

Reseberättelse från konferensen

DAUG - Strategies & Resources: The Repository/Data Warehouse Connection!
Atlantic City, 21 - 24 maj 1995

Intryck och kommentarer sammanställda av Teddy Hector, Posten och Stig Berild, Sveriges Tekniska Attacheer, San Francisco.

(En tidigare DAUG-konferens, hösten 1994, bevakades för Telias och Postens specifika räkning. Med dessa organisationers tillåtelse har vissa textdelar nedan hämtats från denna reseberättelse. Det gäller synpunkter av mer generell karaktär som fortfarande bedöms relevanta och värda att föra ut till en vidare läsekrets.)

1 Begreppet Data Warehouse

Liksom så många andra företeelser inom IT-området har även Data Warehouse genomgått en berg- och dalbana under sin korta historia. Från ett buzzword för en trend eller ett framväxande behov, över massmedialt intensivt intresse med åtföljande mångfacetterade tolkningar, via framväxten av mer eller mindre självutnämnda språkrör och gurus på världsturnéer och i författarskap, till ett spirande nyfiket kundintresse och förstagenerationens stapplande produkter, in i besvikelsens och misslyckandenas mörker, följt av konsolidering och förnuft, samt påbörjad tung men seriösare och mer rimlig väg tillbaka till en andra generationens realism i förståelse, ansats, produkter och genomförande.

Begreppet Data Warehouse börjar röna ett allt större intresse igen. Det är klart på väg ut ur svackan och in i denna andra fas. Huvudsaklig drivkraft är inte längre massmedia och gurus utan de alldeles påtagliga bekymmer, behov och möjligheter begreppet satts att representera.

Industriella drivkrafter som ökad konkurrens, kortare ledtider, snabbare reaktion på förändringar av marknadsindikatorer, mer krävande kunder, kortare produktlivscykler, internationellare konkurrensförutsättningar, mm pressar på behovet av att ha fullödiga, välserverade beslutunderlag.

För övrigt är området Data Warehouse förvånansvärt lite guru-fixerat (undantag Bill Inmon) och desto mer gräsrotsöppet. Området kännetecknas också betydligt mer av pragmatiska ansatser baserade på sunt förnuft och konkreta, stegvis alltmer ambitiösa mål än på lustfyllda visioner och hägrande långtidsperspektiv.

Ämnesområdet som sådant är ju också så nytt att det knappast går att tala om någon högre kunskapsnivå hos någon, utom möjligen hos det fåtal som hunnit skaffa sig reella erfarenheter. Den påtagliga tillämpningsorienteringen varken kräver eller söker teoribildning. Ett i högsta grad tvärvetenskapligt täckningsområde hämtar nödvändig stabilitet och standarder från områden som

databaser, användargränssnitt, informationsmodellering, AI, öppna systemarkitekturer i kombination med avancerade tekniska lösningar.

2. Konferensen i översikt

DAUG är en förkortning av Data Administration User's Group, en organisation som tycks verka som en slags lös samlingsorganisation för ett antal lokala DAMA-organisationer spridda i USA och resten av världen. En avdelning finns förresten i Sverige. DAUGs enda uppgift tycks vara att två gånger om året anordna konferenser inom området dataadministration och närbesläktade områden på olika platser i USA. Nästa konferens går av stapeln i San Francisco, 19 - 22 november 1995. Drivande eldsjäl är Dick Naur.

Vår-konferensen lockade ca 200 deltagare, till alldeles övervägande delen från USA. I en ganska avspänd atmosfär redovisas åsikter, erfarenheter, råd och bekymmer inom konferenstemat. En utställning gav en kompletterande bild av vilka produkter som finns att tillgå och deras kapacitet.

Konferensrubriken pekade ut en temainriktning mot att utnyttja Repositorier eller resurskataloger som stöd för Data Warehouse-arbete. Som vid de flesta andra konferenser är rubrik och innehåll inte alltid överensstämmande. De flesta föredragen var rent DW-inriktade, några berörde repositoryer och ytterligare några hade seriöst nog anammat rubriken genom att redovisa ett samspel.

