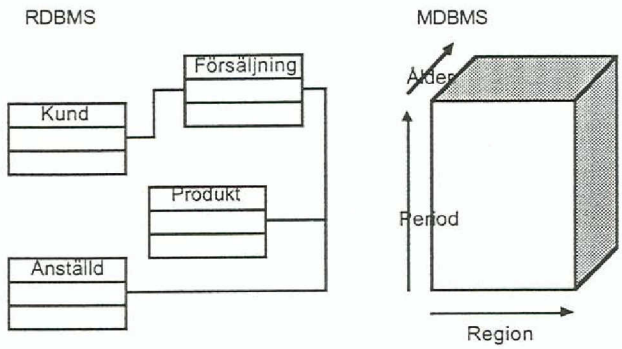
Datavaruhus i stordatormiljö byggs nästan uteslutande med DB2-databaser medan datavaruhus i Unix-miljö ofta baseras på Oracle eller Informix. Fördelen med att basera sitt datavaruhus på en relationsdatabas är att det är beprövad och standardiserad teknik och att det i organisationen i allmänhet finns kunskap om underhåll av relationsdatabaser.

Datavaruhus, åtminstone de stora, kräver i allmänhet en *parallell* databasarkitektur, vilket varit under utveckling under några år hos de stora databasföretagen. Just nu verkar det som om det är Oracle och Informix som klarar parallell databashantering i storlekar upp till 500 GByte på ett bra sätt. Ett exempel på ett databasföretag som har problem med sådana storlekar är Sybase som, enligt Meta Group, förlorat 40 procent av sin marknadsandel inom datavaruhusmarknaden på ett år på grund av att man inte klarar tillräckligt stora databaser.

4.2 Flerdimensionella databaser, MDBMS

De flesta anser att ett datavaruhus bör baseras på en relationsdatabas men det finns förespråkare för en annan teknik som kallas för *flerdimensionella databaser*. Det är databaser som kan representera fler datadimensioner än vad relationsdatabaser klarar av med sina två-dimensionella tabeller.

Fördelen med en flerdimensionell databas är att den har mycket bra prestanda för vissa typer av frågor som ofta förekommer vid analyser, t ex *visa försäljning per produkt och region för de senaste fem åren.*



Figur 7. En flerdimensionell databas är bättre anpassad för beslutstöd och analysarbete.

Figur 7 visar hur produktförsäljning kan representeras i tre olika dimensioner – kunders ålder, tidsperiod och region. I och med att data är representerat på detta sätt kan också komplexa frågor som rör samband i data mycket snabbt besvaras. Samma frågor kan förstås hanteras i en relationsdatabas men då måste specialskrivna program utvecklas som beräknar de olika sambanden.

Express från företaget IRI Software och *Essbase* från Arbor Software är två exempel på flerdimensionella databaser. Nackdelen med de flerdimensionella databaserna är att de inte är standardiserade. Varje leverantör har sin egen lösning.

4.3 RDBMS med OLAP-motor

Ett sätt att komma förbi problemen med OLAP-hantering i relationsdatabaser men ändå bibehålla säkerheten i att använda standardteknik är att lägga ett OLAP-lager ovanpå relationsdatabasen. Microstrategy och Information Advantage är två företag som valt den lösningen.

Deras produkter *DSS Agent* och *Axsys* tillhandahåller OLAP-motorer som kan generera SQL-frågor mot en relationsdatabas. Databasen har då designats enligt en "star-join schema"-princip vilket anses ge bra prestanda för flerdimensionell analys.

5 Konstruktion och underhåll av datavaruhus

I detta kapitel går vi igenom de olika tekniker som krävs för att konstruera ett datavaruhus. En viktig del i att bygga ett datavaruhus är förstås att analysera användarnas informationsbehov och deras verksamhet. Detta kommer vi inte att behandla närmare i denna rapport eftersom det finns många andra sådan rapporter, se till exempel *SISUs Modelleringshandbok*. Istället kommer vi att fokusera på de olika tekniker och produkter som finns för att skapa och underhålla själva varuhuset.

