Datakvalitetsverktyg i datalager

– Hur hanterar datakvalitetsverktyg datakvalitetsfaktorer?

(HS-IDA-EA-03-401)

Veronica Ahl (a00verah@ida.his.se)

Institutionen för datavetenskap
Högskolan i Skövde, Box 408
S-54128 Skövde, SWEDEN

Examensarbete på det dataekonomiska programmet under vårterminen 2003.

Handledare: Jesper Holgersson

Datakvalitetsverktyg i datalager

–Hur hanterar datakvalitetsverktyg datakvalitetsfaktorer?

Examensrapport inlämnad av Veronica Ahl till Högskolan i Skövde, för Kandidatexamen (B.Sc.) vid Institutionen för Datavetenskap.

[2003-06-08]

Härmed intygas att allt material i denna rapport, vilket inte är mitt eget, har blivit tydligt identifierat och att inget material är inkluderat som tidigare använts för erhållande av annan examen.

Signerat: _____

Datakvalitetsverktyg i datalager

–Hur hanterar datakvalitetsverktyg datakvalitetsfaktorer?

Veronica Ahl (a00verah@ida.his.se)

Sammanfattning

I ett företag finns data utspridd i olika datasystem. För att lättare och snabbare komma åt data bör data finnas samlad på ett ställe d.v.s. ett datalager. Om datalagret ska fungera som ett beslutsstöd för företagsledningen är det viktigt att data håller en hög datakvalitetsnivå så att beslut fattas som gynnar företaget. För att kontrollera datakvalitet kan datakvalitetsverktyg användas. Dessa innehåller funktioner för att lösa olika datakvalitetsproblem. I rapporten behandlas hur datakvalitetsfaktorerna korrekthet, datatypsenlighet, konsistens, lämplighet, integration, kompletthet och inga dubbletter hanteras av datakvalitetsverktyg för datalager. En litteraturstudie har genomförts för att se hur datakvalitetsfaktorerna hanteras samt om behandlingen varierar mellan olika datakvalitetslösningar. Litteraturstudien har också använts för att upptäcka om olikheter och likheter i behandlingen av enskilda datakvalitetsfaktorer har påverkat hur väl faktorerna hanteras i lösningarna. Samtliga lösningar i arbetet innehåller datakvalitetsverktyg som beaktar de utvalda datakvalitetsfaktorerna men använder ibland olika tillvägagångssätt. Skillnader finns i hur väl de olika lösningarna hanterar enskilda datakvalitetsfaktorer.

Nyckelord: Datalager, Datakvalitet, Datakvalitetsverktyg, Datakvalitetslösning.

Innehållsförteckning

1 Inledning

Företag idag har ofta en stor mängd data inom sin verksamhet i ett antal olika datasystem. Dessa system växer ständigt i komplexitet beroende på att mängden data ständigt ökar i företaget. För att hantera denna växande mängd data som är utspridd i olika system behövs en samlingsplats för data. En sådan samlingsplats kan vara ett datalager som rymmer all data som ska finnas lätt tillgänglig för företagsledningen inom företaget. Ett datalager har som syfte att fungera som beslutsstöd åt företagsledningen, men ett datalager kan inte förväntas leverera tillfredsställande information om inte datalagrets data är korrekta, pålitliga och trovärdiga. Detta faktum bidrar även till att vikten av hög datakvalitet ökar då effektivitet och kvalitet i beslutsfattande är direkt kopplat till datakvalitet. Att datalagret håller en hög datakvalitet är viktigt för att användarna ska känna en trygghet i att använda datalagret. Om inte användarna kan lita på att informationen som hämtas från datalagret är riktig minskar trovärdigheten för datalagret och användarna slutar använda det. När företagsledningar insett att datakvalitet är ett viktigt faktum för att få korrekt och trovärdig data till beslutsfattandet kommer intresset för datakvalitetsverktyg att öka (Eckerson,2002).

Datakvalitetsverktyg i datalager används för att ge data hög datakvalitet vid införande i ett datalager. Det finns olika verktyg beroende på vilka datakvalitetsproblem som ska hanteras och de olika verktygen är också olika bra på att hantera olika datakvalitetsområden. Gemensamt för verktygen är att de genomför en omfattande granskning av data med hjälp av olika funktioner som exempelvis analysering, standardisering och matchning. Vilka funktioner de olika datakvalitetsverktygen behärskar varierar från verktyg till verktyg men gemensamt är att det inte finns ett verktyg som hanterar alla datakvalitetsfunktioner. Därför erbjuder företag som säljer datakvalitetsverktyg olika paketlösningar där ett antal verktyg har kombinerats till en datakvalitetslösning för datalager. Kombinationen ska möjliggöra att fler datakvalitetsbekymmer ska kunna hanteras på en gång. Samtidigt utvecklas verktygen kontinuerligt vilket användare bör vara medvetna om för att kunna ha en bra datakvalitetslösning som passar den verksamhet företaget verkar i. Utvärdering av datakvalitetsverktyg och deras funktioner behövs då för att ge en bild över vilka datakvalitetsproblem olika verktyg hanterar.

En medvetenhet om att data är en viktig tillgång bör finnas för att förstå vikten av en god datakvalitetsnivå i företagens datalager. Företag med denna medvetenhet kommer förmodligen med stor sannolikt att ha en större chans att överleva dagens tuffa och globala marknad.

2 Bakgrund

Data är en värdefull tillgång i företag. Ofta behöver beslut tas snabbt och det är då av vikt att data snabbt och enkelt går att få fram som underlag för att fatta ett riktigt beslut. Att söka efter information är tidsödande när mängden data är stor och utspridd i många olika system såväl inom som utanför det egna företaget. För att minska denna tidsåtgång behöver data vara samlad på ett ställe för att underlätta sökandet efter information. En sådan lösning kan vara ett så kallat datalager, som fungerar som en samlingsplats där data som är av intresse för företagsledningen vid beslutsfattande samlas. Att samla data på ett ställe för att använda som beslutsstöd förutsätter dels att data som presenteras är sanningsenlig gentemot verkligheten men även att presenterad data inte innehåller några felaktigheter som exempelvis stavfel. Felaktiga eller osanna uppgifter i data som ligger till grund för beslutsfattande kan leda till beslut som blir missgynnande för företaget. Följande kapitel kommer att ta upp några begrepp som belyser arbetets frågeställning och en beskrivning av det valda ämnesområdet för detta arbete.

2.1 Vad är ett datalager?

Ett datalager har som uppgift att fungera som beslutsstöd åt företagsledningen (Connolly & Begg, 2002). Data hämtas från olika system och sammanförs i datalagret. I datalagret lagras och organiseras data på ett sätt som senare möjliggör analys av data. Vid analys kan det vara önskvärt att ha tillgång till data som sträcker sig över en längre tidsperiod, vilket ett datalager gör möjligt då det innehåller såväl historisk som aktuell data (Inmon, 2002). Flera olika definitioner av datalager har gjorts som varierar eftersom datalager kan betyda olika saker för olika personer (Bischoff & Alexander, 1997). Oavsett vilken definition som används av datalager är målet detsamma och det är att integrera data som är utspridd till en och samma lagringsplats där användarna kan använda lagrad data till bl.a. rapporter och analyser (Connolly & Begg, 2002). Definitionen nedan är en översättning från engelska av hur Inmon (2002, sidan 31) definierar begreppet datalager.

> *”Ett datalager är en subjektorienterad, integrerad, beständig och tidsstämplad samling av data med syfte att stödja beslutsfattande.”*

Valet att använda ovanstående definition har gjorts då den ofta förekommer i litteratur om datalager, bl.a. använder Connolly och Begg (2002) ovanstående definition för att beskriva datalager och kallar samtidigt Inmon för datalagrets fader. Även enligt Bischoff och Alexander (1997) är definitionen ovan den klaraste beskrivningen av datalager. I detta arbete belyser denna definition de egenskaper ett datalager ska ha och återspeglar arbetets datakvalitetstänkande. Datalagrets egenskaper subjektorienterad, integrerad, beständig och tidsstämplad förklaras på ett utförligare sätt i följande kapitel.

2.1.1 Subjektorienterad

Med subjektorienterad menas att data i datalagret organiseras utefter företagets huvudfrågor som exempelvis kunder, produkter eller försäljning för att fungera som beslutsunderlag (Connolly & Begg, 2002). Vilka huvudfrågor datalagret organiseras runt varierar då huvudfrågorna skiljer mellan företag genom att alla har olika huvudfrågor beroende på vilket område företaget är verksamt i (Inmon, 2002). Att organisera efter företagets huvudfrågor skiljer datalager från källsystem som koncentreras på data runt företagets olika applikationsområden (Connolly & Begg, 2002). D.v.s. att källsystemen innefattar större områden av företagets data som är fokuserade på transaktioner medan subjektorienterad data i större utsträckning möjliggör analys av data i datalager. Det är därför lättare att finna den data som ska användas i beslutsfattande när data är organiserat runt huvudfrågor.

2.1.2 Integrerad

När data förs samman från olika källsystem säger Connolly och Begg (2002) att data ofta har olika format och integration behöver göras för att möjliggöra en konsistent presentation för användarna. Integration handlar om att kunna sammanföra data från källsystem som använder olika beteckningar för samma sorts data (King, 2000). Till exempel kan kön lagras som m/f eller 1/0 i källsystemen men användes istället X/Y för att beteckna kön i datalagret sker en konvertering av m/f och 1/0 till X/Y i samband med införandet till datalagret (Inmon 2002). Att datalagret ska vara integrerat anser Inmon (2002) vara den viktigaste aspekten av ett datalager och genom att använda samma beteckningar får data en entydig utformning,

2.1.3 Beständig

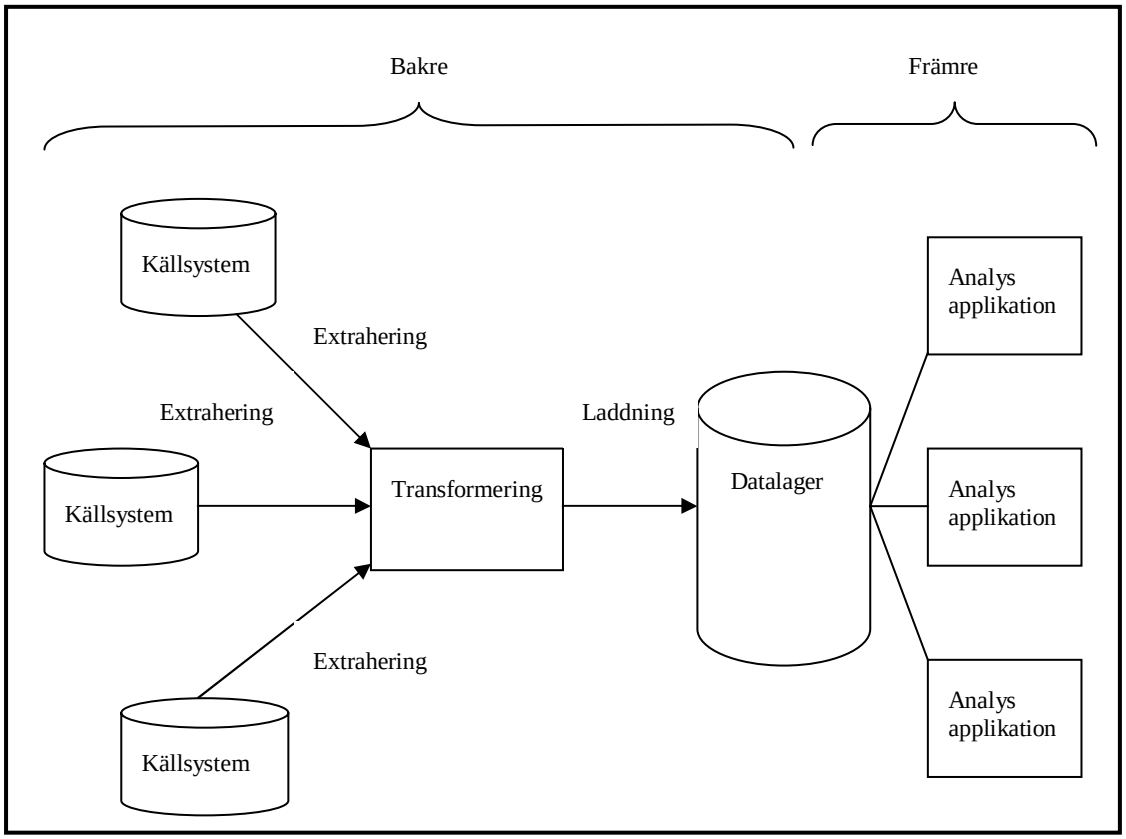
Beständig (eng. nonvolatile) innebär i detta sammanhang att data i datalager saknar behov av uppdateringar och när ny data läggs till integreras den snarare än ersätter existerande data (King, 2000). Det sker ständigt en kontinuerlig tillföring av ny data från källsystemen men istället för att byta ut data så läggs ny data till som supplement till den gamla (Connolly & Begg, 2002). På detta sätt menar Inmon (2002) att en historik över data sparas i datalagret.

2.1.4 Tidsstämplad

Begreppet tidsstämplat (eng. time-variant) innebär i datalagersammanhang en förmåga att lagra data från olika tidsperioder. Varje enhet av data i ett datalager är gällande för en viss period i tiden och för att visa vilken period som data är gällande görs tidsmärkning av data. Datamängden ökar i datalagret genom att inga uppdateringar görs utan data som lagras i datalager är som beskrivits tidigare beständig. Denna ökade mängd av data leder till att analyser över längre tidsperioder kan utföras (Inmon, 2002). Att datalager är tidsstämplat visas också genom den tid data hålls i datalagret (Connolly & Begg, 2002). Ett datalager kan innehålla data som sträcker sig fem till tio år tillbaka i tiden i jämförelse med källsystem som lagrar data i 60-90 dagar (Inmon, 2002).