Det kan vara värt att påpeka att flera av presentatörerna inte gjorde någon distinktion mellan DW och Repository. Det senare verkade kunna vara en avstjälpningsplats för både det ena och det andra.

Enligt den vägledande uppfattningen innehåller DW explicita data hämtade eller härledda från verksamhetens operativa databaser. Repositoryt däremot innehåller uppgifter som beskriver och förklarar en verksamhet och verksamhetens informationssystem. Bland uppgifterna i ett repository kan finnas beskrivningar av vilka typer av uppgifter som de operativa och DW-databaserna omfattar, d.v.s. deras databasscheman, varifrån data härstammar, hur, när, vem, ... som genererar och uppdaterar och med vilka data.

Intressant att notera: Genomgående diskuterades DW utifrån perspektivet att det representerade kopierade data, inte aggregeringar, analyser i operativa databaser. Detta tycks vara en klar trend. Tidigare har man ansett att det är en avvägningsfråga som får lösas från fall till fall. Man har helt enkelt funnit att i det generella fallet DW-tillämpningar mot operativa data alltför mycket stör och försämrar prestanda i den operativa driften samt att ansvarsfrågor och prioriteringar kan bli komplicerade. Dessutom arbetar DW delvis helt andra uppgifter som historiska data och olika grad av sammanställda, filtrerade data. Läger man därtill perspektivet att DW-data kan beröra data i alla upptänkliga databaser i en organisation representerande olika ansvar, ålder, kvalitet, semantik, aktualitet, överlappningar, mm förstår man lätt att DW inte bara är en ny tillämpning att addera till de tidigare. DW innebär ett helt nytt sätt att arbeta med alla de data som finns representerade i en verksamhet och med syftet att generera ny kunskap snarare än att operera på data i dess grundläggande form.

Årets konferens kom att kretsa mycket just kring vikten av semantik och kvalitet.

Den helt avgörande semantiksäkringen när diversifierad information från många källor ska sättas att samverka tycks dock inte vara en källa till akut bekymmer, antagligen därför att DW-databaserna än så länge arbetar med begränsad information, dessutom ofta extrakt ur en viss tillämpning eller databas. Någon gav dock uttryck för att det var just semantikproblematiken som skulle komma att bli den svåra nöten att knäcka framöver, när ambitionsnivåerna växer.

Hand i hand med semantiken går kvalitetsaspekten. Otydlig semantik vid källan får genast återverkan i form av lågkvalitativ output från DW-tillämpningen. Även med gott grepp om hur data ska tolkas har de genererats vid olika tidpunkter av olika tillämpningar och under vitt skilda förhållanden för övrigt. "Garbage in" blir i DW-sammanhang "garbage out" i kubik. Läger vi därutöver till att vi förmodligen har mycket liten kontroll över och kunskap om de principer och villkor de program som hämtar in och uppdaterar DW arbetar efter, speciellt om de köpts in på öppna marknaden, kan vi enkelt konstatera att DW är en riskfull aktivitet som kräver "tungan rätt i mun". Eller som någon uttryckte det "Poor data quality - user dissatisfaction - disuse - warehouse failure."

Alldeles klart kan också konstateras att antalet så kallade "success stories" ännu är synnerligen begränsat. De som tycks lyckas bäst är de som startar i det lilla, dvs utan alltför mycket problem med semantik, kvalitet eller teknik, men som ändå stödjer ett påtagligt behov, dvs har en intresserad och angelägen användarkrets.

För den som är i färd med att ta steget ut i praktisk handling rekommenderades med andra ord att satsa försiktigt men övertänkt. Förankra, ta lärdom, se helhetsekonomin. Under förutsättning att det första steget faller väl ut, expandera stegvis och med täta resultatåterkopplingar. Eller som någon uttryckte det: "Architect globally, implement by increment".

DW ser visserligen förledande enkelt och klart ut i sina principer men är i realiteten mycket komplext, med många fallgropar. Det är i mindre utsträckning en fråga om teknik än om "reculturing" av människor och organisation, om att arbeta med information på ett delvis nytt sätt under nya krav och i ett intensivt samspel.

Än så länge är DW-området en ganska obruten mark, en "fixarens marknad". Specifika standarder saknas, exempelvis för gränssnitt, databashantering, datatransporter till kopiedatabaser, mm.