5.1 Extrahering och transformation av data

DWMS, *Data Warehouse Management System*, är den del av systemet som är ansvarigt för att ur produktionsdatabaserna automatiskt extrahera data och omforma det till beslutsinformation. Det finns idag speciella DWMS-produkter som tar hand om alla steg i denna process, men en del företag väljer att utveckla motsvarande funktioner själva för att skräddarsy lösningen för den egna organisationen. Exempel på DWMS-produkterna är *Prism Warehouse Manager*, *Carleton Passport* och *Extract Tool Suite*.

Sådana produkter innehåller funktioner för att definiera hur data ska översättas och transformeras till beslutstödsinformation. Utifrån det skapas automatisk ett program, ofta ett Cobol-program, som sedan sköter transformationerna.

Vare sig Warehouse Manager eller Passport innehåller en egen databas utan förlitar sig på SQL-standarden. När DWMS-programmen har bearbetat data färdigt levereras data som en uppsättning filer som sedan måste läsas in i den databas som utgör själva varuhuset.

Både Carleton och Prism tillhandahåller verktyg för slutanvändare att söka i metadata för att till exempel slå upp definitionen av ett visst informationsbegrepp. Ett sådant verktyg ingår i Carleton Passport medan Prism säljer en separat produkt, *Dictionary Manager*. Både Carleton och Prism lagrar sina metadata i en relationsdatabas. Databasstrukturen för detta är öppet tillgänglig för de som vill bygga egna verktyg eller tillämpningar som arbetar mot metadata. Se mer om sådana så kallade *informationskataloger* i nästa sektion.

Ett DWMS har också verktyg som hjälper systemadministratören att underhålla datavaruhuset och verktyg för att samla in statistik om datavaruhusets användande. Ett DWMS är ingen billig produkt. Prislappen hamnar runt 500 000 – 1 000 000 kr beroende på konfiguration.

5.2 Hantering av metadata

En viktig komponent i ett datavaruhus är *metadata*, det vill säga information om data i produktionssystemen. Metadata behövs för att översättningen från produktionsdata till beslutstödsinformation ska kunna göras.

Ett problem för närvarande är att metadatahantering inte är standardiserad. Produktionssystemen innehåller ofta någon form av datakatalog med eget format. DWMS-produkter har sina egna metadatakataloger. Dessutom använder beslutstödsprogram ofta någon form av intern metadatahantering. Det innebär att inkonsistenser kan inträffa om metadata ändras.

Den del av systemet som hanterar metadata brukar kallas för *informationskatalog*. Syftet med en informationskatalog är att underlätta utforskning, analys, åtkomst och hantering av information i ett datavaruhus. För slutanvändaren ska informationskatalogen göra det möjligt att förstå vilka data som finns tillgängliga i varuhuset. Katalogen ska hjälpa användaren att förstå betydelsen av dessa data, det vill säga vilka verksamhetsregler som ligger bakom, vilken aktualitet data har, vilka versioner av data som finns tillgängligt och varifrån data i varuhuset kommer.

En informationskatalog ska hjälpa informations- eller systemadministratören att optimera varuhuset för bättre prestanda, administrera säkerhetsfrågor och att ge bättre stöd till slutanvändaren. Det finns ett antal olika verktyg som producerar metadata som kan behöva stoppas in i en informationskatalog – exempelvis modelleringsverktyg, CASE-verktyg, repositories, extraherings- och transformeringsverktyg, datakataloger i relationsdatabaser och DWMS.

Det finns tre olika ansatser för att realisera informationskataloger:

- Dokumenthanteringssystem.
- Dataelementkataloger.
- Data Warehouse Manager.

En informationskatalog baserad på ett dokumenthanteringssystem tillhandahåller typiska funktioner för sökning på attribut, fritextsökning, säkerhetskontroll, versionshantering och arkiveringsfunktioner. Det finns ett par dokumenthanteringssystem som är speciellt avsedda för informationskataloger, *DataGuide/2* från IBM och *InfoHarvest* från Minerva. Det senare var ursprunligen en produkt från Digital men vidareutvecklades och säljs nu av ett separat bolag. Det går också att välja att realisera sin informationskatalog med ett generellt dokumenthanteringssystem som *Saros Mezzanie* eller *Xerox Documentum*.