2.2 Datalagerarkitektur

Ett datalagers arkitektur innefattar alla de element eller delar som ska ingå i ett datalager samt hur dessa är relaterade till varandra. Fokus ligger på att förbereda data och ladda datalagret samtidigt som en acceptabel nivå av datakvalitet hålls (King, 2000). Datakvalitet berör alla processer och alla de system som beskrivs i arkitekturen av ett datalager (Agosta, 2000). Arkitekturen fungerar som en implementationsplan för de olika delarna i datalagret (King, 2000). Kimball m.fl. (1998) delar upp arkitekturen i två större delar som i denna rapport kallas för bakre och främre. Den bakre delen är uppdelad i källsystem, extrahering, transformering och laddning medan den främre delen består av applikationssystem. En illustration över arkitekturens delar presenteras i figur 1.



Figur 1 Arkitektur av datalager (efter King 2000, sidan 40)

2.2.1 Källsystem

Data samlas ofta i någon form av databas vilket Elmasri och Navathe (2000) beskriver som en samling av relaterad data vars innehåll representerar olika aspekter ur verkligheten. Databasen designas, byggs och fylls med data för ett specifikt syfte. För att kunna hantera och få tillgång till data i databaserna kopplas ett program som kallas DBMS (databashanteringssystem) till databasen. Ett DBMS i ett företag som är avsedd att hantera en stor mängd transaktioner som dagligen äger rum i företaget kallas för OLTP (eng. online transaction processing) och är baserade på ”online-transaktioner” (Connolly & Begg, 2002). Vanligtvis har ett företag ett antal olika

OLTP-system och databaser som utgör källsystem till datalagret. Enligt Connolly & Begg (2002) är OLTP-system och datalager byggda utefter olika syften och behov vilket innebär vissa skillnader dem emellan vilket illustreras nedan i figur 2. OLTP-system är ämnade att generera data som är aktuell och detaljerad för att stödja dag-till-dag beslut av ett stort antal användare. Data i OLTP-system är också optimal för transaktioner som är många, förutsägbara, repeterade och uppdaterings intensiva. Datalager däremot innehåller utöver aktuell och detaljerad data även historisk och summerad data vilket möjliggör analyser av trender. Ett datalager fungerar som beslutsstöd för företagsledningen och designen av datalager har gjorts för att stödja ett lågt antal oförutsägbara transaktioner. Dessa skillnader mellan OLTP-system och datalager medför vissa problem då OLTP-system utgör källsystem till datalagret och därför måste data från källsystemen genomgå olika processer innan data kan användas i datalagret (Connolly & Begg, 2002).

OLTP-system	Datalager
Innehåller aktuell data	Innehåller historisk data
Lagrar detaljerad data	Lagrar detaljerad samt summerad data
Innehåller dynamisk data	Innehåller statisk data
Hanterar en stor mängd av transaktioner	Hanterar en mindre mängd av transaktioner
Lämplig för ett förutsägbart och repeterat användande	Lämplig för ett oförutsägbart användande med ostrukturerade svar
Ämnad för transaktioner	Ämnad för analyser
Applikationsorienterad	Subjektorienterad
Stödjer dag-till-dag beslut	Stödjer strategiska beslut
Avsedd för användning av många användare	Avsedd för användning av få användare

Figur 2 Jämförelse mellan OLTP-system och datalager (Efter Connolly & Begg 2002, sidan 1049)

2.2.2 Extrahering

Extrahering är första steget i processen att få data till datalagret. Extrahering innebär läsning och förståelse av källsystemens data samt kopiering av den data som ska lagras i datalagret (Kimball & Ross, 2002). Att plocka ut data ur källsystemen är en stor del av arbetet runt ett datalager och enligt Kimball m.fl. (1998) läggs cirka 60% av arbetstiden på denna process. Den största utmaningen är oftast att avgöra vilken data som ska plockas ut och för att få data med kvalitet till datalagret måste insamlingsprocessen vara väldesignad. För att underlätta extraheringsprocessen ska de verktyg som används kunna hantera följande saker enligt Kimball m.fl. (1998).

Multipla källsystem. Multipla källsystem berör hantering av data från flera olika källsystem. Denna hantering är viktig eftersom data till datalager hämtas från olika system som kan ha olika plattformar och lagringssätt. D.v.s att extraheringsprocessen ska kunna hantera data oavsett vilket eller vilka källsystem data kommer ifrån.

Kodgenerering. Kodgenerering innebär att kod skapas som körs i källsystemen eller ger direktiv till källsystemen för att möjliggöra extrahering av data ur källsystemen. Med andra ord behövs exekverbar kod för att kunna plocka ut data ur källsystemen.

Multipla extraheringstyper. Multipla extraheringstyper handlar om vilken typ av data som extraheras och när den extraheras. Vilken extrahering som görs beroende på vilket syfte datalagret är tänkt att tjäna och utefter syftet görs olika varianter på extraheringar av data. Exempelvis plockas saldot på ett bankkonto ut endast i slutet av månaden medan en annan extraheringstyp kan vara att plocka ut de fält i databasen som uppdaterats sedan förra extraheringen gjordes.

Kopior. Kopior kan användas i ett datalager för kontinuerlig uppdatering och sedan finnas tillgängliga som stöd. Till exempel kan kopior över tabeller i en databas tas under dagen för att finnas till som eventuellt stöd på natten om det skulle finnas behov av tillgång till uppdaterade uppgifter.

Komprimering/dekomprimering. Komprimering kan vara en viktig egenskap när stora mängder data ska flyttas över avstånd då komprimering kan reducera tiden det tar att sända data. D.v.s. att det kan bli flaskhalsar om en stor mängd data ska sändas och att komprimerad data kan minska längden på sändningstiden.

När extraheringsprocessen är över har data som ska lagras i datalagret plockats ut ifrån källsystemen och en temporär fil producerats. Den temporära filen blir i sin tur indata i transformeringssteget (Kimball m.fl. 1998) som är nästa del av ett datalagers arkitektur och beskrivs i kapitlet nedan.

2.2.2 Transformering

Transformering sker efter att data extraherats vilket innebär att extraherad data anpassas till format som passar in i datalagret (Kimball m.fl., 1998). Transformeringsprocessen har som syfte att tvätta, förändra strukturen och reformera data från källsystemen för att verifiera och förbättra kvaliten på data (King, 2000). Syftet med transfomeringsprocessen har delats upp i olika aktiviteter som kallas tvättning, integrering och aggregering vilket förklaras nedan.

Tvättning av data innebär en kontroll av data från källsystemen för att försäkra ett konsistent format och användning av data som ska laddas in i datalagret (Bischoff & Alexander, 1997). En tvättningsprocess lägger koncentrationen på vilket innehåll data har och inte enbart vilket lagringsformat data har när tvättningsprocessen undersöker om källsystemen lämnar giltiga värden (King, 2000). Kimball m.fl. (1998) hävdar att tvättningsprocessen är ett stort problem i skötseln av datalager och Agosta (2000) påpekar att det är viktigt att beaka datakvalitetsaspekter i detta skede.

Integrering är processen som mappar data från olika källsystem till att passa in i datalagret (Bischoff & Alexander, 1997). Som nämnts tidigare i rapporten handlar integrering om att sammanföra data från källsystem som använder olika beteckningar för samma data och ge data en entydig utformning (King, 2000). Det är en stor utmaning att kombinera data från flera olika källsystem till ett datalager och samtidigt hålla hög datakvalitet (Bischoff & Alexander, 1997).

Aggregering är en process som skär ner volymen på mängden data som hämtas från källsystemen genom att summera eller gruppera data till datalagret (Bischoff & Alexander, 1997). Anledningen till att data ofta aggregeras i datalager beror på att det möjligör snabba analyser om data är samlad på ett visst sätt som veckovis, månadsvis eller geografiskt. Till exempel kan försäljningssiffrorna för en viss region summeras för att bli mer lätthanterlig (King, 2000). Aggregering kan även ha en positiv effekt på datalagrets prestanda genom att svarstiden på frågor som ställs mot datalagret snabbas upp (Kimball m.fl., 1998).

Transformeringsprocessen har i uppgift att anpassa data till datalagret för att kunna presentera data på ett kontinuerligt sätt och vara lätt att hitta. Anpassning av data möjliggör ett effektivt användande av datalagret när en hög datakvalitet hålls och därför föreslår Agosta (2000) att det alltid ska finnas en tanke på datakvalitets förbättringar när data transformeras.

2.2.3 Laddning

Efter transformeringsprocessen kan data laddas in i datalagret (Kimball m.fl., 1998) och det är nu datalagret fylls med data. Laddning av ett datalager sker ofta med ett laddningsprogram en s.k bulkladdare (Kimball & Ross, 2002) vilket möjliggör en snabb laddning av datalagret (Kimball m.fl., 1998).

2.2.4 Analysapplikation

Arkitekturens främre del består av analysapplikationer och är den del av datalagret som användarna kommer i kontakt med. Användarna ställer krav på datalagrets applikationer som förväntas kunna förse användarna med den information användarna vill ha (Kimball m.fl., 1998). För att kunna erbjuda användarna den önskade information använder datalagrets främre del olika sätt att presentera och visualisera data för användarna (Agosta, 2000). Två tekniker som används i den främre delen för att göra data lätthanterlig för användarna är till exempel OLAP (eng. Online Analytic Processing) och datautvinning (eng. Data Mining). En kortfattad förklaring över begreppen följer nedan.

OLAP är en flerdimensionell analys där data från datalager presenteras i form av en kub för användarna. Med funktioner som kan minska detaljrikedomen (eng. rollup) och ökar detaljrikedomen (eng. drilldown) på data som presenteras (Connolly & Begg, 2001).

Datautvinning är en process som analyserar stora mängder av data i datalagret. Processen går ut på att söka efter tidigare oupptäckta mönster i ett företags affärsprocesser (Inmon, 2002).

Analysapplikationer tillhör den främre delen av arkitekturen och berörs inte i samma utsträckning som den bakre delen beroende på att rapportens frågeställning inte är fokuserad på den främre delen. En kort beskrivning av den främre delen har getts för att göra bilden av datalagrets arkitektur komplett.

2.3 Vad är datakvalitet?

En stor aktivitet i ett datalager handlar om att fylla datalagret med data från källsystemen. I detta skede extraheras och transformeras data från källsystemen innan den laddas in i datalagret. Data i ett datalager bör hålla en hög kvalitet och enligt Standardiseringskommissionen i Sverige (1992, sidan 167) definierar den internationella standarden ISO 9000 kvalitet som:

> *”alla sammantagna egenskaper hos en produkt som ger dess förmåga att tillfredsställa uttalade eller underförstådda behov”*

Med produkt avser standarden ”resultat av en process” (Sandholm, 2001, sidan 11) och i ett datalager är rapporter exempel på resultat från processer. King (2000) sammanfattar betydelsen av kvalitet med att data i ett datalager ska vara sanningsenlig och inte innehålla felaktiga uppgifter. Innehåller datalagret data som är otillförlitlig kan det leda till att användarna tappar förtroendet för datalagret och istället undviker att använda det. Att en kontroll över datakvalitet hålls i datalager är extra viktigt jämfört med andra system eftersom data till datalagret hämtas från flera olika system. Mängden källsystem medför en större risk, datakvalitetsmässigt, då källsystemen ska sammanföras och används som beslutsunderlag. Vikten av datakvalitet grundas alltså i att kvaliteten och tillgängligheten på data i ett datalager är avgörande för användarnas möjlighet att fatta beslut (Bischoff & Alexander, 1997). Vilken kvalitetsgrad som behövs i en verksamhet beror på vilka behov det finns av noggrannhet på data. Hög kvalitet på data ger användarna en större chans att fatta ett bättre beslut än om användaren blir presenterad data av låg kvalitet. Bischoff och Alexander (1997) anser att låg datakvalitet endast bör tillåtas i ett datalager om dess användare är medvetna om de brister som finns och kan hantera de situationer som kan uppstå. Ett företag kan fungera även utan att känna till viss data istället för att ha tillgång till felaktig data, speciellt om användarna inte känner till att data som presenteras är felaktig (Bischoff & Alexander, 1997).

Beskrivning av datakvalitet har gjorts på många olika sätt och innefattar olika fokus men samtliga innehåller ändå likartade indikatorer. Bischoff och Alexander (1997) tar upp många faktorer för att beskriva datakvalitet och i denna rapport har ett urval av faktorerna gjorts. Detta urval har baserats på likheter mellan Bischoff och Alexanders (1997) faktorer och datakvalitetsfaktorer hos andra författare. De faktorer som är mest relevanta för rapportens frågeställning har därefter valts ut som underlag för att beskriva datakvalitet i detta arbete och presenteras nedan.

1. *Korrekt* och noggrann data är data som överensstämmer med verkligheten (King, 2000). Till exempel att rätt information lagras om kunder. Fel information om kunderna kan bidra till misslyckade försök att ge kunderna erbjudanden att köpa produkter, då dessa inte befinner sig inom kundens intresseområden (Bischoff & Alexander, 1997). Även Eckerson (2002) och Kimball m.fl. (1998) antyder att det är viktigt att bejaka frågor som ifrågasätter om data som ska laddas in i datalagret är korrekt för att hålla hög datakvalitet.