Frågot exv. hur mycket ska finnas i DW, hur länge ska det ligga kvar. Ska det finnas i förberäknat och kondenserat skick eller beräknas vid behov? Hur flexibelt ska man kunna borra ner i sammanställningsnivåerna för att finna grunddata? Hur bör ostrukturerade data hanteras. Fritextsökning? Hur hantera multimediaminformation, exv för analyser baserade på pattern recognition?

Sammanfattningsvis: Alla tycks inse behovet, de flesta har en klar vision, men vägen dit tycks ännu vara stapplande och fylld med mer frågetecken än utropstecken.

2.2 Repository

Repository-området fick en rejäl törn i och med att IBM drog tillbaka sin mainframe-satsning. Förest hade man basunerat ut sin AD/Cycle strategi i slutet av 80-talet och därmed, genom sin "inmutning", tagit loven av andras satsningar inom området. Marknaden inrättade sig i ett vänteläge. När sedan produkter tillsammans med förväntade de facto-standards uteblev, inträdde dels ett produkt- och kunskapsvakuum, dels en allmän besvikelse bland dem som aktivt argumenterat behoven inom sina organisationer. Beslutsfattare och budgetansvariga drog sig undan.

I ärlighetens namn ska också påpekas att de tänkta användarna i de flesta fall inte var mogna att på ett konsekvent sätt hantera denna typ av information.

Kvar fanns dock behoven av att kunna hantera information om en verksamhet och dess informationssystem på ett mer eller mindre samordnat sätt. Kunskapen om hur hanteringen ska läggas upp har ökat. Några av produktleverantörerna har övervintrat och finslipat sina produkter eller snarare anpassat dem till mer realistiska förutsättningar. Dit hör även IBM.

Användningsområdena har expanderat från primärt ett samordnat stöd för systemutveckling och underhåll. Trenden mot öppna system, samordning och komplexare informationshantering ställer ökande krav på kunskap om uppbyggnad, funktion, ansvar kring den inblandade informationsprocessen. Dessa uppgifter passar utmärkt i ett repository. Så gör också allhanda information om verksamheten i sig i och för dess vidareutveckling och anpassning. Detta nya innehåll kan, rätt utformat, ge ett effektivt stöd både vid utveckling av en DW-funktionalitet och som aktiv vägledning vid dess användning. Flera ansåg att en flexibel användning av ett DW helt enkelt kräver ett väl stödjande repository som kunskapsbas, för analys och för vägledning. I nästa steg kommer kravet på fullt aktiva repositorer, dvs sådana som i princip innehåller all processbeskrivning och alla de uppgifter som behövs för att kunna ge avsedd service.

Sammanfattningsvis: Många anser nu tiden mogen för en andra, mer lyckosam fas, för repository-området. Behov och verktyg finns. Kunskapsnivån ökar stadigt. Användningsområdena blir fler.

xx. Införandeproblematiken

Övertyga.
Sälj, sälj, sälj.
Var realist.

Välj påtagliga första-tillämpningar.
Få rejäl färstabudget. Försök inte sälja på att det är billigt utan att det är nödvändigt.
Etablera en kompetent organisation.
Sprid kunskap om att servicen finns tillgänglig. Locka och övertyga.
Följ upp, redovisa.

10 mistakes

xx. Något om produktrender

3. Sammanfattning

Ökat samspel mellan DW och repository.

Avgörande för bådars användbarhet är hög kvalitet. Fokus på semantik, ansvar och ägarskap till data

Dataexplosionen måste kunna tas om hand. DW har möjlighet fylla roll som filter, kunskapskälla, flexibelt stöd.

En lustig paradox. I tillverkande industri vill man undvika lager. Inom informationsbehandling önskar man tvärtom ha så mycket som möjligt i lager. Kanske borde vi anpassa vårt DW mer till specifika behov. Att spara allt blir tungt och kan vara ett tecken på osäkerhet. Å andra sidan utvecklas verksamheten så snabbt att det i realiteten är omöjligt att veta alla behov.

"Tell me what to look for and I will tell you what I need."