En dataelementkatalog gör det möjligt för en användaren att bryta sig ner från en begreppsnivå till fysiska dataelement. Dataelementkataloger innehåller avbildningsregler, det vill säga hur operativa data översätts till varuhusinformation, som är tillgängliga för inspektion, begreppsförklaringar, tidstämplingar, namn på datakällor med mera.

Exempel på produkter som kan sägas vara dataelementkataloger är *Directory Manager* från Prism Solutions, *EasyView* från Brownstone och *Data Shopper* från Reltech. Brownstone och Reltech är två Repository-leverantörer. Reltech har nu köpts upp av företaget Platinum Technology som är mycket stora inom DB2-världen när det gäller programvara för underhåll, konstruktion, laddning med mera av DB2-databaser. Platinum satsar nu hårt på att profilera sig inom datavaruhusmarknaden.

Slutligen har vi Data Warehouse Managers som går ett steg längre och inte bara tillhandahåller information till slutanvändarna utan även innehåller funktioner för att underlätta administratörsarbete. En sådan produkt är *Intelligent Warehouse* från Hewlett Packard.

5.3 Generiska datamodeller

Ett sätt att snabbt komma igång med att bygga ett datavaruhus är att använda sig av en *generisk datamodell*. Det kan beskrivas som en informationsmodell som är generell för alla företag inom en viss bransch. Prism Solution som tycks vara en leverantör som ligger långt fram säljer också generiska datamodeller, *Inmon Generic Datamodels*. Det finns för olika industrier som bank/finans, sjukvård, försäkring, telekommunikation. Dessa modeller anpassas sedan för det specifika företaget.

6 Slutsatser

Datavaruhus har kommit för att stanna. Idéerna bakom datavaruhuset är inte nya. De förekom redan under 70-talet då MIS (Management Information System) skulle förse beslutsfattare med den information de behövde. På samma sätt dök begreppet EIS (Executive Information System) upp för 4-5 år sedan, men fick aldrig något riktigt genomslag. Det som talar för att datavaruhus faktiskt kommer att kunna klara av att leverera bra beslutstöd är dels att prestanda på maskinvara och databashanterare nu förbättrats så mycket att riktigt stora databaser är realistiskt, dels att användarverktygen är så pass mycket bättre än för några år sedan.

Analysföretaget Gartner Group förutspår att en drastisk förändring kommer att ske redan 1996 när det gäller företagsanalyser och konkurrentbevakning. Istället för att ett fåtal speciella analytiker ägnar 100 procent av sin tid åt bevakning och analys kommer alla chefer på alla nivåer att ägna 10 procent av sin tid åt sådana frågor. Utvecklingen går definitivt åt det hållet, huruvida det exakt kommer att inträffa 1996 får framtiden utvisa.

Helt klart är att dataavdelningar och interna databolag kommer att få ägna mycket tid framöver åt att bygga datavaruhus. Analysföretaget *Meta Group*, som specialiserat sig på data warehousing, går så långt att de hävdar att datavaruhus är det enda IT-strategiska som dataavdelningar kommer att hålla på med framöver. Argumenten för det är att dataavdelningens kunskap och kompetens kring de operativa systemen behövs för att bygga datavaruhusen och att varuhuset har en direkt koppling till affärsverksamheten och därför *inte* kan *outsourcas*.

Vår uppfattning är att datavaruhus bör baseras på relationsdatabaser och inte på någon ny typ av databasarkitektur, typ MDBMS. De stora relationsdatabasföretagen som Oracle och Informix kommer i långa loppet att kunna leverera prestanda som överskrider MDBMS även för OLAP-hantering.

Datavaruhusbranschen är en av de hetaste just nu och dessutom mycket turbulent. Många relativt nystartade företag agerar på samma spelplan som jättar som Oracle och IBM. Vi kommer med säkerhet få se en del företag slås ut, andra köpas ut och nya konstellationer bildas.