2. *Datatypsenlighet* innebär att data lagras enligt den angivna datatyp som det aktuella fältet har tilldelats (Bischoff & Alexander, 1997). Till exempel lagras inte bokstäver i ett fält som är avsett för siffror.
3. *Konsistens* är att data har en form och innehåll som är konsistent, vilket tillåter att data integreras och delas av olika användare i olika applikationer och plattformar (Bischoff & Alexander, 1997). Ett konsistent lagringssätt hålls genom att standarder över vilket format olika data lagras på etableras för att befria datalagret från motsägande data (Kimball m.fl., 1998; King, 2000).
4. *Lämplighet* innefattar att data är lämplig för en viss tidpunkt, exempelvis månadsvis (Bischoff & Alexander, 1997). Det som avgör vad som är lämpligt är vad data används till av användarna i verksamheten (Kimball m.fl., 1998). Viktigt är dock att se till att den data som behövs vid en viss tidpunkt finns tillgänglig (Eckerson, 2002).
5. *Integration* handlar om att information som kommer ifrån olika platser går att kombinera med varandra (Bischoff & Alexander, 1997). Begreppet har tidigare beskrivits i kapitel 2.1.
6. *Kompletthet* innebär att all information angående ett visst ärende finns tillgänglig (Bischoff & Alexander, 1997). När datauppgifter är fullständiga blir data användbar och värdefull för verksamheten (King, 2000).
7. *Inga dubbletter* innebär att samma data från samma tidpunkt inte lagras mer än en gång i datalagret (Bischoff & Alexander, 1997).

Som nämnts tidigare i detta kapitel behöver inte data vara helt felfri och uppfylla alla datakvalitetspunkter utan behöver endast hålla den noggrannhet som användarna eller applikationerna kräver (Eckerson, 2002). Vad som krävs och förväntas av datakvalitet varierar inom olika användningsområden och användare (Vassilisdis, Bouzeghoub & Quix, 2000). Därför bör kännas till att olika användare och applikationer har olika krav på nivån av datakvalitet och det finns olika verktyg att använda för att reglera den önskade datakvalitetsnivån.

2.4 Hantering av datakvalitet

Teknologin spelar en viktig roll i utvecklingen mot god datakvalitet och kan hjälpa företag att upptäcka bristande datakvalitet i företagets data. Datakvalitetsverktyg för datalager används för att undersöka, reparera samt omarbeta data vid dess införande i ett datalager (Kachur, 2000). Enligt King (2000) ska datakvalitetsverktygen göra en omfattande granskning av data för att ge detaljerad information om innehåll, struktur och affärsregler. De ska också garantera dataintegritet genom att slå samman och rensa ut redundant data samt omarbeta motstridande data. Datakvalitetsverktyg kan användas på ett företags data för att ledningen lättare ska förstå om det finns problem med datakvalitet i företagets data (Bischoff & Alexander, 1997). Om låg datakvalitet upptäcks med hjälp av datakvalitetsverktyg kan problem rättas till innan data lagras i datalagret och på det sättet reducera kostnader för låg datakvalitet i företaget. Kachur (2000) hävdar att verktyg inom datakvalitetsområdet ständigt ökar i antal och enligt Eckerson (2002) består marknaden för närvarande av ett dussintal verktyg sedan deras debut på 1990-talet. Huvudsakligen har datakvalitetsverktygen tidigare fokuserat på att rensa upp namn och adressdata. Orsaken till denna fokusering beror på att detta

område innehåller mest felaktigheter eftersom det ofta sker förändringar som exempelvis dödsfall, skilsmässa eller flytt som påverkar data som finns lagrad. Fokuseringen börjar dock förändras och datakvalitetsverktyg kan exempelvis innehålla följande funktioner idag:

Analysering används för att identifiera sammansatta strängar som bör vara uppdelade för att passa in i datalagret (Forino, 2001). Datakvalitetsverktygen lokaliserar dataelementen och separerar dem till unika fält d.v.s. om ett fält består av orden John Doe, 18 år delas fältet exempelvis upp i tre fält som är ett förnamnsfält, ett efternamnsfält och ett åldersfält (Eckerson, 2002).

Standardisering behövs när data från källsystemen som matas in skiljer i utformning men betydelsen är densamma. Datakvalitetsverktygen ändrar data till ett konsistent format som används i datalagret (Forino, 2001). Till exempel kan utformningen av adressen St. hötorget 2 från ett källsystem ändras till Stora hötorget 2 för att passa in i datalagrets sätt att lagra adresser.

Validering görs för att verifiera om data värden är giltiga (Forino, 2001).

Matchning drar paralleller med hjälp av data och identifierar olika sorts data som är relaterat till samma person eller företag (Eckerson, 2002). D.v.s. att data från olika källor kan kombineras som innefattar data om samma person eller företag.

Konsolidering liknar matchning då paralleller av data dras. Konsolidering kan dessutom kombinera data för att identifiera länkar mellan data, till exempel personer som ingår i samma hushåll eller barn med samma förälder (Eckerson, 2002).

Antalet och typen av funktioner varierar mellan de olika datakvalitetsverktygen vilket innebär att det inte finns ett enskilt verktyg som klarar av allt (Williams, 1997). Däremot har ofta säljare av datakvalitsverktyg ett antal olika verktyg som kan kombineras med varandra för att hantera de datakvalitetsproblem som kan finnas i ett företag. Exempel på företag som säljer datakvalitetsverktyg är Firstlogic, Trillium Software, Group-1 Software, Vality Technology och Innovative Systems (Forino, 2001). Även QDB Solutions Inc., WizSoft Inc. och Unitech Systems Inc är exempel på säljare av datakvalitetsverktyg (Williams, 1997).

Enligt Eckerson (2002) gör företag satsningar inom olika områden som kommer att leda till behov av verktyg med allt bredare funktionalitet och som kan hantera mera komplexa problem. Bl.a. kommer den ökande marknaden på Internet att vidga marknaden för datakvalitetslösningar (Forino, 2001). I och med att datakvalitetsverktyg ökar i antal behöver företag veta om att datakvalitetsverktyg funktionsmässigt varierar och att en kombination av olika verktyg kan vara nödvändig. Företag bör därför vara medvetna om vad de olika datakvalitetsverktygen har att erbjuda för att kunna välja passande verktyg (Williams, 1997).

3 Problemområde

Företag idag lider av problem rörande låg datakvalitet i datalagren med anledning av att det finns enormt stora mängder data i företagen. Att flytta data från källsystem till datalager är en central aktivitet i ett datalager och samtidigt det mest utmanande, tidskrävande och dyraste steget (King, 2002). Hantering av data när det gäller överföring från källsystem till datalager sker med hjälp av olika verktyg som stöd. Det finns även verktyg som kontrollerar datakvalitet på data som förs in i datalagret. Anledningen till att datakvalitetsverktyg finns grundas i ett behov av att data i datalager är av hög kvalitet för att företagsledningen ska kunna fatta väl grundade beslut med hjälp av data som underlag. I detta kapitel ges exempel på några av de problem som idag finns inom datalager rörande datakvalitet. En beskrivning ges också av ett växande behov av datakvalitetsverktyg och varför utvärdering av dessa är viktig.

Det finns många problem som är kopplade till låg datakvalitet. Enligt Kachur (2000) leder låg datakvalitet i systemen till felaktigheter, konflikter och missad information som i sin tur leder till enorma kostnader för organisationen. Svårigheter med datakvalitet är att den ofta har låg prioritet i organisationer av kostnadsmässiga orsaker. I själva verket kan kostnaderna bli enormt höga av låg datakvalitet. Några exempel på kostnader orsakade av låg datakvalitet, enligt Bischoff och Alexander (1997), är höga kostnader för hårdvara, mjukvara och programmerare/analytiker vilket är resultatet av redundant data. Mycket tid läggs på att rätta till felaktig data. Dåliga beslut fattas som är kopplade till låg datakvalitet och möjligheter går till spillo på grund av att önskad data inte fanns tillgänglig eller inte var tillförlitlig när beslut skulle fattas. Låg datakvalitet kan göra att rapporter i företagen får ett inkonsekvent innehåll och det måste läggas arbete på att lösa detta, vilket är både tidskrävande och frustrerande. Dessa bekymmer syns tydligast hos användarna genom att deras förtroende och förståelse för underliggande data minskar (Bischoff & Alexander, 1997).

Många företag tvekar vid investeringen av datakvalitetsverktyg p.g.a. det höga priset på verktygen. Eckerson (2002) tror att investeringen i datakvalitetsverktyg kommer att förändras då företagsledningar sakta börjar få upp ögonen för vikten av datakvalitet i datalagren. Tidigare har datakvalitetsverktyg fokuserat på att tvätta namn och adressdata men detta börjar förändras då verktygen är under ständig utveckling. Idag hanterar datakvalitetsverktygen fler problem rörande datakvalitet än tidigare med funktioner såsom analysering, standardisering, validering, matchning och konsolidering. Det är viktigt att företagsledningar uppmärksammar och tar del av denna utveckling av datakvalitetsverktygen för att kunna ge företaget en bra datakvalitetslösning. Utvärdering av datakvalitetsverktyg är en viktig aktivitet för att få en försäkran om att verktygen är lämpade för de hävdade ändamålen. Datakvalitetsverktygens funktioner behöver utvärderas för att ge företagsledningar en bild över vad verktygen kan prestera och underlätta sökandet efter datakvalitetsverktyg som är lämpade för det aktuella företagets datakvalitetsproblem.

3.1 Problemprecisering

Detta arbete syftar till att utvärdera datakvalitetsverktyg vid överföringen av data från källsystem till datalager. Datakvalitet i detta arbete har tidigare beskrivits genom ett antal utvalda faktorer från Bischoff och Alexander (1997) och med dessa som utgångspunkt är frågan:

Hur väl hanteras datakvalitetsfaktorerna: korrekthet, datatypsenlighet, konsistens, integration, kompletthet, lämplighet och inga dubbletter av marknadens mest förekommande datakvalitetsverktyg ?

3.3 Avgränsning

De datakvalitetsverktyg som valts att titta närmare på i detta arbete har begränsats till de datakvalitetsverktyg som för tillfället anses vara vanligast på marknaden. Bedömningen av vilka datakvalitetsverktyg som är vanligast på marknaden har gjorts genom att titta på vilka verktyg som ofta återkommer i artiklar från olika verksamheter. Motiveringen till detta urval är att de datakvalitetsverktyg som är vanligast på marknaden också är de datakvalitetsverktyg som flest företag använder idag. Intressant är att se om, på vilket sätt samt i vilken omfattning de mest förekommande datakvalitetsverktygen hanterar datakvalitetsfaktorer.

3.4 Förväntat resultat

Det förväntade resultatet är att med hänsyn till de framtagna datakvalitetsfaktorerna kunna ge svar på om datakvalitetsverktygen hanterar och kan vara fördelaktiga att använda för att hantera datakvalitetsproblem i företag. Resultatet med rapporten förväntas med andra ord att kunna ge upplysning och vägledning i vilka åtgärder datakvalitetsverktyg har på datakvalitetsproblem.

4 Metod

Utifrån arbetets frågeställning behövs ett beslut fattas rörande hur information ska samlas in för att få kunskap om det valda ämnesområdet. Dawson (2000) säger att det är nödvändigt att identifiera och planera vilket arbete som ska utföras för att nå arbetets syfte och mål. Det finns många olika metoder att samla in information på och valet av metod eller metoder att samla information beror på vad som är lämpligast för att få den aktuella frågeställningen besvarad (Patel & Davidson, 1994). För att kunna avgöra vilken metod som är lämpligast att använda är det viktigt att känna till metodernas för- och nackdelar. Kapitlet ger en beskrivning och utvärdering av de metoder som är relevanta att använda i detta arbete med hänsyn till den aktuella frågeställningen. Därefter ges en motivering till det val av metod som gjorts i arbetet och slutligen presenteras ett tänkt tillvägagångssätt.

4.1 Litteraturstudier

En litteraturstudie är en tänkbar metod i detta arbete och innebär att information hämtas från olika dokument (Patel & Davidson, 1994). För att hitta den litteratur som bör ingå i litteraturstudien måste en sökning genomföras. Litteratursökningen innefattar ett systematiskt samlande av publicerad information som är relaterat till ämnet (Dawson, 2000). Litteratur finns presenterat i ett antal olika format och exempel på format som information kan presenteras i är böcker, artiklar, konferensskrifter, manualer, företagsrapporter och dokumentation.

När ett förlag står bakom publiceringen kan informationen anses ha en hög trovärdighet eftersom de källor författaren använt kontrolleras. Böcker skulle då kunna ge en god uppfattning över detta arbetets ämnesområde. Om böcker ska används som informationskälla i detta arbete bör dock en medvetenhet finnas att publicering av böcker tar tid och därför inte innehåller det absolut senaste tänkandet inom området.

En färskare informationskälla för detta arbete är tidskriftsartiklar och konferensskrifter som innehåller det senaste tänkandet inom olika forskningsområden. Artiklar från vetenskapliga tidskrifter är även granskade av personer som är kunniga inom området och kan därför anses vara en mycket tillförlitlig informationskälla. En nackdel med vetenskapliga tidskriftsartiklar är dock att språkbruket kan vara svårtolkat för gemene man (Patel & Davidson, 1994). Konferensskrifter presenterats på nationella och internationella konferenser och kan ibland innehålla mer uppdaterad information än artiklar från tidskrifter (Dawson, 2000), vilket vore intressant för detta arbete. Enligt Dawson (2000) kan dock konferensartiklar variera i kvalitet och en kontroll över vilken deltagarkategori konferensen haft är därför alltid berättigad för att säkra tillförlitligheten på uppgifterna.