Troligtvis kommer de stora databasföretagen snart att agera och köpa upp några datavaruhusföretag för att snabbt komma över nödvändig teknologi för till exempel OLAP-hantering. Vi kommer säkert också snart att få se mängder av konsultbolag som erbjuder sina tjänster att bygga datavaruhus.

En sak som kan ifrågasättas kring datavaruhus är hur beslutsfattande i företag går till. Underförstått i hela resonemanget kring datavaruhus är att beslut fattas baserat på historisk information om verksamheten. Exempelvis att någon studerar försäljning på Europa-basis och bryter ned den till lands-nivå och slutligen hittar en region i något land där försäljningen gått sämre än väntat och sedan fattar ett beslut om någon åtgärd.

Ofta är det så, men i många företag, speciellt i högteknologiska branscher är information om omvärlden, det vill säga konkurrenter, tekniska trender, politiska strömningar etc ett viktigare beslutsunderlag. Det spelar till exempel inte så stor roll för ett programföretag hur försäljningen går i norra Spanien om man får reda på att Microsoft inom ett år kommer att lansera en konkurrerande produkt till halva priset. Då gäller det att snabbt hitta en ny strategi att möta detta hot.

Framöver kommer vi antagligen också att få se datavaruhus som inte är begränsade till enbart hantering av siffror och textsträngar, utan företagens information kommer att finnas tillgänglig i textdokument, och även som bilder och grafer. Det förefaller naturligt att inkludera även denna typ av information i ett datavaruhus. Detta ställer dock nya krav på såväl nya databasarkitekturer som på användarverktyg för sökning och analys. Vi kan då börja tala om *Multimedia Warehousing*.

7 Produktsammanfattning

Det finns en bred flora av produkter som kan användas för att bygga datavaruhus och för att ge slutanvändarna olika typer av beslutstöd. I tabellen nedan sammanfattar vi de produkter som vi tagit upp i denna rapport.

Produkt	*Kommentarer*	*Tillverkare*
Axsys	Utvecklingsverktyg för OLAP-system. Innehåller OLAP-motor som kan generera SQL-frågor.	Information Advantage
Business Objects	Ad-hoc-frågor baserade på verksamhetsobjekt, döljer databasschema, genomtänkt Dictionary-hantering.	Business Objects
Clear Access	Dataåtkomst för användare och andra program.	Clear Access
Crosstarget	Flerdimensionell analys	Dimensional Insight
Data Selector	Frågeverktyg som används tillsammans med Data Shopper.	Platinum Technology
Data Shopper	Dataelementkatalog som arbetar mot Platinums Repository.	Platinum Technology
DataGuide/2	Informationskatalog baserad på dokumenthanteringssystem.	IBM
Dataprism	Ad-hoc-frågor baserade på tabeller, parametriserade frågor.	Brio Data
Directory Manager	Informationskatalog med verktyg för att söka i metadata.	Prism Solutions
DSS Agent	Flerdimensionell dataanalys. Innehåller OLAP-motor som kan generera SQL-frågor.	MicroStrategy
Easy View	Dataelementkatalog.	Brownstone.
Esperant	Naturligt språkgränssnitt för ad-hoc-frågor. Stödjer summeringar och grupperingar.	Software AG.
Essbase	Multidimensionell databas som kombinera kalkylprogramsteknik med databasteknik	Arbor Software.
Extract Tool Suite	Data Warehouse Management System	Evolutionary Technologies.
Forest&Trees	Bygglåda för EIS-system.	Forest&Trees
Gentium	Utvecklingsverktyg för EIS-system.	Planning Science
GQL	Grafiska modeller, verksamhetsnära, parameteriserade frågor, fyra moduler, översikt, många databaskopplingar	Andyne Computing
Holos	Stort utvecklingsverktyg för EIS-system.	Holistic Software
IA Decission Support Suite	Uppsättning verktyg för beslutstöd. Innehåller ad-hoc frågor, flerdimensionell analys och standardrapportfunktioner.	Information Advantage
Impromptu	Ad-hoc-frågor, koppling till PowerPlay.	Cognos
InfoHarvest	Informationskatalog baserad på dokumenthanteringssystem.	Minerva
Inmon Generic Datamodels	Färdiga datamodeller för datavaruhus. Finns för olika industrier som bank/finans, sjukvård, försäkring, telekommunikation.	Prism Solutions
Intelligent Warehouse	Informationskatalog med stöd för att administrera användningen av ett datavaruhus.	Hewlett Packard
Lightship	Bygglåda för EIS-system.	Pilot Software
Pablo	Flerdimensionell analys	Andyne Computing
Passport	Data Warehouse Management System. Funktioner för att extrahera och transformera operativa data till varuhusinformation. Genererar metadata som kan lagras i en relationsdatabas.	Carleton
Platinum Repository	Metadatakatalog.	Platinum Technology
Powerplay	Multidimensionell analys.	Cognos