En annan aktuell informationskälla är manualer som kan förse detta arbete med uppgifter om datakvalitetsverktygens funktioner. Nackdelen med manualer är att de kan vara svåra att komma över eller kosta pengar. Information om datakvalitetsverktygens funktioner är intressant för detta arbete och kan även finnas i

rapporter som är skrivna av företag som säljer datakvalitetsverktyg. Tillförlitligheten i företagsrapporter kan dock ifrågasättas eftersom rapporter skrivna av det egna företaget gärna favoriserar dess egna produkter. Sammanfattningsvis om litteraturstudier kan sägas att oavsett varifrån informationen hämtas är det viktigt att veta att informationen kan vara vinklad ur ett visst perspektiv beroende på författarens avsikt med skriften (Patel & Davidson, 1994).

4.2 Intervjuer

Intervju är en metod som bygger på att samla information med hjälp av frågor och är en annan tänkbar metod i detta arbete. Patel och Davidson (1994) menar att intervjuer är möten där intervjuaren antingen personligen eller via ett telefonsamtal kommunicerar med intervjupersonen och genomför intervjun. Intervjuerna kan se olika ut med avseende på uppbyggnaden av frågorna, exempelvis kan frågorna komponeras i förväg eller under själva intervjuns gång beroende på de svar intervjupersonen ger. Intervjuer bör i första hand användas när det inte föreligger några klara svarsalternativ, utan det finns många möjliga sätt att svara på, och att ämnet gör det lämpligt att följa upp svaren med ytterligare frågor (Andersen & Schwencke, 1998). Metoden kan vara lämplig i detta arbete genom att intervjua personer som är kunniga och har erfarenhet av datakvalitetsverktyg och ge en djupare förståelse då intervjuer ger utrymme för diskussioner.

Personer med sådan kunskap kan ge information och reflektion över hur det är i praktiken att använda datakvalitetsverktyg i en verksamhet. Dessa personer kan även ge en inblick i hur användare uppfattar olika datakvalitetsverktyg och dess funktioner. Problem med intervju är dock att finna personer med rätt kunskap speciellt i detta arbete som kräver information inom ett begränsat område om specifika verktyg. Skulle personer med lämplig kunskap om det valda området i detta arbete finnas är ett annat problem att de ska ha möjlighet att avsätta tid för intervjun. Ännu ett problem är att intervju är en tidskrävande metod att använda och kräver noggranna förberedelser och bearbetning av frågor och svar (Patel & Davison, 1994).

4.3 Val av metod

I detta kapitel ges ett förtydligande över vilken metod som valts samt argumentation för valet. Dawson (2000) nämner ett antal kriterier att ha i åtanke vid val av metod och till dessa kriterier hör resurser, tid och pengar. Utifrån kriterierna och arbetets frågeställning kan ett beslut fattas angående val av metod. Av tidigare nämnda metoder har för detta arbete litteraturstudier valts. Ur arbetets synvinkel är litteraturstudier den mest lämpade metoden då grundkunskap samt aktuell och detaljerad information till detta arbete kan hämtas från böcker, artiklar, konferensskrifter och företagsrapporter. Som nämnts tidigare är böcker en bra start på litteraturstudien för att få en överblick över ämnesområdet för detta arbete. Artiklar och konferensskrifter är sätt att bygga på den baskunskap som inskaffats från böcker då artiklar och konferensskrifter innehåller aktuellare information, vilket nämnts i kapitel 4.1. Företagsrapporter och dokumentation har även diskuterats i kapitel 4.1 och för detta arbete kan uppgifter från företag vara ett bra komplement till artiklar som berör datakvalitetsverktyg. Dock med en medvetenhet att inte enbart basera på

företagens uppgifter då dessa kan vara partiska. Manualer har däremot inte valts som underlag i litteraturstudien för detta arbete beroende på att de är svåråtkomliga eller kostsamma. Litteraturstudier känns som ett naturligt tillvägagångssätt för att få förståelse samt inhämta kunskap om ämnesområdet då rapportens frågeställning kräver många faktauppgifter för att kunna besvaras, vilket finns dokumenterat i någon sorts litteratur. Intervju valdes bort beroende på de problem som medföljer metoden. Som tidigare nämnts i kapitel 4.2 är det svårt att hitta personer som besitter kunskaper som är önskvärda i detta arbete samt att personerna har tid med att bli intervjuade. Intervjuer är också som tidigare nämnts väldigt tidskrävande och kan försvåra detta arbete då det är begränsat till en viss tidsperiod. Dessa problem runt metoden bidrog till att intervju inte valdes som informationsinsamlare i detta arbete.

4.4 Tillvägagångssätt

Avslutningsvis följer här ett tänkt tillvägagångssätt i arbetet som är uppdelat i olika etapper. Första etappen är litteraturstudier där böcker ska ge en grundkunskap om det valda ämnesområdet i detta arbete. När en grunduppfattning finns övergår sökandet mer och mer till artiklar och konferensskrifter som ofta innehåller mer detaljerad information om det senaste tänkandet inom olika forskningsområden. Med en baskunskap om ämnesområdet från böcker ska det även underlätta att finna de artiklar eller delar av artiklar som är intressanta för arbetet. I den andra etappen ska först ett urval göras av vilka datakvalitetsverktyg som detta arbete ska titta närmare på. Därefter ska företagsrapporter och dokumentation användas som ett komplement till artiklar för att underlätta förståelsen och insamlingen av information om olika datakvalitetsverktyg och deras olika funktioner. Tredje etappen är att sammanställa informationen om datakvalitetsverktygen som funnits i litteraturen och bedöma datakvalitetsverktygens funktioner med hjälp av de datakvalitetsfaktorer som satts upp i detta arbete.

5 Genomförande

I detta kapitel ges en beskrivning över hur den valda metoden i kapitel 4 tillämpats i detta arbete.

5.1 Litteraturstudier

Genom litteraturstudierna byggdes en baskunskap upp i böcker om datakvalitet i datalager, vilka faktorer som indikerar hög datakvalitet och vad datakvalitetsverktyg är. Nästa steg togs i sökning av artiklar på Internet där en djupare förståelse av datakvalitet i datalager byggdes upp. Artiklarna bidrog också till en mer ingående beskrivning av datakvalitetsverktyg, vilka funktioner de har samt vilka företag det finns på markanden som säljer datakvalitetsverktyg. Därefter gjordes ett urval av vilka företag som säljer datakvalitetsverktyg som skulle ingå i detta arbete, vilket beskrivs närmare i kapitel 5.2. Efter urvalet hämtades information från företagens webbplatser om vilka datakvalitetslösningar företagen hade till datalager samt dessa datakvalitetsverktygs olika funktioner. Uppgifterna från företagen som säljer datakvalitetslösningar användes som jämförelsematerial till artiklar och företagsrecensioner som beskriver olika datakvalitetsverktyg. De uppgifter som återfanns i samtliga informationskällor om ett verktyg vägdes samman för att beskriva ett datakvalitetsverktygs funktionalitet.

5.2 Val av datakvalitetsverktyg

När klarhet om vilka företag det finns som säljer datakvalitetsverktyg för datalager funnits gjordes ett urval av vilka företag som skulle ingå i undersökningen i detta arbete. De företag som valdes var de företag som påträffades i flest artiklar och återkom i högst frekvens vid sökning efter datakvalitetslösningar i kombination med datalager på Internet. Ett antagande som ansågs rimligt var att dessa företags datakvalitetslösningar används i flera verksamheters datalager och är kända namn inom datakvalitetsområdet. Utvalda företag i detta arbete är Ascential Software, Trillium Software, Innovative Systems och Firstlogic.

5.3 Bedömning

Efter att insamling av information om olika datakvalitetslösningar gjorts sammanställdes uppgifterna. Därefter utgjorde sammanställningen av lösningarna grunden för författarens egna bedömning av hur datakvalitetsfaktorerna korrekthet, datatypsenlighet, konsistens, lämplighet, integration, kompletthet och inga dubbletter hanteras av datakvalitetsverktygen. Dessa datakvalitetsfaktorer sattes upp i en egen bedömningsmall över samtliga datakvalitetslösningar och illustrerar vilka datakvalitetsfaktorer var och en av lösningarna hanterar. Mallen illustrerar även på vilket sätt de hanterar datakvalitetsfaktorerna utefter en bedömningsskala som graderades med siffrorna 1-4. Siffran ett representerar att datakvalitetsfaktorn inte hanteras av datakvalitetslösningen över huvudtaget. Siffran två representerar att datakvalitetsfaktorn hanteras av datakvalitetslösningen och en övergripande beskrivning av behandlingen nämns. Siffran tre representerar en utförlig hantering i beskrivningen av hur datakvalitetsfaktorn behandlas. Siffran fyra representerar en

mycket utförlig hantering av hur datakvalitetsfaktorn behandlas i datakvalitetslösningen.

6 Materialpresentation

Kapitlet beskriver de datakvalitetsverktyg som ingår i de datakvalitetslösningar som valts till detta arbete. Beskrivning ges över om och på vilket sätt de olika lösningarna hanterar de datakvalitetsfaktorer som tidigare beskrivits i kapitel 2.3. Slutligen presenteras en sammanfattning över samtliga datakvalitetslösningar i detta arbete.

6.1 Datakvalitetslösningar

Som nämnts tidigare i arbetet finns det inte ett enda verktyg som hanterar allting som berör datakvalitet i ett datalager. Däremot finns det flera olika verktyg som kan kombineras och hantera de problem som företaget har. De företag som säljer datakvalitetsverktyg presenterar flera verktyg som kan användas ihop med varandra och säljer paketlösningar till företag som behöver få ordning på datakvalitetsproblem i datalager. Dessa paket kan bestå av olika antal verktyg beroende på vilka bekymmer det aktuella företaget har. Exempel på stora företag inom datakvalitetsverktygsmarknaden som valts i detta arbete presenteras i kapitlen nedan tillsammans med deras datakvalitetslösningar för datalager. Varje kapitel nedan avslutas med en beskrivning av hur varje datakvalitetslösning behandlar datakvalitetsfaktorerna korrekthet, datatypsenlighet, konsistens, lämplighet, integration, kompletthet och inga dubbletter.

6.1.1 Ascential Software

Ascential Softwares datakvalitetslösning ”Integrity” hjälper företag att förbättra och bibehålla riktigheten på företagets data (Williams,1997). Verktyget har inbyggda procedurer som undersöker, standardiserar och matchar data från olika källsystem innan data matas in i ett datalager (Uhl, 2002). Arbetet mot datakvalitet sker efter en fyrfas-metodologi (Williams, 1997).

Första fasen är undersökning och här identifieras data som kommer från källsystemen. Undersökningsproceduren analyserar källsystemens data och producerar en rapport som identifierar innehållet i källsystemens olika datafält (Reynolds, 2000). Datakvalitetslösningen delar upp datavärdena i mindre fragment för att känna igen fält i källsystemens data som innehåller likadan information men är organiserad på annat sätt. Fragmenten markeras med olika symboler av verktyget beroende på om det exempelvis är siffror eller bokstäver som utgör fragmenten. Med hjälp av fragmentens markering upptäcks mönster som kan användas för att i ett senare skede avgöra om olika datavärden matchar varandra (Raab, 2000). Avgörandet till matchningen grundas i regler som sätts upp för att passa det aktuella företagets behov av information. (Reynolds, 2000).

Andra fasen kallas standardisering och här används de mönster som kommit fram av undersökningsfasen. Verktyget visar vilka områden i företagets data som är kritiska och behöver standardiseras (Reynolds, 2000). I standardiseringsfasen flyttas ordningen på fragmenten automatiskt om för att likna varandra ännu mer och för att kunna presentera data på ett standardsätt. Ibland kan det vara svårt att avgöra om data

ska matchas eller ej och dessa värden ”flaggas” då av datakvalitetslösningen för en manuell granskning (Raab, 2000). Verktyget grundar standardiseringen på existerande data i datalagret för att se till att data presenteras på ett kontinuerligt vis (Uhl, 2002).

Tredje fasen fokuserar på matchning vilket ser till att dubbletter tas bort genom att sammanföra datavärden med matchande markeringsmönster till grupper. Matchningsprocessen tillåter alltså att dubbletter förs samman genom att jämföra olika mönster och gör en bedömning om data ska grupperas ihop eller ej (Ascentialsoftware.com, 2002).

Fjärde fasen är utformning vilket sker när matchningsprocessen är gjord och här avgörs vilket utformningsalternativ i en grupp som ska användas som standard i datalagret (Raab, 2000). Standardformatet som ska användas i datalagret kan antingen väljas från gruppen eller från datalagret om datalagret redan har ett format som ska användas (Uhl, 2002). Under denna fas tas också felstavningar och förkortningar om hand genom grupperingen (Raab, 2000).

Datakvalitetslösningens matchningsteknik ger användarna en kontroll över hur varje situation hanteras och gör det lättare för användarna att förstå resultatet (Raab, 2000). Datakvalitetslösningen möjliggör också att relaterad data kan länkas samman och att data kan analyseras och jämföras utefter existerande data i datalagret. Dock berikas inte data med geografiska och demografiska uppgifter eller andra externa källsystem (Reynolds, 2000). Datakvalitetslösningen klarar inte heller på egen hand av att hantera extrahering, transformering och laddningsprocesser (Uhl, 2002).

Nedan ges en översiktlig beskrivning över hur datakvalitetslösningen hanterar de datakvalitetsfaktorer som satts upp i detta arbete:

Korrekthet. Datakvalitetslösningen markerar datafragment med olika symboler och grupperar därefter data utefter de mönster som bildas. På det sättet tas felstavningar och förkortningar bort från data. Data delas upp i mindre delar vilket gör det lättare att hantera korrektheten på ett mycket utförligt sätt.

Datatypsenlighet. I undesökningsfasen identifieras innehållet i källsystemens olika datafält på ett utförligt sätt genom att data analyseras och en rapport över vad varje datafält innehåller levereras av datakvalitetsverktyget.

Konsistens. Datakvalitetslösningen visar kritiska områden i data som behöver standardiseras. Med hjälp av mönstermarkeringen hanteras datakvalitetsfaktorn utförligt genom att datafragment kombineras för att få fram ett enhälligt mönster och på det sättet presentera data på ett standardiserat sätt i datalagret. Sättet data standardiseras på grundas på existerande data i datalagret om sådan data finns.

Lämplighet. Datakvalitetslösningen beskriver övergripande att regler anpassas för vilken och hur data ska hanteras utefter företagets behov. Lösningen är därmed flexibel och passar in i olika verksamheter.

Integration. En övergripande beskrivning ges att datakvalitetslösningen kan jämföra och behandla data från olika källsystem och sammanföra dessa källsystem i ett datalager.

Kompletthet. Datakvalitetslösningen kan dra relationer mellan data och länka samman data från olika sammanhang med olika källsystem men som berör samma individ eller företag. Detta sätt kompletterar utförligt de datauppgifter som finns lagrade inom företaget men verktygen tillför inte någon berikning av data från externa källsystem.

Inga dubbletter. Här sammanförs data med liknande mönstermarkering i grupper. Gruppen ges sedan ett gemensamt namn som får representera datasamlingen i gruppen. Genom denna gruppering av data görs en tydlig och utförlig reducering av antalet dubbletter i data som ska laddas in i datalagret.

6.1.2 Trillium Software

Datakvalitetslösningen ”Trillium Software System” är ett paket med datakvalitetsverktyg som kan kundanpassas med regler som passar företagets informationsbehov och möjliggör omarbetning av data i förberedelse inför laddningen av data i datalagret (Schauer, 2000). Verktygspaketet är designat för att tillåta användarna att undersöka, identifiera, godkänna och matcha information från olika källsystem (Sanantoniobusinessjournal.com, 2002; Dmreview.com, 2002). Datakvalitetslösningen har också en stark internationell kapacitet genom att flera olika språk förstås (Hudson, 2003).

Dataundersökning och analyseringsverktyg (eng. ”Trillium Parser” och ”Trillium Software Data Analytics”) i datakvalitetslösningen identifierar och filtrerar data genom att förstå vilken datatyp de olika datafälten har och på så sätt förstå datafältets innehåll. Datakvalitetslösningen återger innehåll och betydelse för företags data, tecken för tecken och rad för rad, för att åstadkomma en kontinuerlig förståelse och syn på informationen i verksamheten (Hillman, 2000). Dataundersökning används för att analysera data och framhäva kritiska områden i data om sådana upptäcks för att anpassa regler som bidrar till standardisering och berikning av företagets data (Hudson, 2003). Verktyget som förstår olika språk förser företagen med en insyn i företagets data genom att avslöja olika dataförhållanden. Exempel på förhållanden som presenteras genom analyser är namn och adresser som är placerade i fel datafält. Andra exempel är frekvensen av händelser i ett fält som kan vara blankrader och nollor eller utformningen av data i ett fält som exempelvis xxx-xxx-xxx för telefonnummer. Analyserna avslöjar också distributionen av företagsadresser jämfört med bostadsadresser samt rader med multipla namn. Analyserna kan sedan användas till att sätta upp regler för hur data ska standardiseras i ett datalager. (Trilliumsoftware.com, 2003).

Standardisering och berikning av data görs när inspektionen är gjord. Standardiseringsverktygen ser till att namn och adressdata, affärsdata och internationell data presenteras på likadant sätt för att ge datalagret ett kontinuerligt utseende. Berikning sker för att verifiera och rätta till data med hjälp av att externa källor integreras för att återge korrekta uppgifter samt bidra till konsistens i datalagret. Berikningsverktyget (eng. ”Trillium Geocoder”) i datakvalitetslösningen är designad

att godkänna, verifiera och förbättra adressdata med hjälp av postkoder och adresskomponenter (Schauer, 2000). Lexikonliknande listor med adressrelaterad information finns i datakvalitetslösningen för att alltid återge kompletta och riktiga uppgifter i datalagret (Bright, 2001). Verktyget kan användas till att matcha all slags data samt även berika data med exempelvis geografisk och demografisk information för att göra datalagret till en komplettare informationskälla (Prillaman, 2001).

Länkning används för att matcha kunder, hushåll, företag m.m. utefter kriterier som det aktuella företaget sätter upp (Yee, 1999). Matchningsverktyget (eng. ”Trillium Matcher”) är ett verktyg som jämför data från olika källsystem för att hitta samband och likheter (Schauer, 2000). När identifiering av relationer mellan data görs kan data standardiseras och eliminera dubbletter som uppstår när data från olika källsystem ska föras samman i ett datalager (Bright, 2001).

Datakvalitetslösningen klarar som tidigare nämnts av att undersöka, identifiera, standardisera, berika och matcha data som ska laddas in i datalagret. En nackdel med datakvalitetslösningen är dock att inget verktyg har funktioner som kan extrahera och sprida data från källsystemen (Clare, 2000) .

Nedan ges en översiktlig beskrivning över hur datakvalitetslösningen hanterar de datakvalitetsfaktorer som satts upp i detta arbete:

Korrekthet. Datakvalitetslösningen förstår flera olika språk vilket ger verktygen en större möjlighet att förstå de flesta uppgifter som finns lagrade. Till hjälp används lexikonliknande listor med adressrelaterad information för att återge riktiga uppgifter men principen med att jämföra mot listor kan även tillämpas på fler områden än adresser. Genom att datakvalitetslösningen har förutsättningar att förstå olika språk samt lexikonliknande listor till hjälp kan datakvalitetsfaktorn behandlas mycket utförligt.

Datatypsenlighet. Genom analys av data från källsystemen identifieras vad varje datafält innehåller. Analysen identifierar och filtrerar data tecken för tecken och rad för rad och upptäcker då data som är placerad i fel datafält på ett utförligt sätt.

Konsistens. Datakvalitetslösningen hanterar datakvalitetsfaktorn på ett väldigt utförligt sätt då data hålls efter standarder genom matchning mot existerande data i datalagret, mot lexikonliknande listor samt via integrerade externa källsystem. Allt detta görs för att se till att data presenteras på likadant sätt i datalagret.

Lämplighet. Datakvalitetslösningen ger en övergripande beskrivning över hur lämplig data fås fram genom att tillåta kundanpassning av regler som uppfyller företagets behov.

Integration. En övergripande beskrivning ges att datakvalitetslösningen klarar av att hantera data från olika källsystem som ska laddas in i ett datalager.

Kompletthet. Data görs komplett på ett mycket omfattande sätt genom matchning mot lexikonliknande listor samt via berikning av data. Berikning görs för att verifiera och rätta till data med hjälp av att externa källsystem integreras med verktygen.

Exempelvis kompletteras postkoder i adresser eller data med geografisk och demografisk information för att göra datalagret till en komplettare informationskälla.

Inga dubbletter. Data från olika källsystem förs samman och samband hittas. Därefter sker en utförlig eliminering av dubbletter genom att länka data med hjälp av relationer mellan data med olika utformning men med samma innebörd.

6.1.3 Innovative Systems

Innovative Systems med datakvalitetslösningen ”i/Lytics” är ett datakvalitetspaket som ska omfatta analyser över källsystemens data och omarbetning av data (Schwalb, 2002). Datakvalitetslösningen identifierar datatyper, kön, relationer, felstavningar, variationer och felplacerade komponenter samt standardiserar data från olika källsystem. Processerna som verktygen utför är regelbaserade vilket innebär att regler sätts upp som är anpassade efter företaget som använder verktygen. Att regler kan kundanpassas medför att verktygen är flexibla och kan passa in i olika typer av verksamheter (Innovativesystems.com, 2003).

Dataprofileringsverktyget (eng.”Data Profiling”) i datakvalitetslösningen ”i/Lytics” ser till att data från källsystemen är korrekt och identifierar automatiskt om det finns felaktigheter i innehållet, strukturen eller relationerna (Database trends and applications, 2001).

I datakvalitetslösningen ingår även ett datakvalitetsverktyg (eng.”Data Quality”) som kan användas för att automatiskt lokalisera och rätta till namn och adressformat, felaktigheter, felstavningar och andra oregelbundna förekomster i data. Verktyget kan också analysera, standardisera och förbättra information som e-postadresser, produkt nummer m.m. med hjälp av kunskapsbasen (Knowledgestorm.com, 2003a). Verktyget levererar en rapport som visar den data som är felaktig samt vad som är felaktigt för att underlätta en fokuserad och tidsbesparande granskning av data (Espiner, 2002).

Datakvalitetslösningens länkningsverktyg (eng. ”Data Linking”) identifierar dubbletter i data genom en förmåga att länka samman kunder som delar samma namn, adress, bankkonto eller annan information (Database trends and applications, 2001). Verktyget levererar en rapport över matchningar som gjorts för att förklara vilken data som grupperats ihop som dubbletter samt varför de betraktas som dubbletter (Espiner, 2002).

Berikning av data som ska laddas in i datalagret löser datakvalitetslösningen med berikningsverktyg (eng. ”Geocoding”) som tillför geografisk och demografisk information. I datakvalitetslösningen finns även ett standardiseringsverktyg (eng. ” CASS correction”) som har standarder över post adresser i USA (Innovativesystems.com, 2003)

Verktygens möjligheter att automatiskt identifiera och rätta till data baseras på en heuristisk kunskapsbas som består av över 3 miljoner namn och adressrelaterade ord

och fraser (Knowledgestorm.com, 2003a). Att den är heuristisk innebär att den har en inbyggd intelligens och är självlärande vilket innebär att kunskapsbasen blir större och effektivare med tiden av den data som passerar verktyget. Allt kan dock inte rättas till automatiskt utan behöver en manuell bedömning. Innovative Systems har då automatiska granskningsfunktioner som guidar användarna till den plats som behöver granskas, vilket underlättar granskningsarbetet tidsmässigt (Innovativesystems.com, 2003).

Nedan ges en översiktlig beskrivning över hur datakvalitetslösningen hanterar de datakvalitetsfaktorer som satts upp i detta arbete:

Korrekthet. Datakvalitetsverktyget undersöker om det finns felaktigheter i innehållet, strukturen eller relationerna och rättar till dessa om felaktigheter hittas på ett mycket utförligt vis med hjälp av den heuristiska kunskapsbasen i datakvalitetslösningen. En rapport som beskriver var felaktig data hittats samt vad som var felaktigt kompletterar det utförliga tillvägagångssättet att behandla datakvalitetsfaktorn.

Datatypsenlighet. Datatyper och felplacerade datavärden identifieras genom en övergripande beskrivning av att data från källsystemen analyseras innan sammanföring till ett datalager sker.

Konsistens. Datakvalitetsfaktorn behandlas på ett mycket utförligt sätt med hjälp av kunskapsbasen som finns i datakvalitetslösningen och ser till att data presenteras på ett kontinuerligt vis i datalagret. Lösningen innefattar även ett verktyg med standarder över postadresser i USA men för övriga världens postadresser finns inte standarder om inte adresserna möjligtvis återfinns i kunskapsbasen.

Lämplighet. Datakvalitetslösningen beskriver övergripande att regelbaserade procedurer anpassar regler för företaget för att få fram data som är lämplig för företagets verksamhet.

Integration. En övergripande beskrivning av att datakvalitetslösningen kan sammanföra data från olika källsystem till ett system visar att datakvalitetsfaktorn behandlas.

Kompletthet. Datakvalitetslösningen hanterar kompletthet på ett utförligt sätt då datakvalitetsverktygen kan förbättra data som ska lagras i ett datalager. Dessutom kan data med gemensamma nämnare länkas samman och information kan kompletteras och berikas med bl.a. geografiska och demografiska uppgifter.

Inga dubbletter. Verktyget kan på ett noggrant sätt identifiera dubbletter genom att länka samman information som på något sätt hör ihop eller relaterar till varandra i grupper. En rapport levereras om vilken data som har matchats samt en förklaring till varför just utvald data har grupperats ihop och betraktats som dubbletter.

6.1.4 Firstlogic

Datakvalitetslösningen ”Firstlogic Information Quality” är ett paket som består av ett antal olika verktyg som bearbetar källsystemens data. Paketet kan kombineras med Firstlogics analysverktyg (eng. ”IQ insight”) som är Firstlogics första steg mot datakvalitet innan åtgärder på företagens data görs.

Analysverktyget (eng. ”IQ insight”) är ett verktyg för bedömning av data och undersöker data för att avgör om datakvalitetsbehov uppfylls (Firstlogic.com, 2003a). Verktyget analyserar och identifierar också de mest kritiska områdena för att företaget ska veta var förbättringar kan behöva göras (Knowledgestorm.com, 2003b).

När en datakvalitetsbedömning är gjord kan åtgärder mot datakvalitet genomföras. Datakvalitetslösningen sammanför data från olika källsystem genom identifiering, standardisering, omarbetning och matchning av data innan inladdning i datalager sker (Bova, 2001).

Datatvättning och dataförhöjningsverktyg identifierar data för att tvätta och standardisera data. Verktyget kan rätta och matcha adressdata för över 190 länder men verktyget klarar även att matcha all slags data och inte enbart adresser (Firstlogic.com, 2003b). Regler kan anpassas utefter företagets behov för all slags data (Dekle, 2002). Verktyget kan även analysera och para in data i passande datafält i datalagret utefter datatyp (Knowledgestorm.com, 2003c).

Ett verktyg i datakvalitetslösningen tillför en lösning som gör tillägg på data (eng. ”Data Appending Solution”) vilket ger en helhetssyn på företagets data. Verktyget sammanför och lägger till information om telefonnummer och e-postadresser för att alltid vara säker på att uppgifterna stämmer överens med verkligheten och har en mottagare (Williams, 2003).