SAS	Utvecklingsverktyg som gör det möjligt att bygga EIS-system och datavaruhussystem. Ursprungligen ett statistikprogram som nu används för att bygga olika typer av beslutstödstillämpningar, i framför allt stora företag.	SAS Institute.
Quercus	Naturligt språkgränssnitt för ad-hoc frågor.	New Business Information
ViewPoint	Ad-hoc-frågor, bra Dictionary-hantering, enbart Informix-koppling, multimediastöd, Motif-gränssnitt.	Informix
Warehouse Manager	Data Warehouse Management System. Funktioner för att extrahera och transformera operativa data till varuhusinformation. Genererar metadata som kan lagras i en relationsdatabas.	Prism Solutions
VPT	Speciell databas för datavaruhus	Red Brick

Appendix: Ordlista

Ad-hoc Query Tools	En speciell typ av frågeverktyg som förväntas få allt större betydelse i bland annat data warehouse-sammanhang. Ett sådant verktyg tillåter användarna att ställa ad-hoc-frågor, dvs ej förprogrammerade frågor, mot databaser. Verktygen arbetar ofta med någon form av visuell frågeteknik, ett fåtal bygger på naturlig språkförståelse.
Data Mart	Ofta kan ett datavaruhusprojekt startas med en delmängd av data, t ex en avdelnings data. Då brukar man tala om en "Data Mart" snarare än ett "Data Warehouse".
Data Mining	Speciella system som med hjälp av AI-teknik försöker hitta tidigare okända samband i stora mängder gamla data.
Data Warehouse	Ett ämnesorienterat informationslager utformat specifikt för beslutstödstillämpningar.
Data Warehousing	Processen att utveckla, införa och underhålla ett datavaruhus.
Drill-down browsing	Att interaktivt bryta ned data till olika detaljeringsgrader.
DSS	Decision Support System, ett kärt begrepp som vägrar dö. Ursprungligen stod begreppet för system som försökte automatisera beslutsfattande. Nu syftar det mer till system som kan bidra till att ge underlag för att fatta beslut.
DWMS	Speciella datavaruhushanterare. Sköter extrahering av data från produktionssystem och översätter dessa till beslutsinformation.
EIS	Executive Information System, en typ av informationssystem som är inriktat på att ge beslutsfattare rätt information.
Information Warehouse	IBM var först ut på plan och myntade begreppet Information Warehouse och varumärkesskyddade det, varför de andra får hålla tillgodo med begreppet Data Warehouse.
Informationskatalog	Databas som lagrar metadata.
MDBMS	Multidimensional DBMS. Speciell typ av databas för beslutstödstillämpningar. Kan utgöra kärnan i ett datavaruhus.
Metadata	Information om data.
OLAP	On-line Analytical Processing, syftar på en speciell typ av databasarkitektur som försöker utvidga relationsmodellen till flera dimensioner, ungefär som fler-dimensionella kalkylark. Ett OLAP-system kan dock använda en relationsdatabas för själva lagringen av data.
OLTP	On-Line Transactional Processing. Transaktionshantering.
RDMBS	Relationsdatabashanterare.
Visuella frågesystem	Användargränssnitt som systematiskt utnyttjar visualiseringar för att hjälpa användare utnyttja databasen.