Datakvalitetslösningen består även av ett verktyg som matchar och sammanför data för att identifierar multipla företeelser och eliminera dubbletter (Wright, 2002). Verktyget kan även identifiera personer som ingår i samma hushåll (Firstlogic.com, 2003c). Verktyget kombinerar och sammanför olika källsystem i ett datalager (Bova, 2001).

Datakvalitetslösningen som identifierar, standardiserar och rättar både inhemska och internationella adresser används för att förbättra matchningsprocessen och korrektheten på data (Knowledgstorm.com, 2003c). En nackdel med datakvalitetslösningen är dock att vissa geografiska variabler behöver passera verktyget flera gånger för att få önskat resultat (Dekle, 2002).

Nedan ges en översiktlig beskrivning över hur datakvalitetslösningen hanterar de datakvalitetsfaktorer som satts upp i detta arbete:

Korrekthet. Datakvalitetsfaktorn behandlas genom att data analyseras och kritiska områden som behöver förbättras identifieras. Verktyget kan även rätta till företagets inhemska adressdata samt internationell adressdata för att på ett mycket utförligt sätt förbättra korrektheten på data.

Datatypsenlighet. Datakvalitetslösningen gör en övergripande identifiering av vilken datatyp data har från källsystemen för att kunna para in data i passande datafält i datalagret.

Konsistens. När verktyget identifierat var förbättringar i data behövs rättas och matchas data från olika källsystem. Datakvalitetslösningen behandlar datakvalitetsfaktorn på ett utförligt sätt genom att hanterar bl.a. 190 länders sätt att presentera adresser för att kunna presentera data på ett standardiserat och kontinuerligt vis i datalagret.

Lämplighet. Datakvalitetslösningen ger en övergripande beskrivning av att regler för all slags data kan anpassas utefter företagets behov. På det sättet kan regler kundanpassas till vilken verksamhet som helst.

Integration. Datakvalitetslösningen beskriver övergripande att data från olika källsystem sammanförs innan data laddas in i ett datalager.

Kompletthet. Datakvalitetslösningen hanterar datakvalitetsfaktorn mycket utförligt genom att sammanföra och komplettera uppgifter som telefonnummer och e-postadresser samt att personer i samma hushåll kan identifieras.

Inga dubbletter. Datakvalitetsverktyget eliminerar utförligt dubbletter genom att matcha och sammanföra data. På detta sätt identifieras tydligt multipla företeelser som uppkommer när olika källsystem kombineras samman i datalagret.

6.2 Sammanfattning

Slutligen kan sammanfattas att samtliga datakvalitetslösningar berör de datakvalitetsfaktorer som valts att beakta i detta arbete. Skillnader mellan de olika lösningarna fanns dock i vilket tillvägagångssätt som valts att lösa hanteringen av vissa datakvalitetsfaktorer. Samtidigt återfanns vissa likheter i lösningarnas sätt att behandla datakvalitetsfaktorer. Nedan i figur 3 illustreras en sammanställning över samtliga företag med datakvalitetslösningar i detta arbete tillsammans med en egen bedömning utifrån litteraturstudierna, utefter tidigare nämnd bedömningsskala i kapitel 5.3, för behandling av varje datakvalitetsfaktor.

Korrekthet har beskrivits tydligt av alla datakvalitetslösningar i detta arbete. Trots olika tillvägagångssätt mellan lösningarna upplevs det att samtliga lösningarna oavsett tillvägagångssätt hanterar korrekthet mycket väl därav betygen 4.

Datatypsenlighet behandlas på likartat sätt av de olika datakvalitetslösningarna men det finns skillnader i beskrivningen av tillvägagångssätt. De lösningar som beskrivit hur analyserna går till har getts betyget 3 medan de lösningar som endast antyder att datatypsenlighet hanteras fått betyget 2.

Konsistens beskrivs också på olika sätt av datakvalitetslösningarna. Här är dock skillnaden att vissa lösningar beskriver ett mer utförligt och noggrant sätt att hantera konsistens på än andra. De lösningar som beskrivit ett så omfattande sätt att hantera konsistens på har tilldelats betyget 4. Övriga lösningar har beskrivit en noga hantering som tydligt visar att konsistens hanteras av lösningen men saknar något extra och har därför fått betyget 3.

Lämplighet har av samtliga datakvalitetslösningar beskrivits på samma övergripande sätt. Detta tyder på att lämplighet beaktas av samtliga lösningar men inget tydligt tillvägagångssätt beskrivs och därav betyget 2.

Integration behandlas enligt beskrivningarna över datakvalitetslösningarna av samtliga lösningar. Däremot redovisas inte hur lösningarna går tillväga för att sammanföra olika system. Det framgår dock att integration behandlas men beskrivning av hur väl detta görs saknas och därför sattes betyget 2.

Kompletthet genomförs av samtliga datakvalitetslösningar genom olika sätt att sammanföra data vilket utförligt beskrivs av respektive lösning. Alla lösningar utom en har dessutom beskrivit resurser som innefattar en ännu mer omfattande hantering av kompletthet. Dessa lösningar har därför fått betyget 4 medan lösningen som saknar dessa resurser fått betyget 3.

Inga dubbletter hanteras på olika sätt av datakvalitetslösningarna. Ingen urskiljning av att det ena sättet är bättre än det andra kan dock göras utan alla lösningar ger en bra beskrivning över hur inga dubbletter uppfylls. Betyget 3 har därför tilldelats samtliga lösningar.

Företag / Datakvalitetsfaktor	Ascential Software	Trillium Software	Innovative Systems	Firstlogic
Korrekthet	4	4	4	4
Datatypsenlighet	3	3	2	2
Konsistens	3	4	4	3
Lämplighet	2	2	2	2
Integration	2	2	2	2
Kompletthet	3	4	4	4
Inga dubbletter	3	3	3	3
Summa	20	22	21	20

Figur 3 Sammanställning av bedömningen av datakvalitetslösningarna.

Bedömningsskala:

1 = datakvalitetsfaktorn hanteras inte

2 = datakvalitetsfaktorn hanteras övergripande

3 = datakvalitetsfaktorn hanteras utförligt

4 = datakvalitetsfaktorn hanteras mycket utförligt

7 Analys

Detta kapitel presenterar en analys av materialet som togs fram i kapitel 6. Följande problemprecisering har i detta arbete presenterats i kapitel 3.1 och lyder:

Hur väl hanteras datakvalitetsfaktorerna: korrekthet, datatypsenlighet, konsistens, integration, kompletthet, lämplighet och inga dubbletter av marknadens mest förekommande datakvalitetsverktyg?

I kapitlet kommer först det material som framkommit om respektive datakvalitetsfaktor att analyseras. Därefter följer en sammanfattande analys för att upptäcka likheter och skillnader i hur väl behandlingen av datakvalitetsfaktorer görs av datakvalitetsverktyg.

7.1 Analys av litteraturstudien

Avsikten med litteraturstudien var att belysa den ställda problempreciseringen för detta arbete. Litteraturstudien har fokuserats på tidigare nämnda datakvalitetsfaktorer i de utvalda datakvalitetsverktygsföretagen. Företagen som använts i detta arbete är Ascential Software som betecknas företag 1, Trillium Software som betecknas företag 2, Innovative Systems som betecknas företag 3 och Firstlogic som betecknas företag 4. Beteckningarna för ovannämnda företag används i detta kapitel för att förtydliga det material som presenterades i kapitel 6 och som visar hur nämnda företag behandlar de tidigare utvalda datakvalitetsfaktorerna för detta arbete.

7.1.1 Korrekthet

Korrekthet är en datakvalitetsfaktor som beskrivs i kapitel 2.3 och innebär att den information data bidrar till ska stämma med hur verkligheten ser ut. Datakvalitetslösningarna i de utvalda företagen för detta arbete behandlar samtliga korrekthet på ett genomtänkt vis men har lite olika sätt att hantera faktorn på. Företag 1 beskriver t.ex. ett mycket noggrant och omfattande tillvägagångssätt genom att dela upp data i mindre delar för att lättare urskilja och eliminera förkortningar och felstavningar. Företag 2 har en annan teori i sin beskrivning som bygger på att datakvalitetsverktyget förstår olika språk och på detta sätt ökar förståelsen av data som ska lagras i datalagret. Samtidigt beskrivs att företag 2 har datakvalitetsverktyg som innehåller lexikonliknande listor att tillgå, vilket leder till att omfattande och vida kontroller av korrektheten i olika data kan göras. Företag 3 har en idé som beskrivs bygga på en heuristisk kunskapsbas och med hjälp av dess innehåll ska hitta felaktigheter i de datauppgifter som ska laddas in i datalagret. Kunskapsbasen byggs upp av data som tidigare passerat datakvalitetsverktyget och växer på så sätt i omfattning, vilket medför att korrektheten blir hög genom att kunskapsbasen hjälper till att hitta felaktigheter i innehållet, strukturen eller relationerna. Företag 4 har en lösning som beskrivs kunna identifiera data med hjälp av analyser över var data behöver förbättras innan laddning in i datalagret. För att öka korrektheten innehåller datakvalitetsverktyget information över hur ett stort antal länder hanterar adressdata, vilket leder till att korrektheten i bl.a. adresser kan kontrolleras i hög grad.

7.1.2 Datatypsenlighet

Datakvalitetsfaktorn datatypsenlighet beskrivs i kapitel 2.3 och innebär att data lagras enligt den datatyp ett visst fält i datalagret är avsett för. Datakvalitetslösningarna har här ett ganska likartat sätt att hantera datatypsenlighet men dock skiljs de åt i vissa avseenden. Företag 1 och företag 2 har lösningar som beskriver att data analyseras genom identifiering av tecken och rader för att komma fram till vad varje datafält innehåller. Detta tyder på noggranna genomgångar för att kontrollera att data inte hamnar i felaktiga datafält i datalagret. Företag 1 skapar dessutom en rapport som visar vad varje datafält innehåller. Företag 3 och företag 4 har lösningar som också analyserar data som ska laddas in i datalagret för att identifiera datatyper. Beskrivningarna som ges är dock så övergripande att de inte återger en känsla av hur väl datatypsenlighet behandlas av datakvalitetsverktygen utan bara att hantering görs.

7.1.3 Konsistens

Som beskrivs i kapitel 2.3 innebär datakvalitetsfaktorn konsistens att data hålls enligt standarder för att uppnå ett kontinuerligt lagringssätt i datalagret. Urvalet av datakvalitetslösningar i detta arbete använder olika metoder för att uppfylla faktorn konsistens. Företag 1 har en lösning som beskrivs använda ett sätt att få datadelar i rätt ordning med ett mönstermarkeringssystem. Markeringarna placeras om för att få fram mönster som liknar varandra i utformning jämfört med redan lagrad data. Tillvägagångssättet som används medför en noggrann och utförlig metod för att se till att data presenteras på ett kontinuerligt sätt i datalagret. Företag 2 tar hjälp av fler källsystem än data som redan finns lagrad i datalagret. Genom att använda lexikonliknande listor och externa källsystem ökar datakvalitetsverktygets möjlighet att finna standardiserade sätt att lagra olika data på om det inte redan finns ett standardsätt i datalagret. Företag 3 har ett omfattande tillvägagångssätt genom kunskapsbasen med stora mängder data som ökar möjligheterna att hitta sätt att lagra data på ett enhetligt vis. Dessutom finns standarder för postadresser i USA redan inbyggt i datakvalitetsverktyget. Företag 4 har en lösning som beskrivs matcha data som ska laddas in i datalagret och innehåller standarder för hur ett flertal länder lagrar adresser. Datakvalitetslösningen använder alltså redan lagrad data i datalagret för att hitta sätt att lagra data enhetligt. Detta innebär att det inte i lösningen finns hjälp om standardiserade sätt om det inte redan finns ett sätt att lagra viss data på i datalagret.

7.1.4 Lämplighet

Lämplighet är en datakvalitetsfaktor som handlar om att data ska vara av det slaget som användarna i en verksamhet har användning av och beskrivs i kapitel 2.3. Samtliga datakvalitetslösningar ger endast en övergripande inblick i hur lämplighet behandlas av respektive lösning. Genomgående för lösningarna är en kort beskrivning av att regler sätts upp för den aktuella verksamheten. På så sätt ger datakvalitetslösningarna ett intryck att vara flexibla och att de kan användas i olika typer av verksamheter då verksamheten avgör vilka regler som är viktiga för att få fram önskad data.

7.1.5 Integration

Datakvalitetsfaktorn integration nämns närmare i kapitel 2.3 och innebär att information kombineras från olika källsystem till ett enda system. Samtliga företag med datakvalitetslösningar som tas upp i detta arbete behandlar integration på ett lika övergripande beskrivningssätt. Beskrivningen som ges av lösningarna säger endast att data från olika källsystem sammanförs till ett datalager. Med den beskrivningen kan bara en övergripande förståelse fås av att integration behandlas i samtliga datakvalitetslösningar men ger inget intryck av hur väl varje enskild lösning genomför hanteringen.

7.1.6 Kompletthet

Kompletthet är en datakvalitetsfaktor som innebär att all data som kan behövas finns tillgänglig, vilket förklaras i kapitel 2.3. Det ges olika utförliga beskrivningar av hur de olika datakvalitetslösningarna hanterar kompletthet. Företag 1 har en lösning som utförligt beskriver hur relationer dras mellan data som finns lagrad och kommer på så sätt fram till data som kompletterar varandra. Företag 2, företag 3 och företag 4 har lösningar som beskriver en liknande grundtanke som företag 1 d.v.s. att data sammanförs på något sätt för att komplettera den information som datalagret innehåller. Lösningarna i företag 2, företag 3 och företag 4 beskriver även en berikning av data med data från externa källsystem, vilket leder till komplettare uppgifter än de som fås av enbart redan lagrad data i datalagret.

7.1.7 Inga dubbletter

I kapitel 2.3 förklaras innebörden av datakvalitetsfaktorn inga dubbletter. Förklaringen lyder att samma data från samma tidpunkt inte lagras mer än en gång i datalagret. Datakvalitetslösningarna ger beskrivningar som alla förklarar hur faktorn inga dubbletter hanteras på ett tydligt sätt. Företag 1 beskrivs använda mönstermarkeringen som nämnts tidigare i kapitlet för att gruppera ihop de mönsterkombinationer som liknar varandra. Denna grupperingsteknik som lösningen använder beskriver ett tydligt och utförligt tillvägagångssätt att reducera dubbletter. Företag 2 ger en tydlig beskrivning över hur dubbletter elimineras genom att upptäcka data med samma innebörd. Företag 3 använder liknande tillvägagångssätt som företag 2 genom att finna data som hör ihop på något sätt och grupperar dessa i grupper som liknar det sätt företag 1 eliminerar dubbletter. Företag 4, använder som företag 2 och företag 3, en taktik som går ut på att sammanföra data med likheter för att få bort eventuella dubbletter innan data laddas in i datalagret. Samtliga datakvalitetslösningar återger beskrivningar som inger en tydlig förståelse över hur datakvalitetsfaktorn hanteras i respektive lösning.

7.2 Sammanfattande analys

Litteraturstudierna har skapat en förståelse för vad datakvalitetsverktyg kan hantera utifrån arbetets problemprecisering. Samtliga datakvalitetslösningar behandlar de datakvalitetsfaktorer som tagits upp i detta arbete. Det kan dock urskiljas vissa olikheter och likheter i hur väl faktorerna hanteras av lösningarna.

Alla datakvalitetslösningar ger en tydlig förklaring över hur korrekthet behandlas av respektive lösning. Av förklaringen framgår att lösningarna hanterar korrekthet väl, dock har lösningarna valt olika tillvägagångssätt men samtliga sätt ger intrycket av att vara noggranna och uppfylla faktorn korrekthet med gott resultat. Skälet till att alla lösningar noggrant beaktar korrekthet kan vara en ökad medvetenhet hos företag av vikten med rätt uppgifter i sina datalager. Detta kan i sin tur ha drivit datakvalitetslösningarna till att utveckla utförliga metoder till att uppfylla korrekthet.

I beskrivningarna av samtliga datakvalitetslösningar framgår också att datatypsenlighet hanteras med samma grundtanke, nämligen genom analys av data. Företag 1 och företag 2 beskriver hur analyserna görs samt ger ett intryck av att användarna av datakvalitetsverktygen får en större inblick i varför verktygen agerar som de gör. Företag 3 och företag 4 inger däremot inte känslan av att data noggrant kontrolleras för att identifiera innehållet i datafälten. Beskrivningen över lösningarnas datakvalitetsverktyg som används i detta arbete indikerar endast att någon slags analys görs och det framgår inte hur dessa analyser genomförs. Identifiering av datafält görs för att inte data ska lagras i ett fält som inte var avsett att innehålla den typen av data. Olikheter i utvecklingen av hur noggrant datatypsenlighet undersöks av respektive lösning kan bero på olika behov och krav av de verksamheter lösningarna varit i kontakt med.

Genom olika standarder kan data skrivas på ett sätt som är konsistent och data kan lagras på ett kontinuerligt vis i datalagret. Dessa standarder kan tas ifrån själva datalagret ifall data av samma slag redan finns lagrad. Finns inte den typ av data i datalagret kan frågetecken uppstå i frågan över vilket sätt data ska lagras på. En omfattande lösning på detta problem har företag 2 och företag 3 hittat i att ha tillgång till stora mängder data i form av listor, externa källor och kunskapsbaser. Företag 4 har endast vissa adressstandarder inbyggda i datakvalitetslösningen och en koncentration på adresser tillför inte en omfattande hantering som är fallet för företag 2 och företag 3. Företag 1 saknar resurser från annat håll än datalagret och kan därför inte kontrollera om det finns något standardiserat sätt att lagra data som inte redan finns i datalagret. Anledningen till att omfattningen av den behandling som görs av respektive lösning skiljer sig åt kan bero på stora variationer på krav av företagen som använder datakvalitetsverktygen. Vissa företag kan beroende av verksamhetsområde vara mer angelägna än andra om att uppgifterna i datalagret följer såväl nationella som internationella standarder. Andra företag kanske inte ser något behov annat än att data i det egna datalagret lagras på likartat sätt oberoende av om sättet är ett standardiserat sätt att hantera data på eller ej.

Att data tas fram som är lämplig gör datakvalitetslösningarna genom att skräddarsy regler. Hur väl detta görs av varje enskild lösning framgår inte av de litteraturstudier som gjorts i detta arbete. Samtliga lösningar indikerar dock att regler sätts upp utefter verksamhetens behov. Detta tillvägagångssätt visar att datakvalitetsverktygen i lösningarna är flexibla och användbara i olika verksamheter. Detta kan bero på att det är viktigt att verktygen kan användas i olika typer av företag med tanke på att behovet av data som passar en verksamhet är helt olikt en annan verksamhet. Om inte denna flexibilitet finns begränsas datakvalitetslösningarnas användningsområde markant och deras chans till överlevnad på datakvalitetsverktygsmarknaden minskar.

Datakvalitetslösningarna beskriver flyktigt att data integreras men visar ändå att integration beaktas och uppfylls. Hur väl respektive lösning genomför integration av källsystem framgår inte av beskrivningarna. Dock kan en slutsats dras att datakvalitetsverktygen uppfyller datakvalitetsfaktorn då data från källsystemen bearbetas för att laddas in i ett enhetligt system, d.v.s. datalagret. Anledningen till att beskrivningarna inte ger en tydlig bild över hur integration går till kan bero på att de företag som är intresserade av lösningen inte känner ett behov av att veta exakt hur detta genomförs. Ett annat skäl kan vara att de företag som använt verktyget inte känner till hur det genomfördes utan endast att det gjordes då denna vetskap kanske ligger på för teknisk nivå för användarna.

Data som relaterar till varandra sammanförs av datakvalitetslösningarna för att presentera så komplett information som möjligt för användarna. Samtliga lösningar ger intryck av att göra detta på ett omfattande sätt. Företag 1 saknar dock resursen att berika data med data från externa källsystem som övriga datakvalitetslösningar i detta arbete kan. Med denna resurs kan information som är kopplad till viss data inhämtas och komplettera data i datalagret automatiskt vilket tyder på att kompletthet i dessa fall tas omhand på ett mycket utförligt sätt. Kompletthet hanteras alltså i stor utsträckning av samtliga lösningar även om vissa visar en mer omfattande hantering än andra. Denna utveckling av lösningarna kan bero på att företag som använder datakvalitetsverktyg är medvetna om vikten av att ha tillgång till omfattande information angående ett ärende när beslut ska fattas. Kan all information som finns om ärendet presenteras på en och samma gång kan snabbare beslut fattas som gynnar företaget.

I detta arbete finns det två olika sätt att hantera datakvalitetsfaktorn inga dubbletter. Antingen grupperas liknande data ihop som är fallet för företag 1 eller så sker identifiering av data vilket företag 2 och företag 4 har valt att använda d.v.s. att data med samma innebörd elimineras. Företag 3 däremot har valt att använda båda tillvägagångssätten som nämnts ovan. Datakvalitetslösningarna har alltså valt att använda antingen ett av sätten eller en kombination av tillvägagångssätten. Vilket som bäst behandlar datakvalitetsfaktorn kan inte avgöras utifrån beskrivningarna som ges men båda sätten förklarar ett genomtänkt hanteringssätt av att eliminera dubbletter. Orsaken till att lösningarna hanterar reducering av dubbletter på likartat sätt kan bero på att detta är de bästa sätten att eliminera dubbletter på. En annan anledning kan vara att ingen annan lösning ännu har hittats och det finns kanske inget behov för tillfället då metoden fungerar tillfredställande för användarna av lösningarna.

Av alla skillnader och likheter som framtagits ovan mellan datakvalitetslösningarnas sätt att beakta datakvalitetsfaktorerna i detta arbete kan en slutledning fattas att lösningarnas hantering inte radikalt skiljer sig från varandra. Den totala bedömningen av datakvalitetslösningarna är ganska jämn och det är inte någon lösning som utmärkts för att vara betydligt bättre eller sämre än någon annan när samtliga datakvalitetsfaktorer beaktas samtidigt. Sammanfattningsvis kan sägas att lösningarna är medvetna om vad som är viktigt att kunna hantera i datakvalitets avseende i ett datalager då hänsyn tagits till samtliga datakvalitetsfaktorer som undersökts i detta arbete.

8 Resultat och diskussion

Det framkomna resultatet i analysen presenteras i detta kapitel som utgår från arbetets problemprecisering. Vidare görs en diskussion som reflekterar över litteraturstudien samt resultatet. Därefter diskuteras fortsatta arbeten inom problemområdet.

8.1 Resultat av litteraturstudien

Med litteraturstudien som utgångspunkt har studien visat på att datakvalitetslösningarna innefattar datakvalitetsverktyg som beaktar de datakvalitetsfaktorer som valts ut i detta arbete. På vilket sätt och hur väl denna beaktning görs framkommer i olika grad av litteraturstudierna över lösningarna, vilket nämns i stycket nedan. Vissa skillnader i de beskrivna tillvägagångssätten har bidragit till olika bedömningspoäng av hanteringen av enskilda datakvalitetsfaktorer som tidigare visats i kapitel 6.

Korrekthet behandlas med olika hanteringssätt, dock framgår att samtliga datakvalitetslösningar insett vikten av datakvalitetsfaktorn. Oavsett vilket hanteringssätt som används beaktas korrekthet väl.

Datatypsenlighet behandlas genomgående med samma grundidé av datakvalitetslösningarna. Samtliga lösningar hanterar datatypsenlighet med hjälp av analysmetoder men Ascential Software och Trillium Software inger ett utförligare hanteringssätt än Innovative Systems och Firstlogic.

Konsistens behandlas i störst omfattning av Trillium Software och Innovative Systems som tar hjälp från flera olika håll om standardiserade sätt saknas i datalagret. Övriga datakvalitetslösningar hanterar konsistens väl men saknar det lilla extra att ta till hjälp om det behövs.

Lämplighet behandlas på samma sätt av datakvalitetslösningarna med kundanpassade regler. Denna lösning ger lösningarna stora friheter i vilka verksamheter och sammanhang de kan användas. Hur dessa regler anpassas framkommer inte och därför upplevs bilden av hur hanteringen av lämplighet görs endast övergripande.

Integration behandlas med likartade tankar av verktygen vilket visar att datakvalitetslösningarna hanterar data från olika källsystem. Hur väl sägs inte i beskrivningarna över lösningarna men det framgår ändå att integration hanteras.

Kompletthet behandlas genom att relaterad data sammanförs och presenteras samlat för användaren, vilket tyder på att datakvalitetsfaktorn tas väl omhand av datakvalitetslösningarna. Trillium Software, Innovative Systems och Firstlogic går dessutom steget längre när dessa lösningar även tillför data, vilket visar ett ännu mer omfattande hanteringssätt av kompletthet.

Inga dubbletter behandlas med två tillvägagångssätt som antingen används separat eller i en kombination av datakvalitetslösningarna. Oavsett vilket tillvägagångssätt som används ges intrycket att datakvalitetsfaktorn uppfylls och inget av tillvägagångssätten verkar bättre eller sämre än det andra.

Avslutningsvis kan konstateras att alla datakvalitetsfaktorer behandlas av samtliga datakvalitetslösningar. Generellt sett skiljer sig inte lösningarna i sin helhet åt i någon större omfattning men däremot behandlas varje enskild faktor i olika utsträckning. För att välja den lösning som bäst passar en verksamhet bör därför en närmare titt på vilka datakvalitetsfaktorer som är viktiga för verksamheten göras. Utifrån dessa behov som då kommer fram kan sedan den lösning väljas som behandlar de önskade faktorerna på lämpligast sätt.

8.2 Diskussion

Syftet med detta arbete var att se om datakvalitetsfaktorer som är betydelsefulla för data i ett datalager hanteras av datakvalitetsverktyg vid inladdning av data i datalager. Arbetet förväntades också ge upplysningar om vilka åtgärder datakvalitetsverktyg har att tillgå vid datakvalitetsproblem.

I de litteraturstudier som gjorts har artiklar hämtats från tidskrifter som är väl förankrade inom arbetets ämnesområde. Dessa tidskrifter kan anses tillförlitliga då de källor som utgör artikeln är kontrollerade. Erfarenheter som fåtts från litteraturstudierna är att kontrollera skribenterna till artiklarna för att inte inhämta information som enbart presenterats av företaget som säljer datakvalitetslösningen. Genom att samla information från källor som kan ge en objektiv bedömning på datakvalitetslösningarna har risken till en subjektiv syn i det insamlade materialet minskats, vilken hade varit större om materialet enbart tagits från respektive försäljare av datakvalitetslösningar. P.g.a. att betygssättningen av hur väl datakvalitetsfaktorerna hanteras grundas i författarens egna tolkning av litteraturen kan bedömningen vara subjektiv. Det är möjligt att någon annan kan tolka annorlunda och dra andra slutsatser än de som gjorts i detta arbete. Litteraturstudien har begränsat möjligheten att utbyta egna idéer och frågor som uppstått under genomgång av litteratur över datakvalitetsverktygens olika funktioner. En intervjuperson med kunskaper och erfarenheter inom området hade möjliggjort chansen att ställa följdfrågor och på så sätt inskaffat information i form av upplevelser som kanske inte framgått i samma utsträckning i litteraturen. Möjligheten att finna just denna person var dock liten och att denna person också skulle ha möjlighet att avsätta tid till en intervju ännu mindre, vilket var anledningen till att detta alternativ tidigt uteslöts ur arbetet. Även om arbetets resultat baserats på enbart litteraturstudie anses inte detta ha påverkat resultatet nämnvärt. Möjligtvis hade mer detaljerad information erhållits om möjligheten till följdfrågor funnits men även medfört en risk i att åsikter från en intervjuperson skulle ha påverkat den egna tolkningen av litteraturen.

I resultatet framgår att datakvalitetsfaktorerna som tagits upp i detta arbete hanteras i olika grad av datakvalitetslösningarna. De som tagits upp är de faktorer som ansetts mest relevanta, vilket innebär tekniskt inriktade då arbetet handlar om vad verktyg kan hantera. Urvalet av vilka faktorer som skulle ingå i arbetet har som tidigare nämnts baserats på att flera författare tagit upp dessa i den litteratur som studerats. Detta innebär alltså att det kan finnas andra faktorer att beakta än de som gjorts och som kan påverka resultatet som framkommit i detta arbete. Studien som gjorts tyder på att datakvalitetslösningar kan underlätta hanteringen av datakvalitetsproblem och höja datakvalitet i datalager. Detta med tanke på att lösningarna bl.a. kan anpassas och

på så sätt användas på all typ av data som kan finnas i olika verksamheter. Vidare pekar studien på att företag med fördel kan använda datakvalitetsverktyg vid laddning av företagets datalager då dessa verktyg beaktar faktorer som bidrar till hög datakvalitet. Vissa datakvalitetslösningar verkar dock mest fokuserade på den amerikanska marknaden men några av verktygen har en mer global syn genom att lösningarna hanterar språk och standarder för andra länder än USA. Att inte alla datakvalitetslösningar är lika internationella antas bero på att marknaden för datakvalitetsverktyg fortfarande är ganska liten. Om förståelsen av datakvalitetens betydelse i företags datalager ökar kommer också efterfrågan på datakvalitetsverktyg att öka. Vilket förmodligen kommer att bidra till att säljare av datakvalitetslösningar också kommer att öka verktygens internationella standard med tiden.

Resultatet som framkommit i arbetet har överrensstämt väl med det förväntade resultatet då svar på hur väl datakvalitetsverktygen hanterar de utvalda datakvalitetsfaktorerna har kunnat ges. Indikationer har framkommit att verktygen har åtgärder mot låg datakvalitet och kan därför lösa datakvalitetsproblem som uppstår i företags datalager.

8.3 Fortsatt arbete

I detta arbete har en studie över datakvalitetsverktygs hantering av tidigare nämnda datakvalitetsfaktorer gjorts. Det finns förmodligen mycket mer arbete att göra inom detta område och förslag på fortsatta arbeten ges nedan.

- Fortsättning på detta arbete är förslagsvis att titta på fler eller andra företag som säljer datakvalitetsverktyg för datalager. Detta för att finna likheter eller olikheter i hur datakvalitetsfaktorer behandlas av ett större antal verktyg än de som använts i detta arbete. En undersökning av andra verktyg skulle kunna komplettera resultatet som framkommit i detta arbete och bidra med en mer omfattande bild över hur datakvalitetsfaktorer hanteras.
- Det kan också tänkas att en testning av enskilda datakvalitetsverktyg i praktiken kan vara ett fortsatt arbete. Ett intressant arbete för att göra en egen bedömning av verktygen genom att själv använda verktygen för att löser specifika datakvalitetsproblem. Arbetet skulle kunna ge detaljerade uppgifter om verktygs tillvägagångssätt samt egna upplevelser som fåtts från testningen.
- En lämplig fortsättning på detta arbete kan också vara att se om datakvalitetsverktygen har utvecklats och hittat andra lösningar på de datakvalitetsfaktorer som tagits upp i detta arbete. Detta vore ett intressant arbete för att se hur utvecklingen på verktygen går. Arbetet skulle också kunna innefatta förslag på olika alternativ som kan användas för att hantera datakvalitet i datalager.
- En annan riktning som ett fortsatt arbete skulle kunna ta är att göra en fallstudie på ett företag som använder en datakvalitetslösning för att göra en egen bedömning över hur lösningen fungerar och hanterar datakvalitetsproblem. Arbetet vore intressant för få en inblick i hur datakvalitetsverktygen fungerar i en verksamhet och hur de uppfattas av användarna.

9 Referenser

Andersen, E. S. & Schwencke, E (1998) *Projektarbete - en vägledning för studenter*, Lund: Studentlitteratur

Agosta, L. (2000) *The essential guide to data warehousing*, New Jersey: Prentice Hall

Ascentialsoftware.com (2002) Introdction to Integrity – Resolving Parts Descriptions,

Tillgänglig på Internet: http://www.ascentialsoftware.com/cgi-bin/dataquality.cgi?URL=WP_Resolving_Parts_3.pdf [Hämtad 03.03.09]

Bischoff, J. & Alexander, T (1997) *Practical advice from the experts*, New Jersey: Prentice Hall

Bova, L. (april, 2001) OSC sharpens its competitive edge with Firstlogic´s information quality suite, *DM Review*, Tillgänglig på Internet: http://www.dmreview.com/master.cfm?NavID=71&EdID=3199 [Hämtad 03.03.08]

Bright, P. (april, 2001) The Trillium Software System creates a cleansed, standardardized, customer-centric view for the Woolwich, *DM Review*, Tillgänglig på Internet: http://www.dmreview.com/master.cfm?NaID=71&EdID=3197 [Hämtad 03.03.08]

Clare, K. (maj, 2000) Norwich union constructs and maintains consolidated customer view with Trillium Software System, *DM Review*, Tillänglig på Internet: http://www.dmreview.com/master.cfm?NavID=71&EdID=2209 [Hämtad 03.03.09]

Connolly, T. & Begg, C. (2002) *Database systems: a practical approach to design, implementation and management.* (Third Edition) Harlow: Addison-Wesley.

Database trends and applications (december, 2001) Innovative Systems Unviels i/Lytics Data Managemaent Suite, *Database trends and applications*, Tillgänglig på Internet: http://www.innovativesystems.com/print/dbt_ilytics.htm [Hämtad 03.03.09]

Dawson, C. H. (2000) *The essence of computing projects: a student's guide*, Harlow: Prentice Hall

Dekle, J. (april, 2002) Firstlogic creates a high-quality data set for Markettouch, *DM Review*, Tillgänglig på Internet: http://www.dmreview.com/master.cfm?NavID=71&EdID=4974 [Hämtad 03.03.08]

9 Referenser

Dmreview.com (januari, 2002) Trillium Software Technology Embadded into Oracle Warehouse Builder, *DM Review,* Tillgänglig på Internet: http://www.dmreview.com/editorial/dmreview/print_action.cfm?EdID=4568 [Hämtad 03.03.09]

Eckerson, W. W. (2002) *Data quality and the bottom line: achieving business success through a commitment to high quality data.* The Data Warehousing Institute. Tillgänglig på Internet: http://www.dw-institute.com/dqreport/ [Hämtad 03.02.07].

Elmasri, R. & Navathe, S. B. (2000) *Fundamentals of database systems,* Massachusetts: Addison-Wesley

Espiner, I. (juli, 2002) Innovative´s customer data management software ensures smooth demutualization, *DM Review,* Tillgänglig på Internet: http://www.dmreview.com/master.cfm?NavID=71&EdID=5407 [Hämtad 03.03.09]

Firstlogic.com (2003a) Tillgänglig på Internet: http://www.firstlogic.com/solutions/iq/dqa.asp [Hämtad 03.04.04]

Firstlogic.com (2003b) Tillgänglig på Internet: http://www.firstlogic.com/solutions/iq/iqSuite/dataCleansing.asp [Hämtad 03.04.04]

Firstlogic.com (2003c) Tillgänglig på Internet: http://www.firstlogic.com/solutions/iq/iqSuite/matching.asp?print=1& [Hämtad 03.04.04]

Forino, R. (mars, 2001) Data e.quality: A Behind-the-Scenes Perspective on Data Cleaning. *DM Review.* Tillgänglig på Internet: http://www.dmreview.com/master.cfm?NavID=198&EdID=3202 [Hämtad 03.02.24].

Hillman, W. (april, 2000) Trillium reduces data anomalies in trading partners database for 3M, *DM Review,* Tillgänglig på Internet: http://www.dmreview.com/master.cfm?NavID=71&EdID=2097 [Hämtad 03.03.09]

Hudson, D. (april, 2003) Trillium Software System helps drive VolvoIT, *DM Review,* Tillgänglig på Internet: http://www.dmreview.com/master.cfm?NavID=194&EdID=6531 [Hämtad 03.04.04]

Inmon, W. (2002) *Building the data warehouse* (Third Edition), New York: John Wiley & Sons.

9 Referenser

Innovativesystems.com (2003) Tillgänglig på Internet: http://www.innovativesystems.com/pdf/iLytics_data_qual_ds.pdf [Hämtad 03.04.04]

Kachur, R. (2000) *Data warehouse management handbook*, New Jersey: Prentice Hall

Kimball, R. & Ross, M. (2002) *The data warehouse toolkit: the complete guide to dimensional modeling* (Second Edition), New York: John Wiley & Sons.

Kimball, R., Reeves, L., Ross, M. & Thornthwaite, W. (1998) *The data warehouse lifecycle toolkit: expert methods for designing, developing and deploying data warehouses*, New York: John Wiley & Sons.

King, E. (2000) *Data warehouse and data mining: implementing strategic knowledge management*, South Carolina: Computer Technology Research Corp.

Knowledgestorm.com (2003a) Tillgänglig på Internet: http://knowledgestorm.com/ActivityServlet?ksAction=displaySolutionPrintVersion [Hämtad 03.04.04]

Knowledgestorm.com (2003b) Tillgänglig på Internet: http://knowledgestorm.com/ActivityServlet?ksAction=displaySolutionPrintVersion [Hämtad 03.04.04]

Knowledgestorm.com (2003c) Tillgänglig på Internet: http://knowledgestorm.com/ActivityServlet?ksAction=displaySolutionPrintVersion [Hämtad 03.04.04]

Patel, R. & Davidson, B. (1994) *Forskningsmetodikensgrunder: att planera, genomföra och rapportera en undersökning* (Andra upplagan), Lund: Studetlitteratur.

Prillaman, D. (mars, 2001) First union national bank gains a compeditive edge with he Trillium Software System, *DM Review*, Tillgänglig på Internet: http://www.dmreview.com/master.cfm?NavID=71&EdID=3103 [Hämtad 03.03.09]

Raab, D. (februari, 2002) Vality Integrity, *DM News*, Tillgänglig på Internet: http://raabassociates.com/a202vali.htm [Hämtad 03.03.23]

Reynolds, S. (april, 2000) Integrity provides data investigation, standardization and consolidation for Pegasus systems, *DM Review*, Tillgänglig på Internet: http://www.dmreview.com/master.cfm?NavID=71&EdID=2098 [Hämtad 03.03.09]

Sanantoniobusinessjournal.com (januari, 2002) Oracle adds Harte-Hanks software to data warehouse package, *San Antonio Business Journal,* Tillgänglig på Internet: http://sanantonio.bizjournals.com/sanantonio/stories/2002/01/21/daily7.html [Hämtad 03.03.09]

Sandholm, L. (2001) Kvalitetsstyrning med totalkvalitet: *verksamhetsutveckling med fokus på totalkvalitet* (Femte upplagan), Lund: Studentlitteratur.

Schauer, J. (februari, 2000) Trillium Software: Quality Products, Quality People, *DM Review,* Tillgänglig på Internet: http://www.dmreview.com/editorial/dmreview/print_action.cfm?EdID=1852 [Hämtad 03.03.09]

Schwalb, S. (mars, 2002) Spotlight On: Innovative Systems, *Internetworld,* Tilgänglig på Internet: http://www.internetworld.com/news.php?inc=crm/03052002d.html [Hämtad 03.03.09]

Standardiseringskommissionen i Sverige (1992*) ISO 9000 international standards for quality management,* Göteborg: Novum Grafiska AB

Trilliumsoftware.com (2003) Tillgänglig på Internet: http://www.trilliumsoftware.com/site/content/product/methodology.asp [Hämtad 03.03.09]

Uhl, D. (april, 2002) Cargill ensures supplier and customer data quality with Vality, *DM Review,* Tillgänglig på Internet: http://www.dmreview.com/master.cfm?NavID=71&EdID=4982 [Hämtad 03.03.09]

Vassiliadis, P., Bouzeghoub, M. & Quix, C. (2000) Towards quality-oriented data warehouse usage and evolution, Stor Brittanien: *Information Systems Vol. 25, No. 2,*89-115

Williams, J. (1997) Tools for traveling data, *DBMS and Internet Systems,* Tillgänglig på Internet: http://www.dbmsmag.com/9706d16.html [Hämtad 03.03.24].

Williams, N. (april, 2003) University reaps the benefits of improved data quality, *DM Review,* Tillgänglig på Internet: http://www.dmreview.com/master.crf?NavID=194&EdID=6524 [Hämtad 03.04.04]

Wright, C. (juli, 2002) Avid puts a wrap on successful information quality initiative, *DM Review,* Tillgänglig på Internet: http://www.dmreview.com/master.cfm?NavID=71&EdID=5386 [Hämtad 03.03.09]

Yee, R. (juli, 1999) Hongkong bank of Canada understands total customer profitability with Trillium Software, *DM Review,* Tillgänglig på Internet: http://www.dmreview.com/master.cfm?NavID=71&EdID=1159 [Hämtad 03.03.09]